Identification of prostate cancer diagnostic and prognostic biomarkers in urine expression data with a focus on extracellular vesicles

Helen Marie Curley

This thesis is submitted for the degree of
Doctor of Philosophy

University of East Anglia
School of Biological Sciences
September 2018

I would like to dedicate this thesis to Andy Ripley.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Helen Marie Curley
September 2018

Acknowledgements

Mostly I would like to thank my supervision team, Professor Colin Cooper, Dr Daniel Brewer and Dr Jeremy Clark. Dr Daniel Brewer was supportive throughout the whole PhD particularly throughout the challenging task of writing the thesis. Dr Jeremy Clark was very helpful and gave me a great amount of insight into the field of prostate cancer research. I would also like to thank the other members of the lab who made this research possible; Dr Rachel Hurst and Marcelino Yazbeck-Hanna.

A huge thank you goes to the friends and family of Andy Ripley, who funded the studentship. I would also like to thank my family and friends who have been there for me throughout.

Abstract

Prostate Cancer (PCa) is a major clinical problem worldwide with considerable variability in clinical outcome of patients. PCa diagnostics and prognostics currently lack specific and sensitive clinical biomarkers and treatment is not well individualised. The *PCA3* test, amongst others, highlights the utility of urine in PCa diagnostics and prognostics. Urine contains cells and extracellular vesicles (EV) that originate in the prostate. There are many areas of the PCa clinical process that could be aided with an expression based urine test, including diagnosis, prognosis and response to therapy.

NanoString data (167 transcripts) from 485 EV RNA samples were collected from PCa patients and used to build models that would aid in PCa diagnosis and prognosis i.e. i) PCa (low- (L), intermediate-(I), and high-risk(H)) vs CB (Clinically Benign/No evidence for cancer), ii) high-risk PCa vs CB, and iii) trend in expression across CB>L>I>H. These models were validated in 235 samples, with AUCs of i) 0.851 ii) 0.897 and iii) 0.709, respectively.

The potential of using urine EVs to predict patient response to treatments was also investigated. In a pilot data set a signature of seven transcripts was identified that could optimally predict progression of patients on hormone therapy ($p = 2.3\text{x}10^{-05}$; HR = 0.04288). Models were also built using NanoString data from 92 cell RNA samples. Intercomparing expression data from matched cell and EV fractions of urine showed that transcripts significantly higher in the EV samples were associated with the prostate, PCa and cancer in general, supporting them as a viable source of biomarkers in the clinical management of PCa.

In conclusion my analyses have demonstrated the utility of examining urine RNA for the diagnosis and prognosis of PCa. My studies have formed the basis of the production of a Prostate Urine Risk test that is currently under development at UEA.

Table of Contents

List of Figures

List of Tables

Abbreviations

A – Advanced

ADT - Androgen deprivation therapy

AED – androstenedione

AIC – Akaike Information Criteria

AMACR - Alpha-methylacyl-CoA racemase

ANOVA - Analysis of variance

AR – Androgen Receptor

AUC – Area under the Curve

BCR - Biochemical recurrence

BPH - Benign prostatic hyperplasia

Bx - Biopsy

CB – No evidence for cancer

CRPC – Castration Resistant Prostate Cancer

CTCs - circulating tumour cells

DHT – Dihydrotestosterone

DNA - Deoxyribonucleic acid

DRE – Digital Rectal Examination

ELISA - Enzyme-linked Immunosorbent Assay

EV - Extracellular vesicle

FDA - U S Food and Drug Administration

FISH - Fluorescent In-situ Hybridisation

FFPE - Formalin-Fixed Paraffin-Embedded

FOV - field of view

GAP1 – Global Action Plan 1

GLM - Generalised Linear Model

GWAS - Genome-wide association studies

H - High risk

HG-PIN – High grade prostatic intraepithelial neoplasia

HR – Hazard Ratio

HT – Hormone Therapy

IHC – Immunohistochemistry

I - Intermediate risk

IQR – Interquartile range

KM - Kaplan Meier

L – Low risk

LASSO - Least absolute shrinkage and selection operator

LHRH – Luteinizing Hormone Releasing Hormone

lincRNA - long noncoding RNA

LPD – Latent Process Decomposition

MFISH - multi-fluorochrome assays

MIP - molecular inversion probes

MMPs - Matrix metalloproteinases

MRI - Magnetic resonance imaging

MS - Mass spectrometry

NICE - National Institute for Health and Care Excellence

NF - normalisation factor

NGS – Next generation sequencing

NNUH - Norfolk and Norwich University Hospital

OOB - Out-of-bag

PCA – Principal Component Analysis

PCa – Prostate Cancer

PCA3 – Prostate Cancer Antigen 3

PCR - Polymerase chain reaction

PPV – Positive Prediction Value

PR- - progesterone receptor negative

PSA – Prostate Specific Antigen

ROC – Receiver Operator Characteristic

RMH - Royal Marsden Hospital NHS Foundation Trust

RNA - Ribonucleic acid

RF – Random forest

RP - Radical prostatectomy

RT - Reverse transcriptase

RTPCR - Reverse transcription polymerase chain reaction

SD – Standard deviation

SKY - Spectral Karyotyping

SNP - Single nucleotide polymorphism

SNV - Single nucleotide variant

TIMPs - Tissue inhibitors of metalloproteinase

TNBC - Triple negative breast cancer

TNM – Tumour, Lymph Nodes, Metastasis

TRUS - Transrectal ultrasound guided

UPGMA - Unweighted pair-group method using arithmetic averages

WHO – World Health Organisation

1

Introduction

1.1 The Research Gap

Prostate cancer (PCa) is the second most common male cancer worldwide and the most common in the UK[1]. The current available biomarkers for PCa lack specificity and/or sensitivity to detect the disease and are unable to distinguish indolent from aggressive disease or predict treatment response. PCa is generally slow-growing, the vast majority requiring no therapeutic intervention at all whilst some of these cancers progress to fatal disease. There is no genetic stratification for treatment unlike many other cancer types, PCa is instead treated with a risk-adjusted patient specific method[2] that aims to improve the control of the cancer whilst reducing risk of complications from treatment. Biopsies

are commonly performed at diagnosis, but can miss the cancerous area of the prostate and thus lead to a misdiagnosis of "no cancer". There are limitations to biomarkers capable of predicting positive subsequent biopsy results. There is an urgent clinical need for biomarkers to determine which patients have PCa, which patients have disease that will progress rapidly, and individualise treatment to optimise response.

1.2 Biomarkers

Biomarkers have become widely used in clinical and basic research. The National Institute of Health defines biomarkers as "characteristics that are objectively measured and evaluated as indicators of normal biological processes, pathogenic processes, or pharmacological responses to therapeutic intervention"[3]. Whilst the WHO (World Health Organisation) have a much broader definition that also includes measurable effects of exposure to chemicals or nutrients that allow for risk assessment[4]. Clinically they are used for diagnosis (identification of disease), prognosis (predicting the likely course/outcome of the disease), treatment response stratification and monitoring treatment response in patients. Examples range from blood pressure to more complex genetic screens of tissues, blood, urine and other samples[5].

1.2.1 Biomarkers in Cancer

Within the field of cancer management, biomarkers are used for risk assessment, diagnostics, prognostics, treatment stratification and monitoring the effects of treatments. Tumour biomarkers are any measurable molecule that is either produced by the tumour itself or through the host's response to the tumour that indicates the presence of cancerous processes. Tumour biomarkers can be proteins, glycoproteins, antigens, hormones, receptors, metabolites, and genetic markers; including DNA and RNAs and their epigenetic changes[6].

Examples of biomarkers in risk assessment include hereditary germ line mutations that increase a person's risk of developing a certain type of cancer, for example, presence of

germ line *BRCA1* or *BRCA2* mutations increases the crude life time rate (number of incidences within a population during a specific time period, not considering subgroups within the population) risk of breast cancer in women from 12.5% to 65% and 45%, respectively. Likewise in ovarian cancer, crude rate risk increases from 0.02% to 39% and 11%, respectively[7]. *BRCA* mutation screens are offered to people with known family history of these cancers and positive results can lead to optional preventive measures (e.g. a mastectomy). Other risk assessment biomarkers include p53 but it's mutant occurrence in such a range of cancers (50% of all cancers) makes it unusable for screening and diagnosis purposes. As, it could be detected but you would not know where the cancer was or if both alleles were mutated. Also, p53 mutation levels differ between cancer types also, for example, only 3-20% of PCas have a p53 mutation detected at diagnosis[8].

An example of a biomarker in use in cancer diagnostics is prostate specific antigen (PSA). Serum PSA is currently the first test for PCa diagnosis in the clinic, as elevated levels can suggest the presence of malignancy. PSA, however, does not have great specificity as discussed later: Section 1.4.1

Tissue inhibitors of metalloproteinase (*TIMP*s) are examples of prognostic biomarkers in cancer. *TIMP*s are glycoproteins able to promote proliferation and block apoptosis by inhibiting matrix metalloproteinases (*MMP*s). Increased levels of *TIMP*s have been shown to correlate with poorer survival in many cancers including multiple myeloma, melanoma, breast, lung, colorectal, gastric and head & neck cancers[9].

Examples of biomarkers in treatment stratification include Human Epidermal Growth Factor Receptor 2 (*HER2*) and Estrogen Receptor α (*ER*α) in breast cancer. *HER2* and *ER*α receptors may be over-expressed in the breast cancer cells and a simple molecular test (Immunohistochemistry (IHC)**)** can determine this. This allows treatments to be applied to target the expression profiles of different biomarkers. Herceptin is a drug that specifically targets *HER2*, whereas Tamoxifen is an *ER*α antagonist.

A biomarker for treatment stratification does not necessarily have to be the drug target. The monoclonal antibody therapies Cetuximab and Panitumumab, which target *EGFR* in colorectal cancers, can only be administered to a cohort of patients who have wild-type *KRAS*. *KRAS* is a signal mediator (extracellular ligand binding and intracellular transduction) between *EGFR* and the nucleus[10]. *KRAS* mutants provide a resistance to these monoclonal antibody therapies. *KRAS* mutations can also occur in response to these treatments and has been shown to be (non-invasively) detectable as early as 10 months prior to radiographic detection of disease progression, allowing administration of *MEK* inhibitors to delay or reverse the resistance[11].

For treatment resistance monitoring in lung cancer patients, a second *EGFR* mutation, Thr790Met, which can be acquired as a result of treatment or can be pre-existing, provides resistance to *EGFR* tyrosine kinase inhibitors and has been associated with a shorter progression-free survival. Therefore could be used to eliminate people out of the *EGFR* tyrosine kinase inhibitor treatment cohort[10].

1.2.2 Problems with the use of current and new biomarkers in clinical diagnostics

There is a striking discrepancy between the efforts made to discover cancer biomarkers and the number of biomarkers that actually make it into clinical practice[6]. Major investments have been made to identify and validate novel cancer biomarkers. Using the search terms novel biomarker cancer and new biomarker cancer, a literature search yields 5,358 hits in 2016 alone. Over the past 5 years (2012-2016), 29,775 papers were published using the same search criteria.

However, very few major diagnostic biomarkers have been put into clinical use in the last 25 years[12]. Clinical programs have promised to revolutionize the diagnosis of cancer and the management of its patients. Considerable improvements to how tumours are characterized at a molecular level have shifted treatments towards the use of

targeted therapies[13]. New PCa tests that have been developed recently include OncotypeDx[14] (section 1.6.3.1), Decipher[15] (section 1.6.4.1) and Prolaris[16] (section 1.6.3.2). However, there is a gap in the number of patients having these tests in clinic to help determine which therapies are suitable for them, and the number of patients that could benefit from these tests. In 2014, the NHS provided 39,298 molecular diagnostic tests for lung, colorectal and melanoma patients in England. Yet the demand was 59,294, leaving 15,929 patients without testing. If this demand was met, it is estimated that 3,552 patients would have been eligible for targeted therapies[17].

Effective cancer biomarkers need to produce a reliable, reproducible clinically useful assay that is cost effective[6]. The process between biomarker identification to a clinical assay used in practice is lengthy, expensive and convoluted; many researchers working on identifying biomarkers are unaware of clinical practice[6]. Even if a useful tumour biomarker is discovered in the lab there still must be commercialisation incentives in place to develop the assays. Before widespread clinical use the biomarker must be tested in many large datasets and trials carried out by pharmaceutical companies in partnership with academics and also optimised to increase predictive power. Therefore, it can be complicated to determine at which point patenting for the biomarker should be awarded. Regulatory authorities also play a crucial role in validation and quantification of biomarker assays to justify the test to health care providers[13].

1.2.3 Biomarkers pave the way for stratified treatment of cancers

The current goal of biomarker research is personalized medicine. It aims to provide targeted therapy for individual patients, given their specific clinical, genetic and environmental state. Cancer treatment success is often limited by the heterogeneity among patients; giving patients with genetically different cancers the same treatments can often lead to failure of response with toxic side effects[18].

Stratified medicine is considered the first step towards personalized medicine. It works by grouping patients via tumour mutations for targeted therapy, using omics

technologies. It has shown good results within breast cancer patients[19], amongst other cancers. Breast cancer patients are often stratified between *HER2*+ and *HER2*-, *ER*+ and *ER*-, *PR*+ and *PR*- and triple-negative groups. *HER2*+ and *ER*+ breast cancers can receive Herceptin and Tamoxifen, respectively: Biological therapies, which are targeted towards those specific receptors.

There is a subset of breast cancers known as triple negative breast cancer (TNBC), where the cancerous cells are *HER2*-, *ER*- and progesterone receptor negative (*PR*-). These cancers have proven to be difficult to treat in the past especially when in their late stages, but promising results have been seen using targeted treatments such as *EGFR* inhibitors and *VEGF* inhibitors that have been previously used for other cancers of different tissues[20,21,22].

In order for stratified medicine to be effective, biomarker assays that can be routinely applied are needed to accurately stratify patients into treatment cohorts. These assays need to be easily performed with minimal risk to the patient and include immediate or rapid return of the results to ensure early initiation of treatment[23].

1.3 Biomarkers in Prostate Cancer

1.3.1 Prostate Cancer

PCa is the second most common male cancer worldwide[24] and the most commonly diagnosed cancer in the UK[25]. In 2010 it accounted for 25% of all cancers diagnosed in men, with 40,975 cases. In 2012 an estimated 307,000 men died from PCa worldwide[24] whilst in the UK 10,721 males died of PCa in 2010; PCa is the second most common cause of cancer death in males. Detected incidence increased by 22% in the last decade and is the fifth fastest increasing cancer in males. Mortality rate, however, has fallen by 11% over the same period[26]; 81.4% of PCa patients survived for five or more years in the UK during 2005 – 2009. Both, the increased incidence rate and the decreased

mortality rate are associated with the use of the PSA test (section 1.4.1). Changes to classification of PCa deaths and improvements in treatment are also likely to have affected mortality rates. 90% of PCas are acinar adenocarcinomas that originate in the gland cells of the prostate[27]. In approximately 75-85% of PCas[2], the cancer originates in the peripheral zone rather than the transitional zone (Figure 0.1). The other 10% of PCas fall into different types: signet ring carcinoma, ductal adenocarcinoma, transitional cell (urothelial cancer), squamous cell cancer, carcinoid of the prostate, small cell cancer and sarcoma/sarcomatoid cancer[28]. These will not be considered in the rest of this thesis.

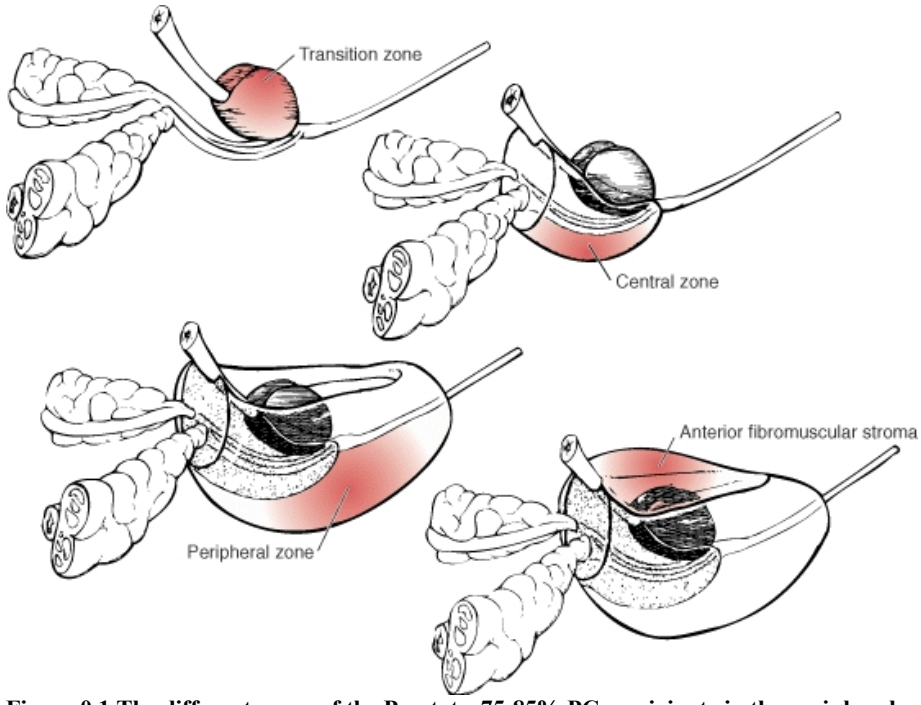


Figure 0.1 The different zones of the Prostate. 75-85% PCas originate in the peripheral zone, whereas, ~25% originate in the transitional zone. Adapted from Akin O., et al 2006[2].

1.3.2 Factors influencing PCa risk of incidence and progression

There are many factors influencing PCa risk including age, race and family history. PCa is primarily found in older men and risk of developing PCa increases with age. Between 2009 and 2011 36% of UK diagnosed cases of PCa were in men aged above 75, whilst

only 1% were in those younger than 50[29]. Men aged over 70 also had a statistically significant association with higher clinical stage and Gleason score[30].

African Americans have a 60% higher risk of developing PCa and mortality is approximately double that of white Americans[31], and a more aggressive form of the disease can be seen in African Americans[32]. In comparison, native Asian men show a much lower frequency of developing PCa; African American men show a 60-fold higher risk than those in Shanghai, China[31], although the incidence in Asian populations is increasing[33]. This extraordinary variation of occurrence across the world is boiled down to genetic and environmental factors, which is thought to largely include a Western diet. American-Japanese men have higher incidence rates of PCa than their counterparts in Japan, and this is independent of if they migrated early or late in life, suggesting that life style can accelerate progression of PCa[31]. Asian-American cohorts still hold a lower rate of incidence than white American men[34].

Evidence of familial risk of PCa has been seen from epidemiological studies, which suggest a two- to three-fold risk increase when there has been a first degree relative diagnosed. Familial clustering patterns have been seen in segregation studies that show high penetrance genetic mutations (including those at the putative susceptibility loci)[31]. PCa aggregates with other familial cancer types (like breast and ovarian). The genes that infer increased susceptibility to these cancers have also shown to increase susceptibility to PCa, e.g. *BRCA1*, *BRCA2*, *CHEK2* and *BRIP1*[1]. Leongamornlert et al., discovered frequent germline mutations in DNA repair genes that were associated with familial PCa as well as a more aggressive phenotype; the cancers were more likely to have nodule involvement, metastasis and be stage 4[1].

Genome wide association studies (GWAS) identified 76 susceptibility loci associated with PCa risk largely within the European population[35]. These occur commonly but with low penetrance and act multiplicatively to substantially increase risk. GWAS are where genetic variants across whole genomes of different individuals are examined to identify if any variants are associated with specific traits. Investigation of >10 million

SNPs in a more diverse ancestry population (European, African, Japanese and Latino) in ~43,000 PCa cases and ~43,000 controls revealed 23 novel susceptibility loci[36]. Combining these 23 novel variants with already known variants, we can now explain 33% of the familial risk of PCa in populations of European ancestry. The per allele effects of the 23 variants ranged from 1.06-1.14 and were consistent with log-additive effects of the 23 variants, 15 were exclusive to the European ancestry population, 7 were multi-ethnic, 17 were associated with earlier onset (<55 years compared to >55 years) and 1 was associated with disease severity[37].

1.3.3 Current clinical practice for the diagnosis of PCa

The current clinical process uses a risk-adjusted patient specific method[2] that aims to improve control of the cancer whilst reducing risk of complications from treatment. The initial step is for a PSA blood test (section 1.4.1.2) to be performed at a GP after a patient has shown symptoms or has other factors increasing their risk such as family history and/or ethnicity. A PSA test is an antibody-based test that measures the concentration of the prostate specific antigen (PSA) in the peripheral blood. A digital rectal examination (DRE) is then performed by a clinician, during which they feel the prostate for any abnormalities. DRE tests have about a 59% overall accuracy[32]. PSA testing is a better predictor of PCa than DRE. In a multicentre trial ($n = 6$) with a total of 6,630 men, 1,167 underwent TRUS biopsies due to PSA>4ng/ml or suspicious DRE result. PSA detected 82% of tumours, whilst DRE only detected 55%, PSA was significantly superior at detecting PCa ($p = 0.001$, PPV for PSA: 32% and PPV for DRE: 21%)[38]. However, a DRE is useful because it can often detect cancers missed by the other tests; especially those with normal PSA levels[32]. It can also be used to investigate other abnormal prostatic conditions such as BPH.

If the PSA test (section 1.4.1.2) result is above normal but below 100ng/ml, then a transrectal ultrasound-guided (TRUS) biopsy of the prostate is performed. Using an ultrasound probe, sound waves are reflected off of tissues and organs providing a black

and white image of the prostate. The probe and biopsy needle gain access to the prostate via the rectum. At the histopathology department, the collected material is examined for cancerous cells and given a Gleason score. In the case of a PSA of greater than 100ng/ml no TRUS is performed, an advanced diagnosis of metastasis is made usually alongside an MRI and/or Bone scan.

The Gleason scoring system (Figure 0.2) is a histopathology score for staging PCa based on how differentiated the cellular structure is in the prostate. This helps evaluate the patient's prognosis, the higher the score the worse the prognosis. It is obtained by combining the scores of the two most common non-normal patterns of histopathology found in the biopsy. The patterns are scored as such: Grade 1 and grade 2 patterns means the tissue is mostly normal; glands are small, well formed and compactly packed, grade 2 has more intracellular space between. Pattern of grade 3 shows recognisable gland units and darker cells that have began to decrease in size and invade surrounding tissue, the invasion is the most defining feature. Grade 3 is the most common identified, followed by grade 4. A grade 4 pattern has few recognisable gland units with many cells invading surrounding tissue, this can be achieved in many ways resulting in this being the most difficult grade to identify. The fifth grade has no recognisable glands with many cells within the surrounding tissue, there are sheets of cells that lack any nuclear arrangement and a complete loss of gland architecture is observed. In common practice no lower than a 3+3 is seen (giving an overall Gleason grade of 6) and this offers a good prognosis. A Gleason score of 4+3 offers a worse prognosis than that of a 3+4[39].

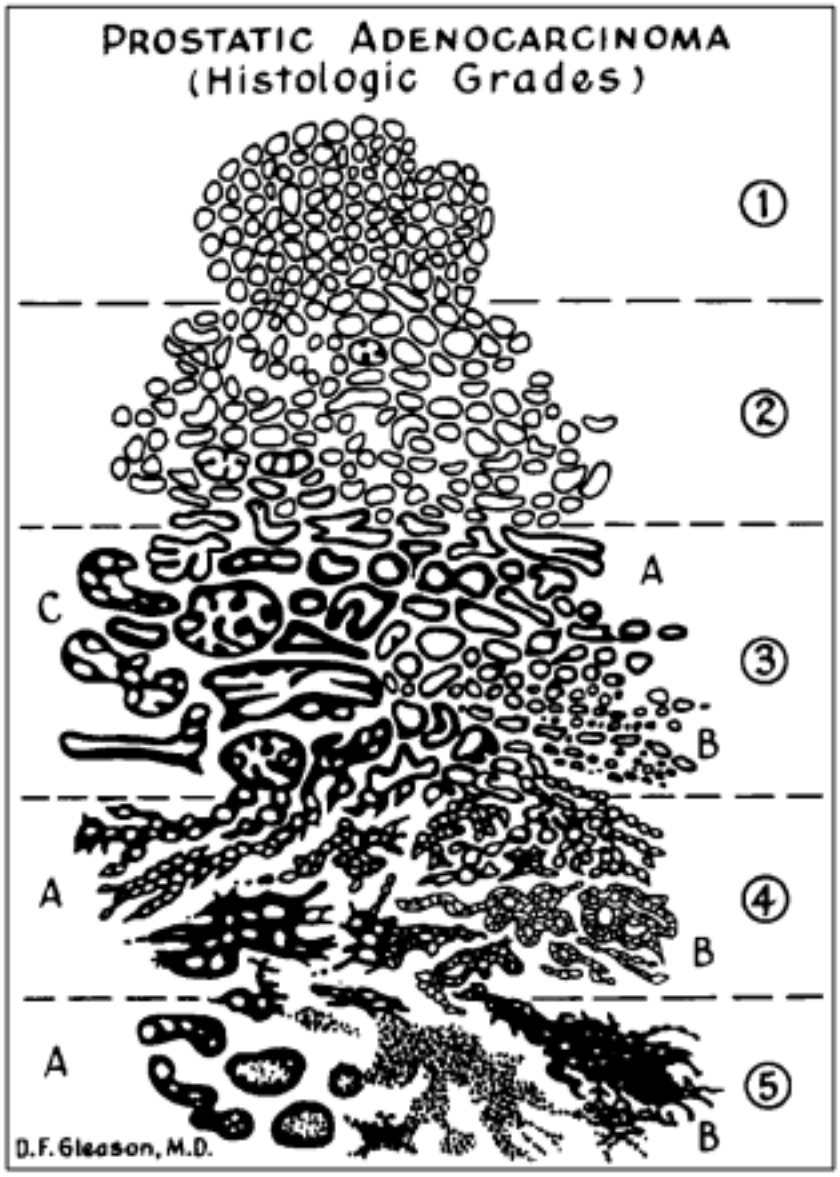


Figure 0.2 The Gleason grading standard drawing. Shows the histopathological pattern of prostate cancers, starting at normal looking prostate cells with normal cellular architecture to fully differentiated PCa cells with no formal cellular architecture. Adopted from Humphrey, P et al., 200439.

Following a negative TRUS biopsy result, if the PSA maintains a high value, a template biopsy can be performed. This differs from the TRUS biopsy as it uses a template or grid over the perineum, which the biopsy needle is entered through to the prostate. However, an ultrasound probe is still used to help guide the needle to the prostate tissue. Generally, more cores are obtained during a template biopsy.

PSA testing lacks specificity and so many men undergo unnecessary TRUS biopsies. TRUS biopsies have risks including serious infection, bleeding, urine retention as well as extra medical costs[40]. Therefore, it is important to identify molecular biomarkers for

PCa detection that are specific and reliable from a non-invasive source such as blood or urine (section 1.5).

1.3.4 Current process for the clinical treatment of PCa

There are many treatment regimes open to patients with PCa. However, there is a lack of specific and accurate biomarker to stratify patients between the different treatments. For many clinical pathways in PCa there is variability in how long the treatment lasts or whether there is any response at all. For example, resistance to hormone therapies (section 1.3.4.2.1) are inevitable but patients will remain responsive for different lengths of time; from no initial response at all to anywhere between 6 months and 10 years. Another example is how long a patient will last on active surveillance (section 0) before requiring treatment. No biomarkers currently exist that are able to detect which patients will have long term response and which patients response will be short lived and therefore, could benefit from receiving a different/ more aggressive treatment more rapidly. This would offer each patient a more effective treatment first time around.

There are many factors taken into consideration when deciding which treatment is best for a specific PCa patient including general health, age, Gleason score, TNM stage, PSA and whether it is metastatic or not. However, there are not any molecular tests currently available.

1.3.4.1 Localised Prostate Cancer

Localised PCa is stratified by their risk of metastasis using the NICE risk categories (Table 0.1**)** which incorporates PSA level, Gleason score and clinical staging in order to decide treatment style for each patient. Each level of risk is offered a different course of therapy.

Table 0.1: PCa risk stratification table. Proposed risk categorization from NICE Guidelines 175[41]

Level of Risk	*PSA*		*Gleason Score*		*Clinical Stage*
Low risk	***<10 ng/ml***	***and***	***≤6***	***and***	***T1-T2a***
Intermediate risk	***10-20 ng/ml***	***or***	*7*	***or***	***T2b***
High Risk[l]	***≥20ng/ml***	***or***	***8-10***	***or***	***≥T2c***
High-risk localised PCa is also included in the definition of locally advanced PCa.					

1.3.4.1.1 Surgery as a treatment for PCa

Radical Prostatectomy (removal of the whole prostate gland) is a treatment considered for men with T1 or T2 PCa (localised to the prostate gland without spread). Side effects can include urinary incontinence, impotence and loss of fertility. Transurethral resection of the prostate (TURP) is considered for men with benign prostate growth (BPH) and for advanced cancer to alleviate symptoms; the inner area of the prostate (that surrounding the urethra) is removed.

1.3.4.1.2 Radiotherapy

Radiation therapy is the provided course of treatment for low-grade, localised PCas (with similar cure rates as those who receive radical prostatectomy). It can also be provided alongside hormone therapy for cancers that have spread out of the gland to nearby tissues, for recurring tumours (post-surgery), and also to advanced patients to reduce tumour size (offering some relief from symptoms). Side effects can include urinary incontinence, impotence cystitis and radiation proctitis.

1.3.4.1.3 Biochemical Recurrence

Men treated with either radiotherapy or radical prostatectomy (RP) can develop biochemical recurrence (BCR), which is characterised by a state of elevating PSA level post treatment and indicates growing tumour or metastases[42]. Within 10 years, of those patients treated with radiotherapy ~30-50% and ~20-40% of patients post RP will develop BCR[43]. Increase in PSA does not necessarily mean imminent death or threat

and can often be treated with hormone therapy. There has been much research in treatment options for these patients, which includes when to administer hormone treatment as well as non-hormonal alternatives including targeted agents and immunotherapies[43] due to the morbidities associated with hormone treatment.

UHRF1 expression in tissue samples has been identified as a potential biomarker for predicting BCR post RP. UHRF1 expression negatively correlates with mean months of BCR-free survival ($p < 0.001$). However, UHRF1 expression was less significant than pre-operation PSA levels and Gleason score[44]. Other studies have identified biomarkers that are linked to BCR; Prx6 (an oxidative stress marker) expression is associated with shortened biochemical recurrence free survival and overall survival in 240 post RP patients ($p = 0.02$ and $p = 0.033$, respectively)[45]. PTEN deletion has been associated with an increased risk of BCR ($p < 0.01$, HR: 3.58)[46]. Metallotheionein-2A (MT-2A), E-cadherin, and cyclin-E were investigated for BCR association by microarray immunostaining. Positive MT-2A and cyclin E expression along with negative E-cadherin expression showed a decrease in biochemical recurrence-free survival ($p = 0.009$ (HR = 2.15, 95% CI = 1.14 - 3.08), $p = 0.037$ (HR = 1.45, 95% CI = 1.02 – 1.92), and $p = 0.047$ (HR = 1.31, 95% CI = 1.03 – 2.21), respectively)[47]. In a multivariate analysis all three were deemed to independently predict BCR[47]. Still, the promise of these biomarkers have not been translated into use in the clinic.

Other clinical features such as tumour volume and percentage tumour volume have also been reported to predict BCR post RP in a meta-analysis of multicentre data ($p = 0.03$, HR: 1.04 and $p = 0.02$, HR: 1.01, respectively)[48].

Active Surveillance, Watchful Waiting and PSA monitoring

To attempt to reduce the number of over-treated patients, programs like active surveillance, watchful waiting and PSA monitoring have been implemented.

A high proportion of PCa are localised and non-aggressive and are unlikely to cause any problems in the patient at all, whereas others progress into more problematic cancers that require more aggressive treatments. Active surveillance is offered to

patients with low-risk localised PCa whom are suitable for radical prostatectomy or radiotherapy as treatment[40]. They monitor the patients looking for indications that their less aggressive cancers are becoming more aggressive problematic cancers. Active surveillance is a close monitoring of the patients and usually involves frequent tests, such as PSA blood tests, DREs, ultrasounds and biopsies.

Watchful waiting is offered to asymptomatic PCa patients for whom there is no curative treatment options or intent. Watchful waiting however is implemented with more aggressive cancers, where treatment would cause problems due to the patients' age or general health. These patients are monitored for disease progression (a rapidly rising PSA or bone pain). Compared to active surveillance, less frequent tests and more reliance on patient symptoms for indication of change are implemented in watchful waiting.

PSA monitoring exists to identify patients who have continual raised PSA in the "grey zone" (PSA between 4 and 10ng/ml) rather than just an intermittently raised PSA on one test. Patients can receive multiple PSA tests to monitor them prior to biopsy. This can help to eliminate the number of unnecessary biopsies if there is a continual PSA>4ng/ml then it is more likely to be due to PCa and thus these patients require biopsies.

1.3.4.2 Metastatic Prostate Cancer

Metastatic PCa is detected in 21% of men at their time of diagnosis[29]. It is usually identified by a PSA>100ng/ml[49] and/or a positive bone scan. Those with metastasis are primarily prescribed hormone therapy agents that block androgen signaling[50].

1.3.4.2.1 Hormone Therapy

Androgens are male hormones, which include testosterone and dihydrotestosterone (DHT), and aid in the signalling for prostate cell growth. Androgen deprivation therapy (ADT) lowers levels of these androgens and/or prevents them from reaching the

prostate cells, resulting in shrinking and slower growth of the cancer. ADT is not a cure but can prolong life.

Luteinizing hormone-releasing hormone (LHRH) analogs and antagonists reduce levels of testosterone released from the testicles by blocking the feedback loop to the hypothalamus. Anti-androgens bind androgen receptors, preventing cell growth signalling, though these are usually added to LHRH treatments when patients begin to become resistant. However, it is a controversial question of when anti-androgens should be added to LHRH treatment to gain full androgen blockage, it is thought in some cases initial hormone therapy should include both LHRH treatments and anti-androgens[51].

Patients receiving ADT develop resistance leading to castration resistant PCa (CRPC), with a median survival of 1-2 years[52]. It is likely that the high level of heterogeneity within the prostate tumour contributes to this resistance[53]. CRPC develops when cells become hypersensitive to the residual levels of testosterone that are left during chemical castration. Castration does not remove all testosterone; the maintenance of intratumoral androgens is due (at least partly) to the intratumoral or intracrine biosynthesis of steroid hormones (adrenal androgens) or potentially de novo steroidogenesis, from cholesterol or progesterone precursors within the tumour[54]. Hypersensitivity to these residual levels of testosterone are believed to be due to androgen receptor (*AR*)- mutations that alter ligand binding, alterations in *AR* co-regulators or *AR* over-expression (considered to be the main driver of CRPC progression)[54]. *AR* over-expression has also shown to convert anti-androgen treatments (like bicalutamide, flutamide and enzalutamide) from *AR* antagonists to *AR* agonists[55,56].

Abiraterone was the first drug in clinical practice to target the production of androgens by the tumour. It irreversibly and selectively inhibits CYP17A activity. CYP17A is a critical enzyme; it facilitates the hydroxylase and lyase activity required in the production of adrenal androgens, DHEA and androstenedione (AED), from cholesterol[54]. Although, abiraterone has had impressive responses in clinical trials, not all men respond and resistance occurs (seen by a rising PSA), the mechanisms for

which are currently unknown. Abiraterones place in the treatment of PCa is so far undetermined and many clinical trials are in place to investigate this.

1.3.4.2.2 Castrate Resistant Prostate Cancer (CRPC)

Once PCa becomes castrate resistant, there are other treatment options available such chemotherapy and vaccine therapy.

Chemotherapies are usually given to PCa patients who have metastasis but are not, or no longer responding to hormone therapies. It is generally not given to patients with early PCa, although studies are currently investigating its use following surgery. The first chemotherapy agent of choice for PCa is Docetaxel (administered alongside the steroid prednisone) and if this doesn't work or stops working, Cabazitaxel is often a second drug choice[57]. Chemotherapy is used again with the focus on increasing life expectancy and/or quality of life for PCa patients (by slowing the growth of the cancer) but is considered unlikely to result in a cure to the disease.

1.4 Known PCa Biomarkers

Biomarkers in PCa fall into different categories: Biomarkers to predict the presence of PCa (screening and diagnosis), biomarkers to stratify patients (into those requiring active surveillance and those requiring more radical treatments), biomarkers for identifying those whom can be treated with biological targeted therapies and predisposition biomarkers for those who are more likely to develop PCa in their lifetime.

1.4.1 Prostate Specific Antigen (PSA)

PSA, a kallikrein like serine protease (coded for by the gene *KLK3*), is a molecular biomarker currently and routinely used for the diagnosis of PCa, as well as roles in prognosis and treatment response. In normal prostate glands, PSA is highly compartmentalized and found at levels 1 million times fold higher within the prostate compared to that in blood serum. However, in prostatic disease it is thought that this

compartmentalization is disrupted resulting in increased levels of escaped circulating PSA[58].

PSA is prostate specific but not cancer specific; elevated serum PSA can be the result of benign prostatic hyperplasia (BPH), chronic inflammation, and infection. Normal and diseased prostatic epithelial cells produce PSA, therefore, weakening its specificity as a cancer biomarker.

Research into men with a PSA less than 4ng/ml has shown that there are many men with low PSA (0.6-1ng.ml) that have PCa (10.1%) and even high-grade (Gleason 7+) PCa (10%)[59]. Evidence suggested there was no PSA threshold for which a man can be assured he has no risk of PCa, but men with <0.5ng/ml PSA do have a decreased risk of developing PCa. Risk of PCa in men with PSA <0.5ng/ml was 6.6%, this increased to 26.9% in men with PSA 3.1-4ng/ml[59]. PSA level effect on the risk of PCa was significant, $p<0.001$ (odds ratio 1.66 per unit increase in PSA, 95% CI 1.50 – 1.85)[59].

PSA levels are affected by both age and race; when deciding on a reference range for diagnosis and deciding which men will undergo TRUS biopsies, it is important to consider these factors. A study on 77,700 men showed that not only does the PSA level rise but also that the range increases with increased age (ages 40-49; mean PSA: 0.83, SD: 0.79, ages 50-59; mean PSA: 1.23, SD: 1.33, ages 60-69; mean PSA: 1.83, SD: 1.94, and ages 70-79; mean PSA: 2.31, SD: 2.35). The differences between the age groups and their variances were significant, $p < 0.0001$ and $p = 0.0001$, respectively[60]. Significant differences in PSA levels were observed between different races also; pairwise differences were seen between white and black people, white and Latino people, black and Asian people, and Asian and Latino people ($p <0.0001$). Black people have the highest mean PSA values in each age cohort[60].

1.4.1.1 PSA - Screening

Due to the lack of specificity that PSA holds, using it for screening purposes has led to over diagnosis and over-treatment as well as downgrading and down staging at

diagnosis and fewer PCa related deaths[61]. A cohort of men diagnosed with PCa, have a form of the cancer that grows so slowly that it is unlikely to pose a threat to the patient. Treating of these cancers is known as over-treatment. PSA's lack of specificity for PCa means it is not recommended for a screening biomarker due to the over-detection and overtreatment costs it would lead to[61].

The National Cancer Institute estimate that screening 1,000 men between 55 and 69 every 1-4 years would result in 100-120 men getting a false positive diagnosis (Figure 0.3). False positive diagnoses lead to anxiety and stress for the patient and his family, as well as extra medical costs in further diagnostic procedures. Procedures include TRUS biopsies, which also add further risk to patients; serious infections are not uncommon. Of the 1,000 men screened, and the 110 patients to receive a true positive result, it is estimated that only 1 man would be saved due to screening, compared to the 4-5 men that would die without screening[62].

1.4.1.2 PSA – Diagnosis

Similarly to its use in screening, PSA makes a weak diagnostic biomarker due to its lack of specificity to cancer. However, it is the current first diagnostic test for PCa. The sensitivity and predictive value of PSA as a biomarker for PCa decreases greatly for patients in the "grey zone". PSA levels in the approximate range 2-10ng/ml is known as the "grey zone" as it is difficult to distinguish which elevations are due to cancer and which are associated to other factors including age and BMI, or due to conditions such as benign prostatic hyperplasia (BPH). Investigations into the PSA grey zone generally use cutoffs between 2/4 to 10ng/ml to define it. For every 5 patients, whose PSA level resides between 2.5-10ng/ml, 4 will have a negative biopsy result, and the predictive value of PSA in the grey zone drops from >90% to <25%[63].

As an individual variable, PSA is a much better PCa predictor than a digital rectal examination (DRE) or transrectal ultrasound[61], but its modest diagnostic accuracy has led to other PSA forms being investigated.

1.4.1.3 Free PSA and Pro-PSA

To improve abilities in distinguishing BPH from PCa in patients who fall in the "grey zone", investigations into the percent free PSA (or ratio of free to complex PSA) and its most significant cut-off for biopsy, and different isoforms of pro-PSA were performed.

Antibodies were developed that could distinguish between and measure the amounts of tPSA and fPSA, a higher ratio of fPSA:tPSA correlates with a lower risk of PCa. This comparison allowed a small yet significant improvement in the ability of PSA to distinguish PCa from BPH (and other benign diseases that raise PSA levels)[64].

A study of 773 men with PSA levels between 4-10ng/ml with confirmed histological diagnosis (379 with PCa and 394 with BPH) resulted with a suggested 25% free PSA cut-off. The 25% free PSA cut-off was able to detect 95% of patients with PCa and was also able to avoid 20% of unnecessary biopsies[64].

PSA is secreted as the inactive enzyme pro-PSA, this can be cleaved at different locations resulting in the mature/active form of PSA. Some remain uncleaved and pro-PSA can have many isoforms. The [-2]proPSA consistently correlates with PCa[65]; it is observed in greater abundance if the prostate is neoplastic (25-95% of free PSA compared to only 6-19% in men without PCa[66]).

Guazzoni et al., showed that the use of %[-2]pro-PSA alone was better at discriminating between PCa and BPH (in patients with PSA ranges 2—10ng/ml) compared to that of total PSA and percentage free PSA, with AUCs of 75.7%, 53%, and 58%, respectively[67]. Using an artificial neural network, Stephan et al., showed that the combination of %[-2]proPSA, %free-PSA, total PSA and age (but not prostate volume) offered highest accuracy (AUC 0.85). It was also shown that %[-2]proPSA was better at discriminating between T2 and T3 tumours as well as Gleason <7 and Gleason >7 cancers[68].

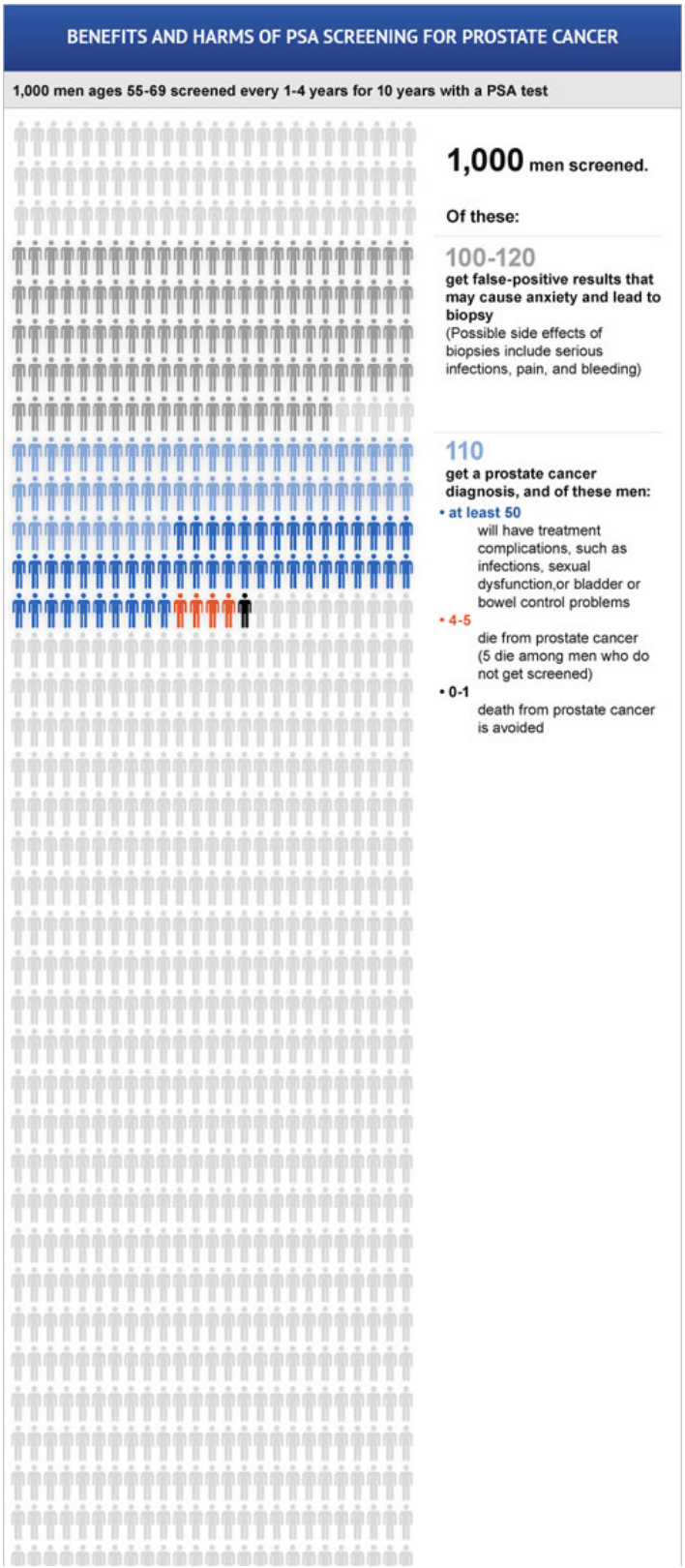


Figure 0.3 The NCI website breaks down the results of PSA screening of 1,000 men between the ages of 55-69. Taken from the National Cancer Institute 2015[69].

1.4.1.4 PSA – Treatment

PSA is commonly used within treatment plans available for PCa, it is a good indicator of progression and drug resistance. PSA levels are routinely and frequently checked in PCa patients; looking for progression in AS patients (section 0), resistance in HT patients (section 1.3.4.2.1) and BCR in radiotherapy or post-radical prostatectomy patients (section 1.3.4.1.3).

PSA is one of the key factors in determining treatment options for patients. A PSA above 100 is indication of metastasis and so hormone therapy is usually provided. PSA also is involved in determining treatment of lower grade localised PCa (Table 0.1).

Investigations into [-2]pro-PSA combined with percentage fPSA identified a correlation for the need of more radical treatment rather than active surveillance[70]. Also, other proPSA isoforms ([-5] and [-7]pro-PSA) correlate with a need for more radical treatments in active surveillance patients, when found in the tissue surrounding the tumour in biopsies.

1.4.1.5 Concluding PSA

PSA is not a specific PCa biomarker, yet it is the first clinical diagnostic test given to patients and is also a determining factor in treatment options and changes. PSA remains a very useful biomarker in following patients with PCa to look for resistance to treatment, further progression and recurrence. Though other biomarkers are unlikely to replace PSA, they are required to improve the sensitivity and specificity of PSA as a PCa biomarker.

1.4.2 PCA3

PCa gene 3 (*PCA3*) is a PCa specific long noncoding RNA (lincRNA), also known as DD3 on chromosome 9q21-22 that is over-expressed in PCa tissue[71]. *PCA3* is not expressed in normal prostate tissue and expression is seen at low rates for hyperplastic prostate tissue, making it the most specific PCa biomarker identified so far. The non-

coding *PCA3* mRNA functions as a polyadenylated RNA transcript, which does not result in a cytoplasmic protein[63].

1.4.2.1 PCA3 – Diagnosis

PCA3 can be found in urine, but only at sufficient levels, after a DRE is performed[72], and that comparing the ratio of *PCA3* mRNA quantities with *KLK3* mRNA (which is the transcript for PSA) quantities (very slightly over-expressed in prostate cells in urine) gave high sensitivity and specificity rates, 67% and 83% respectively[73]. The comparison of *PCA3* and *KLK3* mRNA quantities found in prostate cells in urine is known as the *PCA3* score. An assay was generated to simultaneously detect *PCA3* mRNA as well as *KLK3* mRNA in urine: the uPM3^{TM} assay. The assay was tested on 158 patients with elevated PSA and/or an abnormal DRE, whom provided a sample with a sufficient amount of prostate cells in the urine. The assay identified PCa in 62 of the 158 patients (39%), with sensitivity and specificity rates of 82% and 76%, respectively. The positive and negative predictive values for the assay were 67% and 87%, respectively. Comparably, PSA had sensitivity and specificity rates of 98% and 5%, with positive and negative predictive values of 40% and 83%[63].

The performance of the uPM3^{TM} assay at different PSA levels (<4ng/ml, 4-10ng/ml and >10ng/ml) was examined, with outcome sensitivity levels of 73%, 84% and 84%, respectively, and specificity levels of 61%, 80, and 70%, respectively[63]. A more stable re-designed assay was later designed and evaluated in a multicenter assessment: The assay had between 94%-100% discriminatory rates in samples after a DRE with at least 3 strokes[72]. This test was then applied to 72 men with known biopsy outcomes, of which 17 were positive for and 55 were negative for PCa, at two centers. Taking the *PCA3* score as a continuous variable, a ROC analysis was performed and both sites were able to correctly classify 49/72 (68.1%) of patients, and the AUC were not significantly different (p = 0.9289), this demonstrates significant accuracy between the

sites ($p = 0.0085$)[72], highlighting the *PCA3* assay as an accurate, reproducible test for the diagnosis of PCa.

Another multicentre saw improvement of PCA3 on PSA in the "grey zone" (PSA 3-15ng/ml); AUC increased from 0.57 to 0.66 and specificity increased from 47% to 66% for PSA and *PCA3*, respectively[74]. A study looking at multi-gene expression profiling of prostatectomy tissues yielded and AUC for *PCA3* of 0.85 individually but increased with the addition of *EZH2*, prostein and *TRPM8* to 0.90[75].

1.4.2.2 Repeat Biopsies

The *PCA3* test is effective at identifying patients who were likely to have a positive second biopsy result, after receiving a negative first. A multicentre clinical study of 466 men evaluated the clinical usefulness of the *PCA3* assay for the prediction of repeat biopsy outcome. The study resulted in a suggested *PCA3* cutoff of 25, with patients with a *PCA3* score lower than 25 were 4.56 times as likely to have a second negative result for their repeat biopsy[76]. The PCA3 test is FDA approved but generally only used in private healthcare in the UK.

1.4.2.3 PCA3 Conclusions

Although the *PCA3* assay shows significant improvements in specificity and sensitivity compared to PSA, it is significantly more expensive: A *PCA3* test costs between approximately £300 and £400, (whereas a PSA test costs approximately £7) and this cost will increase with the use of gene panels. In comparison a TRUS biopsy costs £312[50], as you can see the *PCA3* test can be more expensive than just doing the repeat biopsy. The literature and improved sensitivity show that the *PCA3* test is clearly useful but where it fits into PCa diagnostics is unclear at this time. The *PCA3* test is currently available privately but not on the public health care system/NHS in the UK.

1.4.3 AMACR

Alpha-methylacyl-CoA racemase (*AMACR*) is used as an immunohistochemical (section 1.6.7) diagnostic biomarker for PCa. Needle biopsy specimens are stained for *AMACR* during diagnosis of PCa patients[77], as *AMACR* expression is increased in PCa but may decrease with progression[78]. *AMACR* expression alone was not informative for the prediction of metastatic or lethal PCa; age, Gleason score and stage were also indicative[78] and out of 64 prostate adenocarcinomas no significant correlation was seen between *AMACR* expression levels and histopathological grade[79].

AMACR is an enzyme that regulates the metabolism of branched-chain lipids and drugs and is often overexpressed in PCa tissues[80,81]. It is thought that the synthesis of fatty acids and increased use of branched chain fatty acids plays a role in PCa progression. It is essential for optimal growth of PCa cells in vitro and offers a potential treatment target complementary to hormone therapy. *AMACR* is also frequently seen in tumours of patients with hereditary links to PCa[79].

1.4.4 AR

The Androgen receptor (*AR*) binds androgens leading to the development and survival of prostate epithelial cells. In PCa it allows survival and growth of the tumour and is a known contributor to its progression. Whilst PCas show great heterogeneity, it is obvious that *AR* plays an important role in the survival of the bulk of prostate tumour cells[82].

Hormone therapies work by blocking androgen-*AR* signalling, inhibiting growth and survival of the tumour. *AR* transcriptional reactivation/rearrangements are fundamental to the inevitable resistance of PCa to hormone therapies and androgen-independent activation of the *AR* pathway. One resistance mechanism is the production of *AR* variants that lack the canonical ligand-binding domain[82], allowing the transcription of *AR* target genes without the initiating signal of androgen binding. 17 of these *AR*

variants have been identified, all containing a common core of the DNA binding domain and then NH2 terminal domain and lacking the ligand-binding domain. There are several mechanisms for the production of *AR* variants including: proteolytic cleavage, genomic alterations, and altered exon splicing.

Levels of specific *AR* variants observed in clinical samples are highly variable and not all variants are equivalent at predicting progression and resistance. As well, clinical studies using *AR* variants are limited by the lack of clinically validated assays for the detection of the individual variants. However, these limitations are currently being addressed[82], suggesting potential clinical use of *AR* variants as biomarkers.

1.4.5 SPOP

SPOP, otherwise known as E3 ubiquitin ligase adaptor speckle-type poxvirus and zinc finger (POZ) domain protein, interacts directly with and regulates SRC-3 (p160 steroid receptor coactivator-3). The p160 SRCs play fundamental roles in the cell proliferations and *AR* transcriptional activity as well as resistance to androgen deprivation therapy[83]. SPOP binds wild-type AR leading to its degradation; this is promoted by anti-androgens but antagonized by androgens. Whereas, *SPOP* mutants and *AR* alternative splicing leads to *AR* stabilization suggesting a key role in acquiring ADT resistance[84].

A new molecular subtype of PCa can be defined by mutations in *SPOP*; *SPOP* mutations are found in PCas that lack ETS family rearrangements[85,86]. *SPOP* missense mutations within the substrate-binding cleft were identified in 13% PCas and were the most common mutations in 111 prostate tumours that underwent exosome sequencing[87]. This substrate-binding cleft harbours many residues that can be mutated in PCas (Figure 0.4*B)*. The cleft central F133 is the most common site of mutations (Figure 0.4*A*).

Exome sequencing of 50 lethal heavily pre-treated CRPCs and 11 treatment naïve high-grade localized PCas', showed that four CRPCs had *SPOP* oncogene mutations; 2 point mutations, 1 frame-preserving indel and 1 copy-number call increase[88]. *SPOP* mutations correlate with somatic deletions at chromosome 5q21 and 6q21. *CHD1*, *FOXO3* and

PRDM1 are found at these chromosomal regions and are also correlated with *SPOP* mutated PCas[89]. As well as *TMPRSS2:ERG* fusions, *SPOP* does not appear to be mutated in cancers with *Tp53*, *PTEN* and *PIK3CA* mutations[87].

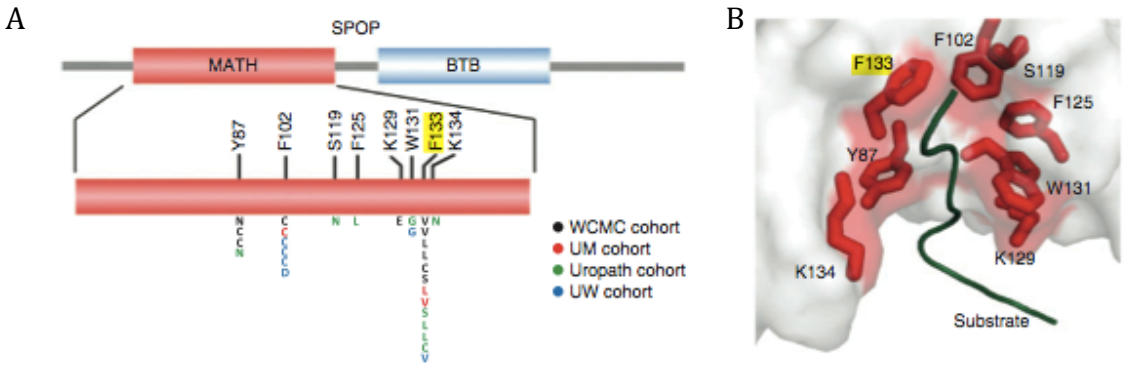


Figure 0.4: SPOP frequency of substitutions and substrate binding cleft[87]. A) the frequency of substitutions in SPOP across four PCa cohorts from Weill Cornell Medical College (WCMC), University of Michigan (UM), Uropath and University of Washington (UW). B) the substrate-binding cleft of SPOP with the positions of all eight residues that can be possibly mutated. Adopted from Barbieri, C. E. et al. 2012[87]

SPOP associations with *AR* highlight the need for examining *SPOP* mutation frequencies in men whom do not initially respond to, or very quickly acquire resistance to PCa; *SPOP* mutation detection could potentially be used to stratify patients out of hormone therapy as a treatment.

1.4.6 TMPRSS2:ERG

TMPRSS2:ERG is a fusion gene that is formed as a result of structural chromosomal rearrangements. *TMPRSS2* is an androgen responsive, prostate specific gene and *ERG* is a transcription factor oncogene belonging to the *ETS* family, both located on chromosome 21. *ETS* family genes are involved in proliferation, differentiation, angiogenesis, inflammation and apoptosis. The fusion occurs via a translocation of sequences that can involve deletion of the intervening sequences between *TMPRSS2* and *ERG*[90].

ERG has been identified in fusion genes in other cancers; leukaemia and Ewing's sarcoma. *ERG* knockdown inhibits cell growth and invasion and oppositely over-expression leads to invasion and the induction of PCa like lesions on *in vivo* models.

ERG has also been identified to work with mutated members of the *PI3K* pathways leading to the progression of PCa in animal models.

TMPRSS2:ERG fusions are seen in ~50% of PCas[91]. *TMPRSS2* also fuses with other members of the *ETS* family (*ETV1*, *ETV4* and *ETV5*) in PCas but at much lower frequency (Figure 0.5). Diversity is also observed in the splice variants of *TMPRSS2:ERG* (Figure 0.5) not only between PCas but also within an individual PCa. The most commonly identified *TMPRSS2:ERG* fusion is *TMPRSS* exon 1 fused with *ERG* exon 4, this is described as T1/E4, the second most commonly found is T1/E5[92].

It remains controversial for if *TMPRSS2:ERG* fusions are implicated in a poor clinical outcome. A number of studies now suggest it is not the major factor of clinical outcome, but that in a combination of copy number gain and other genetic aberrations (like *PTEN* loss) it can offer prognostic information[92]. Yet many papers still suggest that *TMPRSS2:ERG* fusions are implicated in mediating advanced PCas[93]. However, it has also been shown that early cancers and HG-PIN can also harbour *TMPRSS2:ERG* fusions.

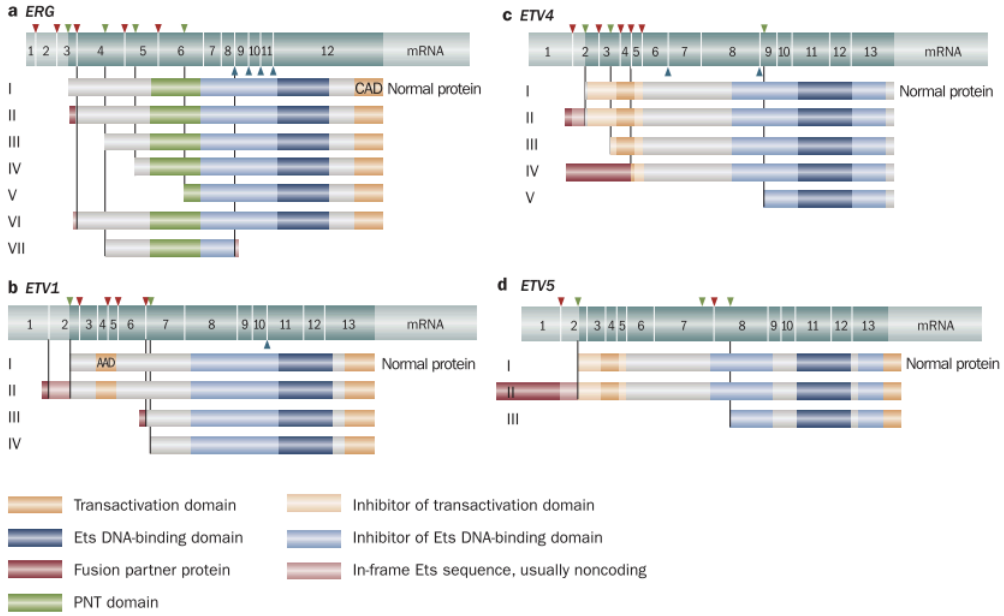


Figure 0.5 *ETS* family partners for *TMPRSS2* fusion and their splice variant diversity. Adopted from Clark, J *et al*., 2009[92].

1.4.6.1 TMPRSS2:ERG as a Therapeutic Target

During the 1990's, a leukaemia fusion; BCR-Abl (the Philadelphia chromosome) emerged as a target of treatment (Imatinib) in Philadelphia chromosome positive (Ph+) myeloid leukaemia[94]. *TMPRSS2:ERG* has a prevalence of approximately 50% and is one of the commonest of all cancer fusion genes in solid tumours, making it a good potential therapeutic target. However, studies have shown that *TMPRRS2:ERG* does not increase cellular proliferation or anchorage-independent growth, but instead induces a transcriptional program associated with invasion[95]. Knockdown of ERG transcriptional programming in ETS-positive cancers lead to an inhibition of invasion in the VCaP cell line. Direct over expression of *ERG* in both VCaP and benign prostate cells mediate cellular invasion through engagement with plasminogen activation pathway components, potentially showing a downstream target that could be used as a drug target[96]. *TMPRSS2:ERG* fusions have also been implicated in signalling pathways and ion channel genes creating further opportunities for therapeutic targeting of these fusion positive cancers[92].

Shao *et al*., have shown that targeting the most common and clinically significant alternatively spliced isoforms of the *TMPRSS2:ERG* fusion using siRNAs delivered by liposomal nanovectors resulted in the inhibition of tumour growth *in vivo*[97]. The mice with orthotopic or subcutaneous xenograft tumours (with the target fusions) also showed no sign of toxicity. Therefore, *TMPRSS2:ERG* targeting could be a potential future therapy for PCa.

1.4.7 Biomarkers for pre-disposition to PCa

Family history has been significantly associated with a higher risk of PCa ($p = 0.01$, odds ratio, 1.39; 95 percent confidence interval, 1.07 to 1.79;) in a study of 2,950 men, all with an initial PSA of less than 4ng/ml. Of the 2,950 men, 477 were family history positive and 2,473 were family history negative. After a seven-year follow up, 449 men were diagnosed with PCa; 94/477 (19.7%) that were family history positive and

355/2,473 (14.4%) that were family history negative[59]. Family history in a first-degree relative (brother, father, or son) is said to double a man's risk of developing PCa, with increasing risk as the number of affected relatives rises[98].

BRCA2 mutations increase relative risk by 5-23 fold in men above 60 years of age, however, the frequency of *BRCA2* mutations is low and can only account for a small number of PCa susceptibility cases[99]. *BRCA2* mutation carriers are in higher risk of developing PCa than *BRCA1* mutation carriers and studies into *BRCA1* mutations suggest they have limited contribution to PCa risk[100]. Breast cancer linkage consortium studies (BCLC) found that *BRCA2* carriers risk was also based largely on age and the mutation location[100].

Genome-wide association studies (GWAS) have led to the identification of more than 46 single nucleotide polymorphisms that have low penetrance in PCa[99]. As discussed by Goh et al., these include SNPs at loci or close to loci known to be involved in PCa such as *KLK3*, *AR*, and *AR* transporter genes[99]. Low penetrance genes were investigated because evidence has suggested that the risk of developing PCa is likely related to a combination of loci conferring low to moderate risk of the disease and, not so commonly, alleles with higher risk such as *BRCA2*[99].

As targeted therapies and screening for PCa becomes more widely used, the use for pre-disposition biomarkers will become increasingly important[1].

1.5 Urine and Exosomes

The PCA3 test (section 1.4.2), as previously discussed, proves that urine contains PCa specific biomarkers. The anatomy and location of the prostate make urine a viable source of prostate biomarkers; urine from the bladder passes through the middle of the prostate, where secretions from the prostate glands can enter the urine (Figure 0.6). DRE manipulates a more abundant release from these glands allowing prostate and PCa specific markers to be detectable in urine (such as *PCA3*, *KLK3* and *TMPRSS2:ERG*)[72].

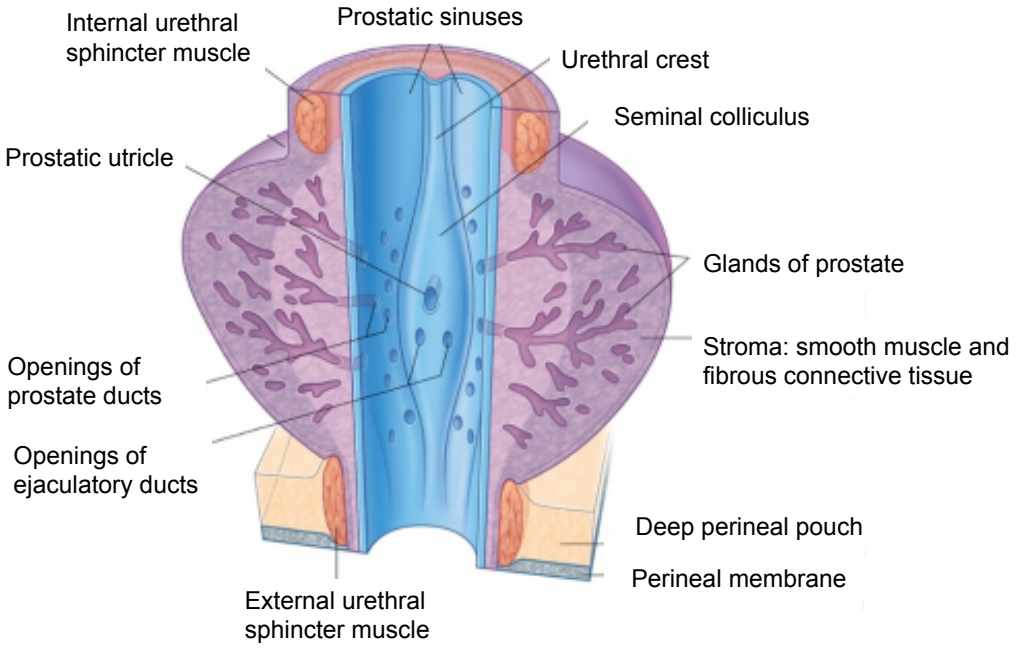


Figure 0.6 Anatomy of the prostate. Adapted from Drake *et al*., 2015[101].

Urine holds an advantage over tissue biopsies in that it potentially allows an overview of all foci of cancer in one go. More than ~80% of cancerous prostates have more than one tumour focus[102], and each cancer focus will have a number of variant tumour clones with divergent genetic and epigenetic changes. Biopsy sampling is incapable of capturing the diversity of cancer within a prostate.

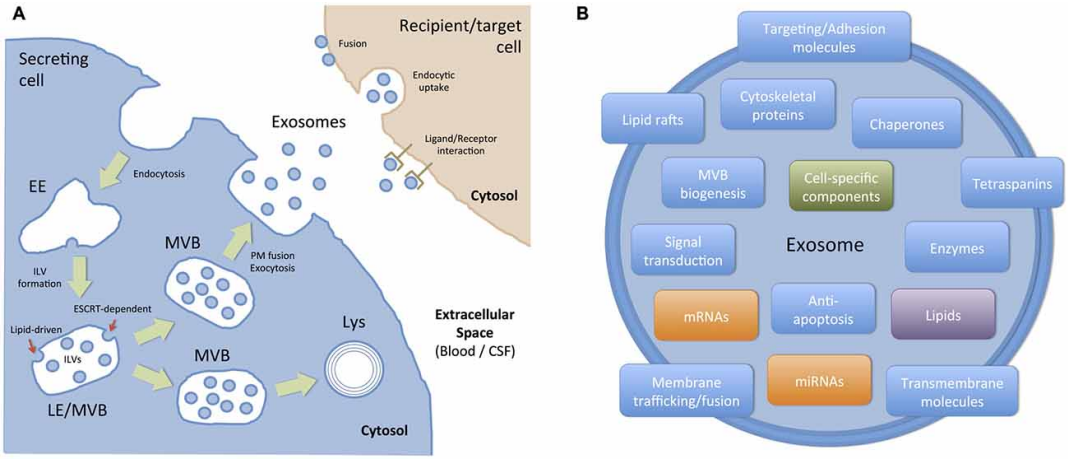


Figure 0.7 Tumour cells send signals to distant cells through exosomes. A) Production of exosomes and how they can be sent to recipient cells. B) The different materials that can be found inside exosomes. Adopted from Bátiz, L.F., 2016[103].

Exosomes are endocytic membrane derived microvesicles 30-120nm in size. They can be found in many biological fluids including those that are easily attainable like blood and urine, which also see elevated exosome secretions during malignancy[104]. Exosomes are a key component of biological trafficking across membranes and play a key role in cell homeostasis. In cancers, aberrant exportation of proteins and RNAs via exosomes can lead to miss-expression in cells that take up the exosome. Exosomes contain proteins, lipids and nucleic acids that can be involved in cell-to-cell communication (Figure 0.7), through their release into surrounding cells. Exosomes derived from tumour cells have roles involved in tumourigenesis, metastasis, and response to therapy by transferring mRNA, miRNA and proteins between cancer cells and the tumour microenvironment[105]. Also ligand binding can trigger a signalling cascade in the target cell. Exosomes have the ability to cross talk/influence key tumour-related pathways (such as those involved in the hallmarks of cancer[106]) including hypoxia driven EMT, evading immune responses, angiogenesis and metastasis[107]. The content of exosomes (miRNA, proteins and mRNA) have been shown to cause changes in a) neighbouring cells, b) the tumour microenvironment and c) in distant cells. "Exosomal shuttle RNA" can be transferred via exosomes from the cell of origin to a recipient cell where it can be translated[107]. Exosomes originating from tumours have been shown to educate non-transformed cells in host tissues to create a pro-metastatic phenotype pre-metastasis.

Hoshino *et al*., showed that treatment of organ-specific cells with lung-tropic model derived exosomes can redirect metastasis of bone-tropic tumour cells[108]. Specific exosomal integrins are associated with organ-specific metastasis and so could be useful in predicting which organs metastasis will occur in. Costa-Silva *et al*., showed that exosomes derived from pancreatic ductal adenocarcinomas was able to create a pre-metastatic niche in livers of naïve mice and also increased the metastatic burden within the liver[109].

Thus, it could be said that looking for biomarkers in exosomes is like raiding cancers' letterbox. The molecular composition of exosomes vary with cell and tissue of origin[107] and can also be altered by pathophysiological changes in the cell of origin, meaning exosomes have great potential for cancer biomarkers.

Some RNAs are enriched within the exosomes at several 100-fold compared to cells, and transcripts that may have very low copy numbers in tumour cells could be detected at much higher relative levels within exosomes[110]. Nilsson *et al*., were able to show that exosomes in urine contained genetic information that is directly from PCa cells[111]. Both *PCA3* and *TMPRSS2:ERG* transcripts were detected in the exosomes. Dijkstra *et al*., showed that the genetic content of exosomes differs from that of the cell sediment[112]. Exosome membranes can resist ribonuclease and DNase digestion of their contents allowing a better-protected RNA inside in comparison to cell RNA. Exosomal RNA will be similar on harvest as when it left the cell, in contrast to cellular RNA which will be altered on loss of cell:cell contact and entry into the non-life sustaining environment of urine. These points make exosomes a stable, viable, and more promising source of PCa biomarkers than cells harvested from urine.

1.6 Methods in Biomarker Discovery

Over the past two decades extensive investigations have proven that cancer is a heterogeneous disease with diverse genomic aberrations[113]. These genomic aberrations

consist of gains, losses and rearrangements of chromosomal fragments, specific gene mutations and epigenetic alterations including methylation. These can lead to aberrant transcript expression and incorrect protein production at differing levels between disease and benign states.

Many cytogenetic and molecular tests have been developed to detect such aberrations. As technologies advance, more effective, less time consuming and cheaper methods are available for biomarker discovery and their validation (Table 0.2).

Table 0.2: Cost of different technologies available for biomarker discovery.

Technique example	*Number of transcripts*	*Batches of Samples*	*Amplification Required*	*RNA usage*	*~Cost/Sample*
NanoString	***<800***	***12***	***Y***	***20ng***	***£50/sample***
Microarray	***30,000***	***1***	***Y***	***20ng***	***£400/sample***
Sequencing	***All***	***1***	***N***	***100ng****	***£1,000/sample***
qRTPCR	***1+***	***1***	***Y***	***20ng***	***$35/sample***
Targeted Sequencing	***250***	***1***	***N***	***1ng****	***$50/sample***

*RNA not amplified and used directly in technology.

1.6.1 Nanostring

The Nanostring nCounter gene expression system was made available in 2008 and is capable of capturing and counting individual mRNA transcripts. It provides direct count data for each of the target genes via a two-probe system: A capture probe and a reporter probe. Both probes are hybridised to the mRNA, the reporter probe hybridising to sequence adjacent to the capture probe (Figure 0.8*A*). The reporter probes are specifically labelled with a series of fluorescent 'beads' that are unique for each gene. The capture probe is biotinylated and the mRNA/probe combination is captured by binding to a streptavidin coated slide. The DNA on the slide is then subjected to a voltage which stretches out the molecules on the slide (Figure 0.8*B*). The slide is then washed to remove excess probes, and the slide is photographed. The bead codes are counted to give the frequency of each mRNA in the sample (Figure 0.8*C*)[114].

In comparison to microarrays (section 1.6.4), NanoString technologies allow quantification of small amounts of starting materials (100ng), and mRNA levels can be measured without the need for amplification. By allowing the customer to choose specific targets, use of NanoString over array can work out cheaper per sample. Microarrays will provide >34,000 targets and cost ~£400-500, however, if you want a select cohort of genes (maximum 800 per analysis), NanoString can allow a cheaper overall experiment. NanoString is also more specific and has a better dynamic range than microarrays. The reaction is performed in solution and not fixed to a solid surface allowing the reaction to be driven to completion and so boasts higher sensitivity. The Nanostring nCounter system also allows a pure digital readout of transcript counts that claim to have less background noise, and be less ambiguous in downstream analyses than those that use analog signals, like microarrays[115].

A disadvantage is that due to the barcode system it utilises, there is a limited number of probes (capped at 800 for a custom codeset)[116]. Again, like microarrays unknown mutations are not identified via Nanostring, and so for the identification of these, sequencing is still preferred and similarly to microarrays and PCR, the quality of the data is dependent on the quality of the probe.

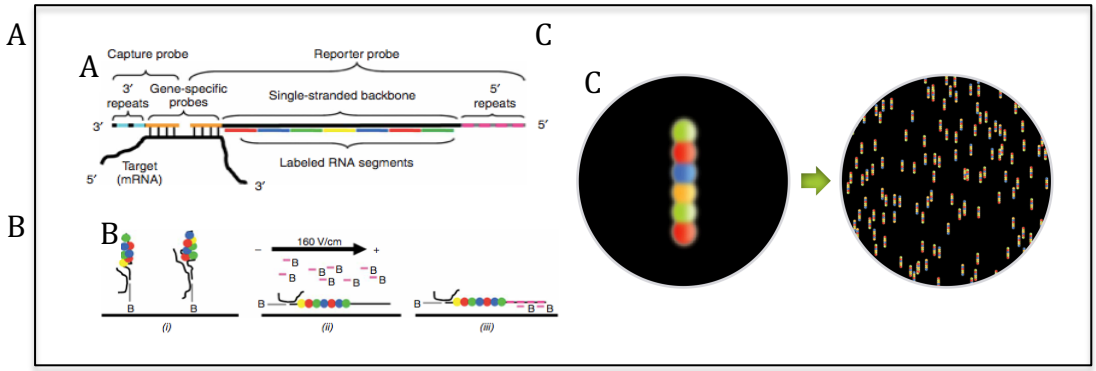


Figure 0.8 NanoString Ncounter system. A) The set up of the two probes (capture and reporter), one target system. B) The elongation and fixing of probes using a current for imaging. C) Imaging of the uniquely labelled reporter probes. Adapted from Geiss, G et al., 2008[115].

1.6.2 Sequencing

In 1977 Frederick Sanger published Sanger sequencing, a method using the incorporation of chain-terminating dideoxynucleotides by DNA polymerase, which cause base-specific termination during DNA synthesis[117]. This was a fundamental breakthrough for science and allowed a monumental accomplishment: the finished grade human genome sequence in 2001. Since then, sequencing technologies have advanced and become considerably cheaper: In 2001 it cost $100 million to sequence a genome and since late 2014 it is ~$1,000. The biggest price drop occurred in 2008 (Figure 0.9) and was a consequence of the introduction of commercialised next-generation sequencing (NGS) technologies. In 2015 the production of Illumina's HiSeq X Ten allowed the first $1,000 sequenced genome[118].

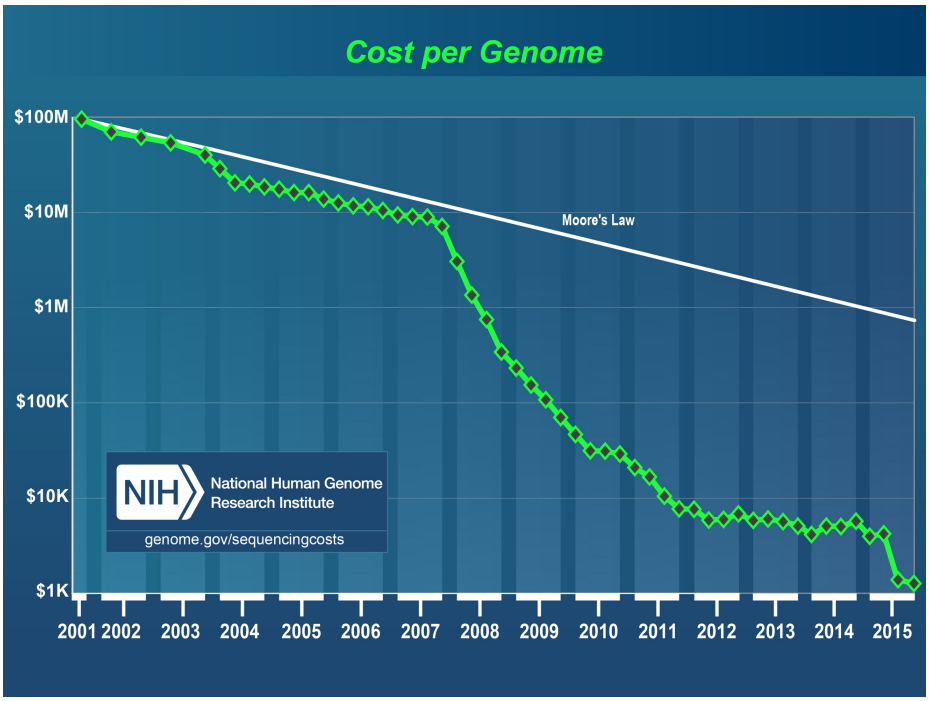


Figure 0.9 Sequencing cost per genome from 2001 to 2015. Sudden drops seen in ~2008 and again in 2015. Adapted from National Human Genome Research Institute (NHI) 2016[119].

1.6.2.1 Next Generation Sequencing (NGS)

Next generation sequencing began with the discovery of the pyrosequencing method using luminescent for measuring pyrophosphate synthesis. This was a two-enzyme

process whereby ATP sulfurylase converts pyrophosphate into ATP. ATP is the substrate for luciferase, which produces a proportional amount of light to the amount of pyrophosphate produced as each nucleotide is washed over template DNA that is fixed to a solid phase. This method is still, similarly to Sanger sequencing, a sequence by synthesis method. Benefits included using natural dNTPs, and being observed in real time without the need for electrophoreses. A disadvantage of this was that identification of more than 4-5 identical nucleotides proved to be difficult. Further improvements in methodology including using beads for DNA attachment and enzymes for degraded unused dNTPs (removing the lengthy wash step), led to the first commercial NGS technology by 454 Life Sciences. This allowed massive parallelisation of sequencing reactions, meaning the amount of DNA sequenced in one run was significantly increased[120].

Following the success of 454's high throughput sequencing machines, a number of new techniques were developed, including the Solexa method of sequencing, which was later acquired by Illumina. The Solexa method used bridge amplification, where DNA molecules were run across complementary oligonucleotides bound to a flowcell. Here, the original flow-cell binding DNA strands arch over to prime the next round of polymerisation for neighbouring oligonucleotides to create clusters of clonal populations by solid phase PCR. This is another example of sequence by synthesis, although here modified dNTPs with a fluorescent 'reversible-terminator' occupies the 3′ hydroxyl position. These fluorophores needs to be cleaved prior to the next polymerisation step, allowing sequencing in a synchronous manner (

Figure 1.10).

Illumina created the first Paired end sequencing, improving efficiency and accuracy when aligning to a reference genome by providing positional information[121,122] and decreased sequencing costs per template[123]. Paired-end sequencing enables improved biological applications, allowing genome-wide identification of gene fusions, insertions, deletions and translocations and spliced exons because it retains information on the distance and relationship between two ends of DNA fragments[121,123].

Illumina's HiSeq series then used a further improved method to allow longer read length and depths. Disadvantages include substitution errors (commonly after "G" incorporation), under-representation of AT-rich and GC-rich regions (due to amplification bias) and a 2.5% false positive rate for novel single nucleotide variants (SNVs)[124]. Illumina is the most commonly used sequencing platform: The HiSeq series is still used commonly for genome sequencing, whilst Illumina's other machines are used for other applications. MiSeq is used for experiments that require lower-throughput and longer read lengths with a faster turn around[121]. NextSeq machines are desktop sequencing tools with fast turn around time used for transcriptome and targeted re-sequencing and thus is commonly used for clinical settings.

Although there are many NGS platforms (Roche/454, Illumina, and Pacific Biosciences, etc.), all use spatially separated, amplified or single DNA molecules, in a flow cell that are massively parallel sequenced[125]. NGS technologies have provided us with an ability to produce enormous amounts of data at a relatively cheap cost. The ever-increasing amounts of DNA sequenced, longer reads and faster turn around times are constantly improving the sequencing technologies.

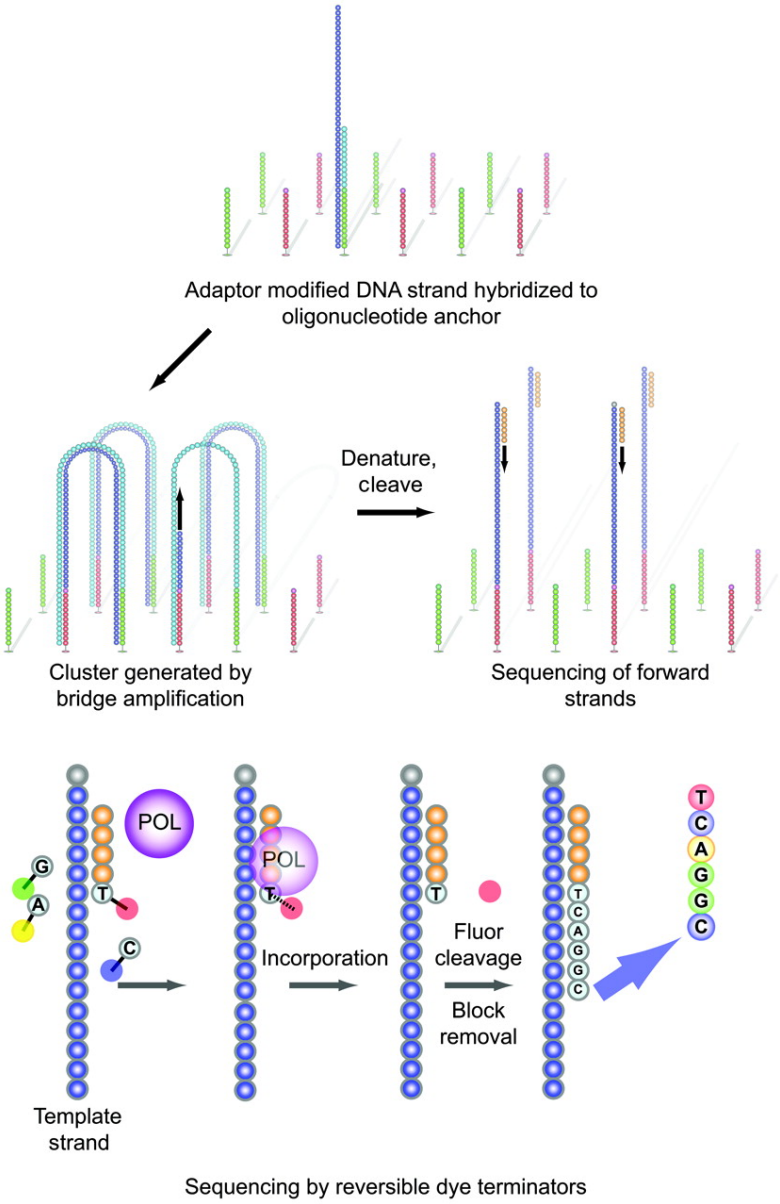


Figure 1.10 Solexa's sequencing methodology using bridge amplification. DNA strands bound by complimentary oligonucleotides to a flow cell arch over to prime the next round of polymerization. This creates clusters of clonal populations via PCR. Fluorophores that can be cleaved between steps show the incorporation of the next dNTP. Adapted from Voelkerdig et al., 2009[126]

1.6.2.2 Third Generation Sequencing

The Oxford Nanopore, Pacific Biosciences' (PacBio) Single Molecule Real Time (SMRT) and Illumina's Tru-seq Syntheic Long-Read are the three commercially available third generation sequencing technologies[127]. Third generation sequencers can be considered as those that are capable of sequencing single molecules (SMS), which negates the need for DNA amplification[121], and can produce much longer reads (generally between 5,000-15,000 bp)[128]. Pac Bio's SMRT was the first, released in

2010, and the reads generally had a raw error rate of 10-15%. However algorithmic techniques and a 50x long read coverage (for *de novo* genome assembly) can allow correction. The main limitation is cost compared to second generation technologies[128].

Illumina's Tru-seq Synthetic Long-read was released in 2012. Long DNA molecules are clonally amplified and barcoded prior to sequencing using a short read instrument this results in synthetically produced long reads from the short read sequences. This technology boats a high accuracy without the need for correction but the standard illumina shortcomings are the same; high GC content and tandem repeats remain troublesome. For de novo genome assembly, cost can be even greater than that of PacBio's SMRT because for 30x long read coverage you need 900x – 1500x short read coverage[128].

Oxford Nanopore's MinION is the newest, released in 2014 and is a handheld device. It works by measuring the small disruptions to an electric current as DNA molecules flow through a nanopore. The MinION has low accuracy and throughput compared to the other third generation technologies. Accuracy can be improved with correction algorithms like those used for the PacBio SMRT. A major benefit of the MinION is its size, cost, and speed, allowing its use in remote areas and for breakout classification[128]. Further improvements on accuracy can make the MinION a powerful tool for the future.

1.6.2.3 Exome sequencing

It is estimated that 85% of mutations that cause disease can be found in coding and functional regions of the genome, and therefore, can be identified through exome sequencing rather than whole genome sequencing. Sequencing only the exomes provides a lower cost per genome/exome and whole exome sequencing provides coverage of more than 95% of the exons[129]. Therefore, exome sequencing can be used to identify the majority of cancer biomarkers at a much lower cost. Exome sequencing can also be used to target non-coding elements such as microRNA and lincRNA[130]. The exome can be captured using either solution based or array based technologies. Solution

based exome capture is most commonly used. Bioitinylated oligonucleotide probes to target regions in the genome are used to capture fragmented DNA. Streptavidin beads then bind the probes and untargeted DNA is washed away. PCR is used to amplify the captured target DNA and this is then sequenced[131]. Solution based capture is most commonly used even though array based capture was the first to be used, this is likely due to less input DNA requirements. However, array based capture has proven to be useful in low GC content regions and SNP detection[131].

1.6.2.4 RNASeq

RNASeq, first used in 2008, is when next generation sequencing approaches are used to sequence total cDNA, allowing quantitative expression scores (similar to microarrays). However, the entire transcriptome can be observed (without prior knowledge requirements for probe production), including novel transcribed regions and transcript structures, such as alternatively spliced isoforms, can also be identified[132]. Due to the desire to determine differential splicing activity, antisense transcription and novel transcriptional regions in eukaryotes, RNASeq has been key milestone for biological experiments in these organisms. The resolution and sensitivity that can be achieved and the range of different changes that can be observed give RNASeq advantages over microarrays. However, there is a significant extra cost, bioinformatics requirements and data storage required for RNASeq experiments[133]. Due to the role of NGS in RNASeq experiments, the limitations of NGS technologies are still present (section 1.6.2.1).

RNASeq experiments have allowed a better understanding of transcription initiation sites, improved detection of alternatively spliced variants, and fusion genes as well as a better identification of sense and antisense transcripts. All of these things are key to cancer research[134]. Developments in RNASeq methods to allow low-input (cDNA pre-amplification) and the use of unique molecular identifiers (UMIs) have allowed single cell RNA sequencing experiments that can identify transcriptomic variation between genetically homogenous cells. This is very important in cancer research where cancer

cells are known to have subpopulations with heterogeneous mutations and transcriptomes[135].

1.6.2.5 CHiPSeq

ChiPSeq, chromatin immuno precipitation sequencing, is the sequencing of DNA fragments that co-precipitate with a DNA binding protein. The most common of the DNA binding proteins investigated with CHiPSeq are transcription factors, chromatin modifying enzymes or modified histones that interact with the DNA. DNA segments that are associated with a specific DNA-binding protein can be identified with ChiPSeq in an unbiased manner, without existing knowledge of precise DNA binding sites[136]. ChiPSeq allows experiments to study gene regulation.

1.6.2.6 Targeted Sequencing

The decreasing cost and improvements to second-generation sequencing technologies mean sequencing of complex organisms will eventually become routine. Currently, sequencing large numbers of whole genomes of Eukaryotes routinely is not yet feasible and thus enrichment for areas of interest can reduce time and cost[137]. There are a number of methods to selectively "capture" genomic regions for sequencing, known as target-enrichment; each has their own advantages and drawbacks. These include PCR (Section 2.1.6), molecular inversion probes (MIP), on-array hybrid capture and in-solution hybrid capture[137].

PCR has been widely used prior to sequencing in experiments. It boasts high sensitivity, good specificity, uniformity and robustness. However, there are issues such as cost, difficulty to multiplex (with the simultaneous use of multiple primers, high levels of nonspecific amplification are observed due to interaction between primer pairs), and an upper limit to the generated amplicon size. Also, in practice not all amplification reactions yield products, which is a key problem when working with clinical samples[137].

MIP uses the enzyme ligase to circularize single stranded oligonucleotides formed of a common linker flanked by target-specific sequences. Exonucleases are then used to digest uncircularised species, leaving only the circularised oligonucleotides to be amplified via PCR, using primers targeting the linker. DNA polymerase is used to "gap fill" between target specific MIP sequences. Gap fill and PCR can occur in small volume, aqueous solution, meaning they are easy to scale to large numbers via a 96-well plate. Another advantage is that barcodes for identifying purposes can be incorporated into the primers allowing pooling of multiple samples and input requirements can be as low as 200ng[137]. Issues include capture uniformity, which have been improved modestly but remain this technique's biggest downfall.

Hybrid capture is performed using immobilised specific probes that hybridise the shotgun fragment library and the un-targeted DNA strands are washed away whilst those captured are eluted. Arrays can hold 2.1 million probes per array with the ability to capture 34Mb[137]. Compared to PCR based approaches, array techniques are quicker and less laborious. Hybrid capture also has its drawbacks including expensive hardware, high starting material requirements (10-15µg) and limits to a) the number that can be performed in a day and b) the number of samples in a study (large numbers aren't feasible).

In solution capture is similar to array capture, with an excess of probes allowing less starting material. Again this technique can be used in 96-well plates meaning it is readily scalable without the need for specialist equipment[137].

1.6.3 Polymerase Chain Reaction (PCR)

PCR is an important laboratory technique that is capable of amplifying a single DNA sequence to make thousands/millions of copies. The PCR procedure has multiple heating and cooling steps:. The reaction mix (DNA, dNTPS, DNA polymerase, buffer) is heated to 94-98°C to denature double stranded DNA and then cooled to enable the

sequence-specific hybridisation of the primers to the single stranded DNA. DNA polymerase then makes a complementary DNA strand extending from the 3'end of the hybridised primer. These heating and cooling steps can then be repeated to create more and more copies of the DNA.

PCR can be used to detect presence/absence of a specific target as well as to quantify the amount of target present. Presence/absence can be observed via gel electrophoresis, using a ladder of known sizes to obtain product size. Quantification is generally performed using fluorescent dyes.

There are multiple uses for PCR, for example real-time PCR can monitor the amplification of target nucleotide sequences in real time by either using fluorescent dyes that intercalate between dsDNA in a non-specific manner or by using target specific probes that are fluorescently labelled. The number of cycles required for the product to exceed a predetermined fluorescence threshold is measured (as a cycle threshold- or ct-value) to infer the amount of starting target material. Quantification can be also be performed post-PCR. Nested PCR uses using two sets of primer pairs in sequential reactions. It is used to reduce non-specific probe binding, and increase sensitivity: PCR product from a first PCR is used to seed a subsequent PCR containing a second set of 'nested' primers that hybridise to sequences 3' to the first round primers in the amplified product. This improves specificity as it is unlikely that DNA other than the intended target sequence would hybridise to both primer pairs.

PCR has been used to detect mutations and biomarkers, and to diagnose cancer. Leading up to the development and cost reduction of NGS (section 1.6.2.1), many scientists were using PCR-based investigations into cancer biomarker discovery: A reverse transcription-PCR assay of 761 transcripts was used for the discovery of colon cancer biomarkers[138]. Comparisons with targeted NGS have shown that real-time PCR and NGS have significant concordance (96.3 to 100%) for detecting *EGFR*, *KRAS* and *BRAF* mutations in FFPE materials. However, NGS was capable of identifying seven non-synonymous SNVs and an indel in EGFR that was not detected by the real-time

PCR method[139]. PCR is also commonly used for target enrichment for targeted sequencing of genes or specific transcript splice variants (section 1.6.2.6).

1.6.3.1 OncotypeDx

OncotypeDx[14] is a multi-gene expression array that uses quantitative reverse transcription polymerase chain reaction-based assay. It is used clinically to give prognostic and predictive value in early stage ER+ breast cancers, to predict the benefit of chemotherapy with adjuvant hormone therapy[14].

1.6.3.2 Prolaris

Prolaris is another quantitative reverse transcription polymerase chain reaction-based assay. It can be used (alongside patient and tumour information) to predict the aggressiveness of PCa. It utilises thirty-one cell cycle progression genes and fifteen housekeeper control genes for PCa tissue[16]. The expression of the thirty-one cell cycle progression genes are correlated with PCa proliferation to serve as a risk-stratification tool: a lower score means lower risk and these men may be prime candidates for AS and a higher score represents those needing treatment[140]. It can also be used to predict ten-year PCa specific mortality and ten-year PCa BCR[16]. Cell cycle progression genes have also been used in the prognosis of other cancers[16].

1.6.4 Microarrays

Prior to the affordability of sequencing as a method to identify biomarkers microarrays were frequently used, and are still incredibly valuable due to cost and availability of standard pipelines for analysis. Microarrays are significantly cheaper than sequencing and so are still often used today. Microarray technology allows the user to assess DNA copy number or RNA expression levels in cells or tissues in different disease states. They are relatively cheap, not considerably time consuming and the array data can be re-investigated for many different questions. Microarrays have been utilised in gene

discovery and regulation involved in physiological, developmental and pathological processes, in diagnosis and drug discovery[141].

Microarrays are an array of specific DNA sequence 'probes'. Fluorescently labelled DNA/cDNA samples are hybridised to the probes, the excess DNA washed off, and the quantity of each DNA sequence is assessed by the strength of the fluorescent signal that remains attached to each probe. on the array, However, cross-hybridisation is an issue in microarray experiments, leading to false positives, and masking of eg down-regulated transcript signals. Analyses can use either single-channel (one sample hybridised) or two-channel microarrays (two differentially labelled samples hybridised at the same time).

Two-channel microarrays have been used to directly compare gene expression between two different conditions, e.g. cancer cells with normal cells to identify genes that have expression changes in cancer. The two samples (one cancer and one normal) are labelled with two different fluorophores (often cy3 (green) and cy5 (far-red)) the two samples are then mixed and simultaneously applied to the microarray for hybridisation (Figure 0.11).

Single-channel microarrays provide intensity data for each specific target DNA/cDNA that hybridises to its matching probe. It provides data on the relative abundance of each probe sequence in a sample and can be compared to data from multiple samples A downside to single-channel microarrays is that unless great care is taken in consistency of sample preparation, microarray hybridisation and washing conditions etc., then error rates can be higher than those achieved from two-channel microarrays.

Oligonucleotide microarrays, like single channel microarrays use one fluorescent label for all of the samples (Figure 0.11). They use short genomic ssDNA fragments that allow sequence coverage of an entire genome, and therefore, can be used for extensive genetic profiling and mutational analysis by providing absolute yield values for each specific target gene. They are capable of providing a presence or absence call for each gene, but two separate arrays are required to allow the comparison of healthy controls to

cancer patient samples[142]. Affymetrix is the major producer of microarrays: They provide standard arrays for many species, for example, the human genome U133 array, which contains 45,000 probe sets for 39,000 transcripts from 33,000 well-substantiated human genes. Affymetrix also produce of custom arrays for a wide variety of different uses. A standard affymetrix arrary contains oligonucleotide probes, 25 bases long, specific to targets are fixed to a glass wafer, in set locations. Each oligo is present in millions of copies to allow accurate interpretation of expression levels, from measured intensities of fluorescence given by the tagged hybridised nucleotide sequences.

1.6.4.1 Decipher

Microarrays have been used in cancer biomarker platforms such as Decipher[15]. Decipher is a classifier score calculated from a gene expression microarray analysis of 22 coding and non-coding RNA probes[15], that predicts metastatic PCa progression/high risk of recurrence and PCa related mortality within 5 years of RP. High-risk of recurrence is defined by extra-prostatic extension, seminal vesicle invasion, positive margins or biochemical recurrence. Whilst the 22 specific probes are unknown, the panel represents known pathways involved in aggressive PCa, including cell proliferation, cell structure, immune system modulation, cell cycle progression and androgen signalling[143].

The score ranges from 0-1, with every 0.1 increase representing a 10% increase in risk of metastatic progression. The score is then more generalised into three categories; low-risk 0-0.44, intermediate risk 0.45-0.59 and high risk 0.6-1.

Whilst Decipher was originally established as a predictor for metastatic progression post-RP, there have been further applications since. Decipher has been evaluated for its ability to ease decision making between adjuvant and salvage radiation therapy (second-line treatments); Dalela et al., showed that any two or more risk factors from pT3b-T4, G8-10, lymph node invasion or >0.6 decipher score showed a 4-fold reduction in metastatic progression at 10 years with adjuvant therapy. A score >0.6 had up to 80%

reduction in metastatic progression if adjuvant radiation therapy was received[144]. The PRO_IMPACT was a multi-institutional study that showed decipher could significantly decrease decision conflict and patient anxiety. Decisions on adjuvant and salvage therapy were altered with the addition of a decipher score in 18% and 32% of cases, respectively[145].

Another key finding was that decipher could also be performed on small amounts of genetic material like that obtained from biopsy and also including FFPE tissues. Decipher was tested on the biopsy material of 219 men who then went on to have RP to validate the findings, this gave HR = 7.3 and HR = 11 when moving from low-risk to high-risk on multivariate analysis[145]. In a second study, decipher was applied to the biopsy material of 57 men, who proceeded to undergo RP and also had long term follow up. Here decipher was capable of predicting metastatic progression as an individual predictor with AUC = 0.72[146]. This highlights its potential use in aiding decision making for primary treatment also, helping to identify those who are safe for active surveillance and those who should receive treatment more swiftly.

A limitation of decipher in the aid of primary treatment decision making, is that it relies on biopsies, and so carries the same limitations of a biopsy: PCa is often a multifocal disease and the lower-grade or lower decipher scoring foci could be picked up by biopsy, whilst the higher-scoring foci is missed. Leading to a less severe prediction occurring. Decipher has shown great promise as a second line treatment informer and clearly has a role here in PCa management.

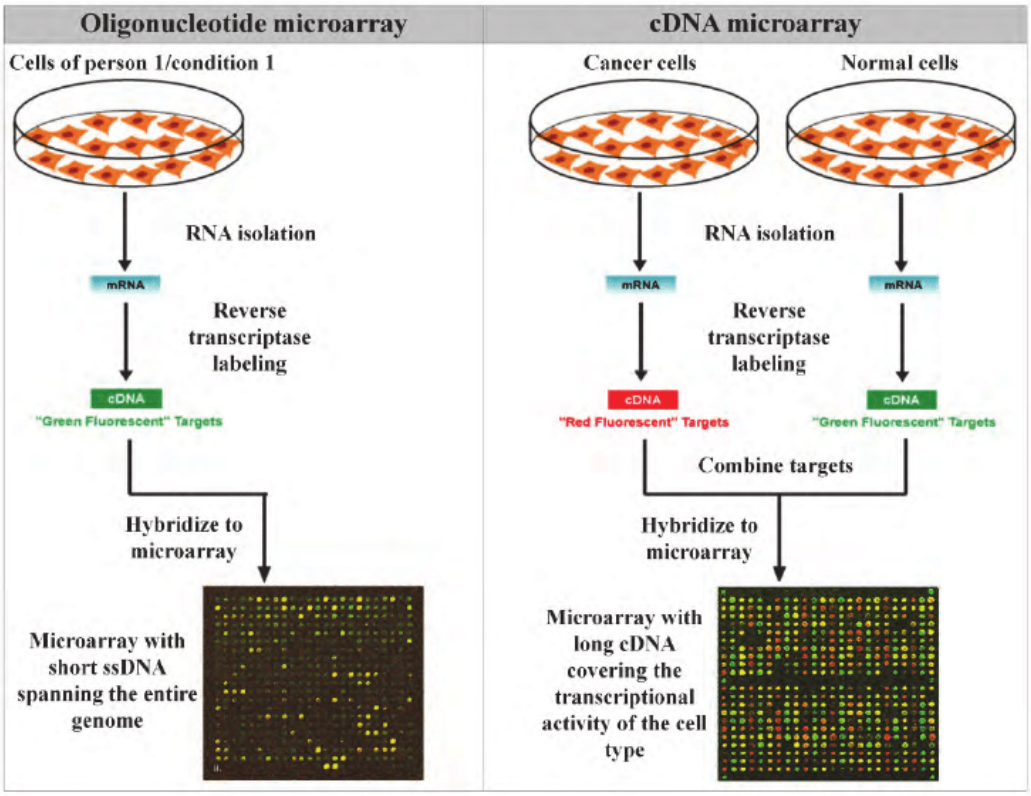


Figure 0.11 A schematic for oligonucleotide and two-channel microarrays. Both show RNA isolation from the cells of interest, followed by reverse transcriptase labeling to create cDNA from RNA and then hybridisation to array. In two channel arrays, cDNA from the normal cells and the "condition" cells are combined prior to hybridisation. Adapted from Vermeeren et al., 2011[142].

Sequencing trumps microarrays with its ability to provide further information about specific unknown mutations. Mutations can be detected via microarrays; however, the probes must be designed to hybridise that specific mutation as the target gene, meaning the mutation must first be known. Cross-hybridisation problems in microarrays also mean that SNVs will be unable to be detected. Sequencing can also detect novel gene fusions. However, if you only require count data microarrays hold some advantages in comparison to sequencing: They are cheaper and the analysis of the data produced is easier. There are well-known analysis pathways to take, whereas, the best method for sequencing data analysis is still being investigated. Also, there is a lot of data available to the public from standardized platforms, which can be utilised for comparison with your own samples.

1.6.5 Mass Spectrometry

Mass spectrometry (MS) has shown great promise in proteomics and the identification of protein biomarkers. Proteomics is the large scale analysis of proteins including their structure and function; it provides information about the complex end products of a gene[147]. MS has positioned itself as one of the key technologies for the unbiased identification of cancer biomarkers[148]. Combining MS with liquid chromatography allows easy profiling of bodily fluids (samples which generally are less invasively obtained from patients) whilst MALDI-MS (matrix-assisted laser desorption/ionization) is useful for identifying biomarkers in FFPE tissues[149].

MS works by bombarding molecules with electrons (ionizing) to create charged molecules and measuring their mass-to-charge ratio, by accelerating them and applying an electric or magnetic field. Ions of the same mass-to-charge ratio are deflected at similar amounts and can be detected via an electron multiplier. Results are available as a "spectra", which can be correlated to previously known masses to identify atoms or molecules present in the sample.

Proteomics has an advantage over genomics, as it will be clear if a mutation is making a big difference to the protein, which can never truly be proven with genomics, just inferred[148,150]. However, MS suffers from insufficient sensitivity when detecting low-concentration biomarkers in a sample with a high-abundance of proteins, making depletion of abundant protein fractions and enrichment of biomarkers imperative to improving MS sensitivities[150]. MS also suffers from accuracy and reproducibility problems caused by software issues, meaning samples need to be run, typically, 10 times, increasing the amount of material required. These issues need to be addressed in order for Mass Spectrometry to develop as an efficient tool for biomarker discovery. Also, unlike other biomarker detection technologies, MS is limited to providing information of presence/absence and levels of the proteins cannot be determined unless targeted MS is performed. Targeted MS means the experiment must be focused on a

small subset of protein targets to achieve their quantification, thus reducing the scope of proteins that can be quantified within the experiment[151].

1.6.6 Fluorescent In-situ Hybridisation (FISH)

FISH, a cytogenetic technique, can be used to detect chromosomal abnormalities; changes in chromosomal structure and numbers (including genomic deletion and fusion genes) can be observed when viewing cells or chromosome preparations upon a slide. Chromosomal abnormalities are common in many tumours, and Some of these abnormalities can be used for diagnostic and prognostic purposes[152]. FISH was being used to identify specific chromosomal regions and loci (chromosomal mapping) by the late 1980s. It works by labelling DNA with fluorophores, which emit light detectable by microscopy. FISH probes are capable of hybridising to DNA and RNA of circulating tumour cells (CTCs) and FFPE tissue sections that are fixed. This allows FISH to be useful for solid tumours as well as hematological cancers. Probes are designed for specific target sequences, and usually consist of cloned DNA sequences in the form of BACs, PACs, fosmids and cosmids, but can also be PCR products. The DNA probes can be tagged directly with fluorophores, or with biotin or DIG that can be bound post-hybridisation with streptavidin linked fluorophore or anti-DIG antibodies bound to fluorophore. Short DNA fragments are added to block repetitive DNA sequences and then the probes are applied to the cell preparations on a glass slide. Hybridisation requires approximately 12 hours followed by several wash steps in order to remove non-bound or partly bound probes. After which a microscope can be used to excite the dye and record the images for location and quantification of aberrations.

Improvements in fluorescent dyes and advances in microscopy and imaging allowed for MFISH (multi-fluorochrome assays), particularly SKY (Spectral Karyotyping). This new method allowed for entire metaphase spreads to be investigated using 24-colours, which showed the chromosomal origins of structural rearrangements[153].

FISH is a reliable, simple and specific assay for biomarker detection, and because of this, even though it is a low throughput method, it remains to be a cornerstone in genetic labs and even in clinical practice for the diagnosis, treatment stratification and prognosis of cancers. Whilst high-resolution molecular profiling techniques (microarrays and sequencing) are advanced in identifying novel chromosomal abnormalities, FISH remains a reliable validation method for any potential biomarkers identified[154].

1.6.7 Immunohistochemistry (IHC)

IHC is still commonly used in cancer diagnosis, and can validate biomarkers identified from other methods. Now that molecular, quantitative, global methods exist for novel biomarker identification, it is used much less to identify these but more to locate where in the cell the biomarker is and to validate its presence/absence in cancer tissues[155].

For IHC, first tissue needs to be collected, fixed (commonly with paraformaldehyde) and often embedded in paraffin wax. The tissue is the sliced (4-40μm), mounted on slides and dehydrated with alcohol washes and cleaned with xylene before imaging via microscope. Blocking buffers are often used to reduce background staining. Positive and negative controls are required, a tissue known to express and a tissue known not to express the specific protein. Antibodies specific to the target antigen must be extracted from animals; the protein of interest is injected into the animal to elicit an immunological response producing the desired antibody. Therefore, this can make IHC time-consuming. Monoclonal or polyclonal antibodies can be used, targeting one epitope or multiple, respectively. Antibodies are often linked (using biotin) to reporter molecules. Reporter molecules can either be fluorophores or enzymes allowing fluorescence or chromogenic detection. There are two methods of antibody detection: Direct and Indirect. The direct method is where the labelled antibody directly binds the antigen, and the indirect uses an intermediate antibody to bind the antigen to which the labelled antibody binds.

In order to compete with the new molecular methods, IHC will need to be quantitative[156]. IHC is specifically useful for the validation of protein biomarkers, similar to ELISA (section 1.6.8).

1.6.8 Enzyme-linked Immunosorbent Assay (ELISA)

Similarly to IHC (section 1.6.7), ELISA is used for protein detection. ELISA works by using enzyme linked antibodies to capture antigens, and colour changes from the enzyme binding its substrate provide detectable signals, which are proportional to the amount of antibodies bound to the antigens present. IHC and ELISA have their own advantages; ELISA is fully quantifiable and easily standardised with quality assured measurements obtained. IHC is at best semi-quantitative but allows insight into tissue heterogeneity and can be performed on both frozen and paraffin-embedded tissues (section 1.6.7)[157].

There are three main types of ELISA: Indirect, Sandwich and Competitive, all use 96-well microtitre plates as the immobilising surface, allowing moderately high-throughput investigations. ELISA is a versatile and robust tool and so ELISA is often used for validation of biomarkers[150].

ELISA only allows the detection of a single antigen and often requires a large amount of sample. This along with the narrow dynamic detection range means it not useful in biomarker discovery. ELISA is costly and its quality is dependent upon antibody quality, users skill and experience, and shows problems with accuracy and reproducibility. ELISA's downfalls mean it is less useful at biomarker validation when there are multiple proteins, which is often the case.

1.6.9 Methylation Assays

Epigenetic gene regulation, such as methylation and histone modification, is an important factor in normal development and disease states like cancer[158]. These are modifications to our gene expression that is not encoded for by the DNA, but inherited

mitotically[159, 160]. Hyper-methylation of promoter regions is commonly seen in cancers to knock-down the expression tumour suppressor genes[161]. Methylation is the addition of a methyl or hydroxymethyl group to the C5 position of cytosine, which occurs at or around CG dinucleotide regions (known as CpG islands and shores)[162]. Methylation is known to aid in cell cycle regulation and cellular differentiation processes[163,158]. The role of DNA methylation has been well established in many cancers including PCa[163, 164, 165]. Hyper-methylation of several genes, including GSTP1, is commonly observed during the transition between intraepithelial neoplasia to carcinoma[166]. Hyper-methylation detection has shown promise as biomarkers for the diagnosis and prognosis of cancers.

There are many methods for the identification of methylated sites, which method you chose can be based on many things, but importantly is the biological question you are asking: There are different methods available for whole genome methylation profiling, identifying regions of differential methylation status, or for determining the methylation status of specific genes of interest[167]. Other factors to include when choosing a method are the amount and quality of the sample, the sensitivity and specificity requirements for the experiment, robustness and simplicity of method, and its bioinformatics analysis, as well as the availability of specialist equipment and overall experiment cost[167].

1.6.10 Supervised and Unsupervised Analyses

Due to the developments in genomic technology more and more biological data is being developed that needs to be analysed; to identify patterns and trends and understand what the data means to the biological question. The application of statistics on such data can be called statistical learning or machine learning, and can roughly be separated into two categories: supervised and unsupervised analyses[168]. These can be referred to as classification and clustering, respectively[169].

For supervised learning, or "classification" of observation x, an observation with multivariate *p* dimensions (also called features) and associated with class *c*. The

purpose is to "learn" a mathematical function that when evaluated with the input *x* provides a prediction of its class *c*. In general practice data is subset into training and test datasets. The training set is used to "set" the mathematical function to correctly predict the class for each observation provided. This function, with the parameters set from the training data is then applied to the test data to observe its ability to correctly classify the data without bias[169]. Examples of supervised machine learning are generalized linear models (glm) (section 2.6.1), probit regression (polr) for ordered multivariate models, random forest (section 2.6.3). These methods can be accompanied by a shrinkage method to reduce over fitting and thus improve predictability; examples of such methods are Lasso (section 2.6.2) and Step (section 2.6.4).

Unsupervised analysis, or "clustering" can also be referred to as class discovery. A key difference between unsupervised analysis and supervised analysis is a lack of training set for the former and thus no cross-validation. A second important difference is that clustering algorithms are set using optimality criterion and there is a lack of guarantee that the global optimal solution is found, and therefore a heuristic approach is often taken. A choice of a) features to be used, b) similarity metric, and c) algorithm needs to be made for many methods[169]. Unsupervised learning can be further partitioned into hierarchical clustering (section 2.5.2), (which can then be subdivided into agglomerative and divisive) and partitioning. Hierarchical will cluster data into a tree like feature and then to achieve a desired number of clusters one can cut the dendogram at a desired height. However, partitioning generally requires the user to specify the number of clusters prior to clustering[169]. Examples of partitioning are k-means clustering (section 2.5.3), principal component analysis (section 2.5.1), and latent process decomposition (section 2.5.5).

1.7 Summary and Aim

1.7.1 Summary

Whilst the introduction of the PSA test has decreased mortality from PCa, the increased incidence rate that can also be attributed to it comes with problems of over-diagnosis and over-treatment. Highlighting the need for additional biomarkers for the diagnosis of PCa. A need for biomarkers for hormone therapy response prediction, BCR prediction, further treatment stratification, and prognosis were also highlighted.

The heterogeneity of PCa means that there have been a lot of potential biomarkers discovered, but also that they are not always consistent in the tumours. Meaning a limited number are capable of being used for the diagnosis and prognosis of PCa. However, combinations of biomarkers in a panel could be of great clinical use. The utility of urine in PCa biomarkers is well established via the PCA3 test (section 1.4.2) and the role of exosomes in cancer development and metastasis (section 1.5) has highlighted a resource to be investigated.

The development of NGS technologies and the continuous advancements in sequencing technologies are making it possible to investigate a large number of genes across a large number of samples, at continuously decreasing costs. Sequencing is an important technology for the discovery of novel biomarkers, as it is capable of identifying expression changes and mutations at high-throughput. The reducing costs of sequencing are closing the gap between data production costs and data processing costs, it is said that there may come a time when processing the data will become more costly. Bioinformatic analysis of the data is still under on-going development to identify the optimal pathways for analysis. Currently, cheaper methods for high-throughput expression analysis (microarrays and Nanostring) still hold a firm place within biomarker discovery. Whilst, mass Spectrometry for proteomic biomarker discovery holds massive potential, however there are issues still to overcome.

Older techniques of biomarker discovery hold great sensitivity but are at considerably low-throughput, making them very good for validation and clinical detection after a few potential biomarkers are selected from higher-throughput methods. These include: FISH for gains/losses, rearrangements and chromosomal instability investigations, IHC and ELISA for the validation of particularly proteins and to see where in the cell these biomarkers are gained to or are lost from and PCR-based methods for confirming mutations in the biomarkers.

Knuutila *et al.*, compared NGS, aCGH, FISH, PCR and IHC methods for specific biomarker analysis of FFPE tumour tissues. Their conclusions suggest that NGS has the potential to replace all other methods tested for the analysis of tumour biomarkers, especially as the reducing costs and required sample material decreases to that near of FISH or PCR. NGS allows the investigation of mutations, gene fusions and copy number changes in one single analysis[170]. However, NGS has not currently reached the position where it is commonly used in clinical practice.

1.7.2 Aims & Objectives

PCa diagnostics and prognostics currently lack specific and sensitive clinical biomarkers and treatment is not well individualised. The PCA3 test highlights the utility of urine in PCa diagnostics and prognostics. The aim of our work is to interrogate PCa patient's urine samples, mostly the exosomal fraction to identify novel biomarkers or sets of biomarkers to aid in PCa management. My objectives are as follows:

O1: To determine whether RNA expression from urine extracellular vesicles in prostate cancer patients are a viable target for the development of biomarkers through the use of Nanostring technology.

O2: To determine an optimal combination of probes to predict cancer presence and aggression in prostate cancer patients.

O3: To determine whether an optimal combination of probes can predict response to hormone therapy treatment.

O4: To evaluate the differences between urine fractions (extracellular vesicles and cell sediment) and determine whether cell sediment can be used to predict cancer presence and aggression in prostate cancer patients.

Below are described more detailed aims for each chapter.

3.1.1.1 Chapter 3: NanoString Data Analysis 1: The Pilot Study

This chapter encompasses the analysis of the pilot study of samples sent to NanoString to investigate exosomal RNA expression level changes of 57 target sequences. The RNA was extracted from the EV fraction of urinary samples collected at the NNUH as part of the Movember study. The aims were to primarily determine if the transcript content of urinary exosomes contained any PCa derived transcripts and if transcript level could be utilised for risk stratification. Also, it was important to investigate if NanoString was a suitable method for obtaining expression data from these cDNA-amplified samples and to determine suitable methods for analysis.

3.1.1.2 Chapter 4: NanoString 2 Analysis: The Movember GAP1 Project

A second analysis for the Movember study. RNA was extracted from the EV fraction of urinary samples that were collected from multiple centres (NNUH, Norwich, St James' Hospital, Dublin, Royal Marsden Hospital, London, and Emory Healthcare, Atlanta). 864 samples were sent to NanoString for the quantification of 167 transcripts. The aims were to primarily identify optimal models capable of predicting PCa and to risk-stratify PCa without the need for biopsy. Models were built to answer four important clinical questions: 1) determine which samples were from PCa and which were from samples with no evidence of Ca 2) determine which samples were from high-risk PCa only and which were from samples with no evidence of cancer 3) determine if there was a trend in expression that corresponds to a trend in risk category (CB>L>I>H) and 4) determine if there was a trend in expression that corresponds to a trend in patient type (CB>Ca>Metastatic cancer).

3.1.1.3 Chapter 5: Response to treatment

Many cancers have benefitted from treatment stratification due to expression of certain genes, however not yet PCa. With hormone therapy (HT) it is known that patients will inevitably progress to castration resistant prostate cancer (CRPC). How long each patient will last on HT varies widely from months to years. It is our aim to use the NanoString data of the advanced patients in the pilot study (n = 32) to see if a significant predictor of early progression in patients on HT can be built and whether this predictor improves on current clinical information collected (e.g. PSA, Gleason score and bone scan). The NanoString 2 data can then be used for validation of this predictor.

3.1.1.4 Chapter 6: Analysis of Cell Fraction and comparison with exosomal fraction

The use of RNA extracted from EV fractions and cell sediment fractions were used to compare the transcriptome profiles from PCa patients and controls (taken from patients with no evidence of cancer (CB)). The aim was to identify if both fractions contained similar expression profiles of genes and if either contained higher levels of prostate or PCa associated transcripts. The fraction with the highest level of these transcripts is likely to be a better source of material for PCa diagnosis and risk stratification. Data from microarray of samples collected from NNUH, Norwich and Royal Marsden Hospital, London was used.

Secondly, I am to use NanoString data from cell sediment fraction derived transcripts (collected only from NNUH, Norwich) to identify optimal models to answer the four important clinical questions asked of the EV derived data (Chapter 4).

2

Materials and Methods

2.1 Sample Collection and Processing

Overview: Urine samples were collected from patients attending hospital clinics. Extracellular vesicle (EV) RNA was harvested by urine microfiltration (Section 2.1.2). EV and cell pellet RNA was extracted (Section 2.1.3), converted to cDNA and amplified as cDNA (Section 2.1.4), ready for NanoString expression analysis (Section 2.1.5).

Not all the procedures in this section were performed by me but were included in this thesis as essential information relative to the study.

2.1.1 Sample Collection

Note: The procedures in this section were not performed by me.

Urine samples were collected in a 30ml Universal tube (Sterilin) from urology clinics at the Norfolk and Norwich University Hospital (NNUH, Norwich, UK), St. James Hospital (Dublin, Republic of Ireland) and from primary care and urology clinics of Emory Healthcare (Atlanta, USA), between 2012 and 2015. Most samples were collected as first void post-DRE but a few matching pre-DRE samples were collected for comparison (these were labelled as such). All samples were collected from treatment naïve patients. Control samples were collected at a micro-haematuria clinic at the NNUH, again first void post-DRE in a 30ml Universal tube. Microvesicular RNA was harvested by ultracentrifugation (section 2.1.2), extracted (section 2.1.3), converted to cDNA and amplified (section 2.1.4). RNA from the cell pellet was also processed, using either the Qiagen Allprep DNA/RNA mini kit cat no: 80204 or RNeasy micro kit cat no: 74004 according to manufacturer's instructions).

The lab also had access to urine samples collected as part of the active surveillance prospective study at the Royal Marsden Hospital NHS Foundation Trust (RMH) between 2009 and 2012. The active surveillance prospective study collected samples, first void post-DRE, specifically from men with untreated, low-risk prostate cancer. Low-risk PCa defined as having clinical stage T1/T2a, Gleason 3+3 (or 3+4 of older than 65), PSA<15 and <50% positive cores. Three of these samples were collected pre-DRE from post-radical prostatectomy patients for comparison. Microvesicular RNA was harvested as above.

The study was given favourable ethical opinion by the NRES Committee East of England – Norfolk on 21st August 2014 under the study title "Urine biomarkers for detecting prostate cancer". Ethics was approved to Dr Marcelino Yazbek Hanna of NNUH with REC reference: 12/EE/0058 and IRAS project ID: 96199.

2.1.2 Micro-filtration harvesting of Urine Extracellular Vesicles

Note: The procedures in this section were generally not performed by me (I performed these procedures on ~20 samples).

Urine samples were processed within four hours of samples collection. Urine was centrifuged at 1200g for 5 mins, and then the supernatant was transferred to a 50ml tube and re-centrifuged at 2000g for 5 mins. Supernatant was decanted and then filtered through a 0.8μM filter (Millipore), transferred to an Amicon Ultra-15 100KDa MWCO microfiltration unit and spun at 3400g r/t for 15 mins or until the volume was reduced to below 500μL. PBS (15ml) was added to the sample followed by further centrifugation until the volume containing the EVs was reduced to 200μL. Transmission electron microscopy (TEM) was performed to confirm the presence of EVs.

2.1.3 Qiagen RNA Extraction

Note: This section was generally not performed by me (I performed this step on ~20 samples).

The Qiagen Micro RNA RNeasy kit was used for RNA extraction from EVs and cell pellet as per the manufacture's manual. 700μL of buffer RLT was added to the cell pellet or EV samples. The cell pellet/RLT mix then had an extra step, which was to disrupt the cells using a QIAshredder spin column for 2 mins at full speed (~12,000 rpm) in a microfuge. From this point onwards the cell pellet and EVs were treated the same. 70% ethanol was added and the mixture pipetted into a MinElute spin column and centrifuged in a microfuge (15 seconds, >10,000rpm). 350μL of buffer RW1 was added to the MinElute spin column before re-spinning (15 seconds, >10,000rpm). Then 80μL of Qiagen DNase mix I was directly applied to the membrane and left to stand at room temperature for 15 mins to complete DNA digestion. The wash step with RW1 was then repeated followed by the addition of 500μL of buffer RPE and re-spun (15 seconds, >10,000rpm). 500μL 80% ethanol was added and then spun (2 mins, >10,000rpm). To dry the membrane the column was spun for a further 5 mins with an open lid. The column was transferred to a fresh collection tube, and the RNA was eluted with 14μL of RNase free water and centrifuged (1

minute, 12,000 rpm) in a microfuge. Nanodrop and Bioanalyzer were used to confirm that RNA was of a good quality.

2.1.4 Nugen Amplification of RNA as cDNA

Note: I performed the amplification step for 286 samples.

Amplification was performed using the Nugen Ovation picoSL WTA V2 kit as per the manufacture's instructions[171]. The kit works via the following mechanisms: Firstly, the first strand of cDNA was generated using a DNA/RNA chimeric primer mix (containing a mix of random and oligo dT primers such that priming occurs throughout the whole transcript) and reverse transcriptase (RT). The RT extends the 3' end of the DNA for each primer resulting in a cDNA/mRNA hybrid containing a unique RNA tag sequence known as the SPIA tag at the 5' end of each cDNA strand. The SPIA tag was used for a priming site for the SPIA process.

Secondly, fragmentation of this cDNA/mRNA complex was required to provide priming sites for RNA polymerase to synthesise a second cDNA strand. This includes DNA complementary to the 5' SPIA tag and results in a double stranded cDNA with a DNA/RNA heteroduplex, which corresponds to the SPIA tag. Finally, strand displacement amplification occurs that uses a DNA/RNA chimeric primer (SPIA primer), DNA polymerase and RNase H in an isothermic assay. RNase H removes the RNA part of the heteroduplex SPIA tag allowing the SPIA primer to bind. DNA polymerase can then synthesise from the 3' end of the primer displacing the existing forward strand with new cDNA. Priming with the chimeric SPIA primer then in turn makes a new heteroduplex SPIA tag, which becomes the new substrate for RNase H and can initiate the next round of cDNA synthesis. These last few steps were repeated in a highly processive manner allowing rapid accumulation of μg of amplified cDNA from ng of total RNA.

The actual process was as follows: samples were diluted with RNase free water to ensure all contain 20ng of total RNA in a PCR tube. The first strand synthesis primers were added

(2μL) to each sample and they were heated to 65°C for 2 mins. The first strand buffer and enzyme were pre-mixed, 2.5μL and 0.5μL per sample, respectively. 2.9μL of this mix was then added to the sample tubes and samples were returned to the thermal cycler for program 2: 4°C for 2 mins, 25°C for 30 mins, 42°C for 15 mins, 70°C for 15 mins and hold at 4°C. Second strand synthesis required second strand buffer and enzyme to be pre-mixed; 9.7μL and 0.3μL per sample, respectively. 9.5μL of this mix was added to each sample, mixed via pipetting (5x) and returned to the thermal cycler for program 3: 4°C for 1 minute, 25°C for 10 mins, 50°C for 30 mins, 80°C for 20 mins and hold at 4°C. The cDNA must then be purified using the magnetic beads provided in the NuGEN kit; 32μL of the beads were added to each sample and mixed via pipetting (10x) and were incubated at room temperature for 10 mins.

Following this the samples were placed into the 96 well magnet and were incubated at room temperature for a further 5 mins. Using long thin pipette tips, 45μL of buffer was removed as the beads (that were attached to the cDNA complexes) were pulled to the tube side via the magnet. The tubes were then washed with 70% ethanol (200μL) three times and left to air dry at room temperature (roughly 25 mins but until there was no liquid left in the tubes). The last step was then the SPIA amplification; where the SPIA buffer, the SPIA primer mix and the SPIA enzyme were pre-mixed in order; 20μL, 10μL and 10μL per sample, respectively. 38μL of the SPIA mix was added to each sample and the samples were returned the thermal cycler for program 4: 4°C for 1 minute, 47°C for 75 mins and 95°C for 5 mins. At a different bench, the PCR tubes were returned to the magnet for 5 mins, and the liquid that contained the amplified cDNA was collected.

TE was added to a final concentration of 0.2xTE and yields determined via Nanodrop or Qubit.

2.1.5 NanoString

The NanoString nCounter gene expression system technology uses 2 probes; a capture and a reporter probe[115]. The probes were designed to have a complementary sequence to the specific transcripts we wished to study. Each transcript reporter-probe had a distinct string of fluorescent coloured beads attached[115]. Up to ~800 different bead string combinations are available, and so up to 800 different transcripts can be detected in one analysis.

The probes were hybridised to the complementary nucleic acid sequences in each sample, forming a tripartite complex of the 2 probes and target mRNA or cDNA. The complexes were then pulled down and immobilised onto a capture surface, unbound sequences were washed away and an electric field was passed across the surface to stretch out the nucleotide and bead complexes. The bead-complexes were then imaged and the number and type of each string of coloured beads counted. This provided a direct measure of RNA or cDNA counts per transcript[115]. Twelve samples loaded onto the NanoString machine at a time (in each cartridge).

2.1.6 PCR (Polymerase Chain Reaction)

Note: I performed PCR detection of *TMPRSS2:ERG* fusions for 113 samples.

TMPRSS2:ERG fusions were detected by primary PCR and confirmed with a secondary PCR that used nested primers. A master mix was made using the following components (volumes provided for a single PCR): 2.5μL 10x PCR buffer, 1μL 50mM $MgSO_4$, 0.5μL 10mM dNTP mixture, 0.5μL Primer 1 (10μM), 0.5μL Primer 2 (10μM), 0.1μL Platinum Taq (Thermo Fisher) and 19μL HPLC H_2O. Primer 1: CAGGAGGCGGAGGCGGA (*TMPRSS2* exon 1 Forward). Primer 2: GGCGTTGTAGCTGGGGGTGAG (*ERG* exon 6 Reverse). The master mix was pipetted into a clean 0.25ml tube and 1μL of the template cDNA was added. PCR conditions were as follows: 94°C for 30 seconds, followed by 35 cycles of 94°C for 20 seconds to denature and 68°C for 60 seconds to extend. A second master mix was created using the same reagents but using 0.5μL of

the nested primer 1 and 0.5μL of the nested primer 2. Nested primer 1: GGAGCGCCGCCTGGAG (*TMPRSS2* exon 1 nest Forward) and nested primer 2: CCATATTCTTTCACCGCCCACTCC (*ERG* exon 6 nest Reverse). This master mix was aliquoted and 0.25μL of the primary PCR was added. PCR conditions were as above but with a 66°C annealing temperature instead of 68°C. The resulting amplification products of the primary PCR and the nest PCR were run in adjacent wells on a 2% agarose gel with a 100bp DNA ladder (New England Biolabs (N3231L)) to determine product sizes (Table 2.1) and thus infer which of the *TMPRSS2:ERG* fusions were present in each sample.

Table 2.1 PCR product sizes for *TMPRSS2*_exon1 (T1) and *ERG*_ex6 (E6) PCR primers (nests are 139bp smaller than primaries)

	Primary	*Nest*
T1/E4	***596***	***457***
T1/E5	***379***	***240***
T1/E6	***227***	***88***
T2/E2	***856***	***717***
T1/E3, - , 5, 6	***465***	***326***
T1/E2, 3, 4, -, 6	***661***	***522***
T2/E5	***450***	***311***
T3/E4	***891***	***752***
T4/E5	***760***	***621***
T5/E4	***1098***	***959***

2.2 Clinical Data Collection

Note: I completed part of the clinical data collection.

Clinical data was collected for NNUH samples from a number of different NHS databases such as ICE (the NNUH database), Somerset (the NNUH cancer database) and also from the patient's forms completed for the study within the clinic. Information from the patient's forms were manually typed into an Excel sheet and uploaded to a pseudo-anonymised online database for the Movember project. A clinical NHS colleague and I updated and checked over clinical information for the majority of the samples, this included information

such as age, initial PSA reading, Gleason score and further biopsy information, scan conclusions, prostate volume, family history, health altering habits, general health and current medications as well as subsequent information (ensuing PSA readings for example). Clinical data for samples from other centres were provided by them and uploaded into the Movember database.

2.3 NanoString Pre-processing

2.3.1 Normalisation

The NanoString output data file provides the nCount data for 6 spiked non-human positive control probes and 8 non-human negative control probes for each of the samples being analysed. The six positive control probes matched to spiked-in RNAs and was used to calculate a normalisation factor (NF): the average nCount for each samples' positive controls were calculated and this number was divided by the sum of all samples' averages. Each nCount value was then multiplied by the sample-specific NF. This results in a shift of all samples so that the means of the positive controls was identical across samples. Background correction and background subtraction using the negative controls was found to be inappropriate for this data.

2.1.1 Normalisation by *KLK2* and *KLK3*

Normalisation using *KLK3* and *KLK2*, separately, was conducted as follows. For *KLK2*, a ratio was determined (Equation 2.1) and then applied to the data, this data was referred to as *KLK2* ratio data.

Equation 2.1 KLK2 ratio normalisation, similar to the normalisation of PCA3 by KLK3 in the PCA3 test

$$\left(\frac{(x_{ij})}{(\bar{x}_{KLK2})}\right) * 1000$$

Additionally, for *KLK2* and *KLK3* an adjustment normalisation was conducted using the median and IQR (Equation 2.2). For this data, any samples observing low *KLK2* or *KLK3* levels, respectively were removed from the data set prior to adjustment. The threshold for "low" expression was determined using a density plot and the Brent method to find the minima of the curve. For *KLK2*, and *KLK3* the same nineteen samples were identified and removed for low expression. As well as removing low Kallikrein expression samples, six CB samples that had high *TMRPSS2:ERG* expression were also identified and removed. Samples with high *TMPRSS2:ERG* expression were again identified through density plots and the Brent method.

Equation 2.2 Kallikrein adjustment of data using median and IQR. Where *i* is the sample and *j* is the transcript.

$$\left(x_{ij} - \frac{median(x_j)}{IQR(x_j)}\right) * IQR(KLK) + median(KLK)$$

2.3.2 Normalisation by housekeeping genes

Five previously identified housekeeping transcripts were included in the NanoString1 pilot study: *ALAS1*, *B2M*, *HPRT, GAPDH*, and *TBP*. *RPLP2* was added in NanoString2. Tukey tests (section 2.4.7) were used to identify transcripts that were not significantly different between any clinical category ($p < 0.05$). ANOVA (section 2.4.6), variance and IQR (section 2.4.8), and Pearson's correlation (section 2.4.3) were also utilised, to identify novel transcripts to use for housekeeping purposes. In NanoString2 EV data, *RPLP2* and *GAPDH* were selected to normalise the data, whereas for the NanoString2 cell data, *RPLP2* and *TWWAST1* were selected.

For each sample, the mean of the two transcripts was calculated, as well as the mean of those means across samples. Each sample was then multiplied by a normalisation factor (ratio of the mean of means with the individual sample mean).

2.3.2 NanoStringNorm and NanoString QC Pro

The quality of the normalisation was evaluated using the R packages NanoStringNorm[172] and NanoString QC Pro[173].

2.3.2.1 NanoStringNorm

NanoStringNorm[174] investigates the normalisation of the data as well as identification of samples and transcripts that were outliers. The first test performed by NanoStringNorm was to plot the mean verses the standard deviation (SD) with a Loess curve of best fit. Positive controls and potential housekeeping probes should have high means and low SD, whilst negative controls should have low means and low SD. Batch effects and potential confounding were also tested for using sample summary features, including mean, SD, proportion of missing (0 counts) or positive/negative control counts. These features were plotted independently by NanoStringNorm, where the location of the point relative to the horizontal line shows how different it was from the others and the size of the (green) dot was proportional to the level of its significance. Orange dots were not significant. Potential influencing outliers were identified by looking into the normalisation factors: if the normalisation parameters extended beyond 100% difference from the mean, it was flagged as a potential outlier.

2.3.2.2 NanoString QC Pro

NanoStringQCPro[173] (an R library) was conducted to check the quality control of the NanoString data, specifically looking at the control probe metrics and count probe metrics (similar to NanoStringNorm) but additionally looks at other metrics. The field of view (FOV) was a discrete area of each lane being imaged by the ncounter® digital analyser. Within the FOVs, bubbles and insufficient oiling can make unsuccessful imaging attempts. A low ratio between successful and unsuccessful attempts can be indicative of low imaging performance. NanoStringQCPro highlights any samples with less than 80% successful imaging attempts. If the binding density was too high in a sample, there can be overlapping

of the barcodes, which leads to errors in correctly imaging the number of probes. According to NanoString a binding density of less than 0.5 and higher than 2.25 can lead to these errors. NanoStringQCPro flags samples that have binding densities outside of the recommended thresholds.

Positive controls were spiked into each NanoString experiment, they should show linearity with positive control A having highest values, down to positive control F with the lowest. Control range and Interquartile range (IQR) were also examined. The counts were also examined; any samples with unusually low counts were flagged using cutoffByMMAD to identify the threshold. This was based on the median of the data and the upper and lower thresholds were counted using median (x) – d * mad (x) and median (x) + d * mad (x), respectively (where d was a scalar).

2.3.3 Log and Square-root Transformation

Sometimes, to obtain a more normal distribution of the data, it can be useful to transform the data. Many inferential statistical tests assume that the data was of normal distribution and violating these assumptions can cause an increase in both type 1 and type 2 errors. For regression-based models, the relationship between input and output variables should be approximately linear (so the input variables have a normal distribution and the output has constant variance, thus the variance of output variables was independent of input variables). Two transformations that have been used in this project were log transforming the data and square root transformation. Square root transformation has been shown to be appropriate for transforming count data[175]. However, square-root transformation of data has its drawbacks; if your data contains both values greater than 1 and values between 0 and 1, these two types of values will be treated differently.

2.3.4 ComBat

Batch effects occur in many high-throughput experiments, they can be caused due to laboratory conditions, reagent lots and personnel differences. ComBat was determined to be

the best performing of six methods for removing batch effects in microarray data[176]. The ComBat function is an empirical Bayes method, where location and scale model adjustments are made as follows:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

Where α_g is the overall gene expression, X is a design matrix for sample conditions, β_g is the vector of regression coefficients corresponding to X, $\varepsilon_{ijg}$is the error terms (which are assumed to follow a normal distribution) with expected value of 0 and variance σ_g^2. The γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for gene g.

The adjusted data is then given by:

$$Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\widehat{\beta_g} - \widehat{\gamma_{ig}}}{\widehat{\delta_{ig}}} + \hat{\alpha}_g + X\beta_g$$

Where $\hat{\alpha}_g, \widehat{\beta_{g,}}\gamma_{ig}$ and δ_{ig} are estimators for the parameters $\alpha_g, \beta_g, \gamma_{ig}$ and δ_{ig} based on the above model. The ComBat function of the sva R package was used with R version 3.2.1.

2.4 Basic Statistical Tests

Basic statistical functions used and described below were part of the R stats package and were used with default settings, under R version 3.2.1.

2.4.1 Mann-Whitney U test (Wilcoxon Rank Sum test)

The Mann-Whitney U test was a non-parametric log-rank test capable of identifying differential expression of genes between two different states, for example, cancer vs. non-cancer. The test works by assigning a rank to each individual value from 1 to *n* (where *n* was the number of samples) and 1 was assigned to the smallest value. It then compares the sum of the ranks in the first group (R_1) to the expected sum of the ranks given the sample size of group 1 and then the sum of the ranks in the second group (R_2) was compared to the expected sum of the ranks given the sample size of group 2 (these values were considered

U_1 and U_2, respectively, see Equation 2.3). The smallest of these numbers was then used to calculate the significance.

Equation 2.3 Mann-Whitney U test

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

One advantage was that more accurate results than a *t*-test were obtained when used on data with a non-normal distribution[177].

2.4.2 Spearman's Correlation

Spearman's rank correlation coefficient was a non-parametric test of statistical dependence between two sets of data, most commonly two variables. It measures the relationship between these variables providing a value between -1 and 1, where 1 or -1 means complete dependence, whilst 0 means that no dependence was observed. Spearman's correlation uses the rank of the variables rather than exact values (as used in Pearson's correlation). The covariance of these ranks was divided by the standard deviation of the ranks also (Equation 2.4). Here, d_i was the difference in ranks for variables x and y, r_s was the notation for the coefficient for a sample statistic and n was the number of samples. Spearman's correlation was preferred over Pearson's correlation typically when one of the variables was ordinal and the other was continuous or if the relationship was non-linear[178].

Equation 2.4 Spearman's Correlation

$$r_s = 1 - \frac{\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

2.4.3 Pearson's Correlation

Pearson's product moment correlation coefficient was calculated in a very similar method to that of Spearman's correlation in that the covariance of the two variables was divided by the standard deviation of those variables. The key difference was that the exact values were used instead of their ranks (Equation 2.5). Here the correlation coefficient was noted by r and x_i and y_i were the *i*th individuals of x and y variables. Pearson's correlation was

typically used when both variables were continuous, normally distributed (as extreme values can bias the strength of a relationship), and the tested relationship was linear[178].

Equation 2.5 Pearson's Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}}$$

2.4.4 Pearson's Chi-Squared

This statistical procedure was typically used to identify if the frequency distribution of events was independent from the labels assigned to the event. It can be used to suggest if two groups of variables were related or not, for example in clustering, to see if the clusters were significantly related to the clinical category, a frequency distribution table can be produced. To calculate what frequencies were likely to occur from chance, the number of observations (O_{ij}) was divided by the number of cells in the table, this gives what was known as the theoretical frequency (E_{ij}). This can then be used to calculate the test statistic (Equation 2.6) and with *n*-1 degrees of freedom, the *p*-value can also be determined[179].

Equation 2.6 Pearson's Chi Square test

$$\mathrm{X}^2 = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

2.4.5 Welch *t*-test

The Welch *t*-test was a parametric test to measure how the means and variance of two groups differ in normally distributed data, where the variances of the two populations were assumed to be non-equal. The mean of the data points in-group A and B, along with the squared sums $(\sum x)^2$ and also the sum of the squares $\sum(x^2)$ were used to calculate a *t* value (Equation 2.7). This provides a t value, which was comparable to values designated using different degrees of freedom (dependent on the number of samples in your two groups of data). Combining the *t*-value with the relevant degrees of freedom (sum of the variables in each group minus 2) yields a *p*-value.

Equation 2.7 The Welch *t*-test, comparing the mean and standard deviation between two sets of data to conclude if they were significantly different from each other

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\left[\frac{\left(\sum A^2 - \frac{(\sum A)^2}{n_A}\right) + \left(\sum B^2 - \frac{(\sum B)^2}{n_B}\right)}{n_A + n_B - 2}\right] \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}}$$

2.4.6 ANOVA – Analysis of Variance

Analysis of Variance (ANOVA) was a procedure used to analyse the differences among group means. In this work, it has a similar function to the *t*-test but allows the analysis of more than two subgroups. Firstly, the mean sum of squares within each group, MSS_w (Equation 2.8) and the mean sum of squares between the groups, MSS_B (Equation 2.9), was calculated. The ratio of these then provides the test statistic, $F\left(F = \frac{MSS_B}{MSS_w}\right)$. Combining the *F* statistic with the degrees of freedom allows a *p*-value of significance to be determined. There were two degrees of freedom to calculate in ANOVA, $df_B = k\text{-}1$ and $df_w = n\text{-}k$, where *n* was the total number of samples and *k* was the total number of groups.

Equation 2.8 Mean sum of squares within each group of data, where *n* was the total number of samples, *k* was the total number of groups, *g* was the value and *G* was all of the values across all groups.

$$MSS_w = \frac{\sum g \in G\ (x - \bar{\bar{x}}_g)^2}{n - k}$$

Equation 2.9 Mean sum of squares between each group of data, where *n* was the total number of samples, *k* was the total number of groups, *g* was the value and *G* was all of the values across all groups and n_g was the number in each group.

$$MSS_B = \frac{\sum g \in G\ n_g(\bar{\bar{x}}_g - \bar{\bar{x}}_G)^2}{k - n}$$

2.4.7 Tukey test

The Tukey test allows us to make multiple mean comparisons within the data with just a one step procedure (Equation 2.10). It was essentially a *t*-test that takes into consideration multiple testing. By assigning known groups to the data one can infer if these groups have significantly different means from all other groups within the data. Pairwise comparisons of all the possible groups' means were made and the difference was compared to the standard error.

Equation 2.10 The Tukey test, where Y_A was the greater of the two means, Y_b was the smaller of the two means and SE was the standard error of the sum of the means.

$$q_s = \frac{Y_A - Y_B}{SE}$$

2.4.8 Kruskal-Wallis

The Kruskal-Wallis test is a one-way ANOVA on ranks, it is essentially an extension to the Mann Whitney U test. Similarly to Mann Whitney U, the Kruskal Wallis test uses ranks, and therefore, is a non-parametric test useful for non-normally distributed data. Additionally, similarly to ANOVA, the Kruskal-Wallis test can allow testing of >2 categories of data.

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

where n_i is the number of observations in group i, g is the number of groups, r_{ij} is the rank of observation j from group i, N is the total number of observations across all groups, $\bar{r}_i$ is the average rank of all observations in group i and $\bar{r}$ is the average of all the r_{ij}.

A p-value can then be approximated from H from the table of X^2 distributions and the degrees of freedom (g-1). The function kruskal.test from the stats R package was used in R version 3.2.1.

2.4.9 Variance and IQR

Variance of a dataset can be measured as the sum of the squared distance of the data points from their mean. The IQR of the data was the lower quartile (the data point at 25%) subtracted from the upper quartile (the data point at 75%). The IQR was useful when data was not normally distributed.

2.4.10 Log rank Test

The log rank test is used to compare the survival experience of two different experimental statuses. It tests for the null hypothesis that there is no difference for the populations for the probability of an event at any time period, unlike survival curves, where a comparison at arbitrary time points are given.

For each time the number of events in each group are calculated and compared to the number expected if the null hypothesis were to be true. For each group the test statistic is calculated using $(O-E)^2/E$, where O is the number of observed events and E is the number of expected events. The comparison is completed using a X^2 test (Section 2.4.4) and from the X^2 distribution tables, a p-value can be provided allowing acceptance or rejection of the null hypothesis[180].

The log rank test has advantages such that the whole follow up period is utilised, and no information about the shape of the survival curve of distribution of survival times is required.

The log rank test was completed using the survdiff function of the survival R package[26].

2.4.11 Shapiro-Wilk

The Shapiro-Wilk test was used to determine if a sample came from a normally distributed population. The null hypothesis was that the data was from a normally distributed population and so was rejected if the *p*-value was less than the chosen alpha value (typically 0.05). Equation 2.11 was utilised to determine the W statistic, where $x_{(i)}$ was the *i*th smallest number in the sample (the *i*th order statistic) and a_i were the constants derived from the covariance matrix of the order statistics[181]. The algorithm used in R also has the ability to calculate a *p*-value from W[182]. This was used with standard settings, under R version 3.2.1 for all Shapiro-Wilk testing.

Equation 2.11 The Shapiro-Wilk test, where $x_{(i)}$ was the ith smallest number in the sample (the *i*th order statistic) and a_i were the constants derived from the covariance matrix of the order statistics

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2.4.12 Brent

The Brent method is an algorithm that combines three root-finding algorithms. It is as quick and more reliable than most other bisection methods. It is an iterative method that moves inwards from two points known to be on a quadratic curve until the root that provides the optimal bisection is discovered. The function optim from the stats package in R was used, with the argument method=Brent, for the optimal bisection of density plots.

2.4.13 Benjamini – Hochberg Multiple Testing Correction

In order to limit false discovery rates when completing multiple tests, multiple testing correction is completed. This particularly is useful for removing false positive hits, but has the trade-off of creating false negatives. The Benjamini-Hochberg method is a widely used procedure when completing multiple statistical tests, like testing each gene or probe between two groups. The correction starts by assigning a rank 1 to N, where 1 is assigned to the smallest p-value. Each p-value is then given a Benjamini-Hochberg critical value, using the formula (i/m)Q, where i is the assigned rank, m is the total number of tests and Q is the false discovery rate (chosen by the user). A comparison between the p-value and it's critical value is then made by finding the largest p-value that is smaller than its critical value. Any p-value above this is then considered significant by the Benjamini-Hochberg method, and a new p-value is assigned.

The function p.adjust from the R package stats was used the Benjamini-Hochberg method passed as an argument.

2.4.14 Receiver Operator Characteristics (ROC)

ROC curves were a graphical plot to show the diagnostic ability of a classifier system as its discrimination threshold was varied. There was a trade-off between true predicted positives (sensitivity) and true predicted negatives (specificity) in the outcomes as this threshold was

varied. ROC aims to identify the best threshold to give the best balance between the specificity and sensitivity. Each ROC also gives an AUC (area under the curve), which was a value between 0 and 1. Where 1 was a perfect model with all positives classed as positive and all negatives classed as negatives, and 0 shows there was no predictive value of the model at all. Generally, an AUC above 0.8 was valued as good. Two packages were used to produce ROC curves. For the HT chapter, ROC was performed using ROC, also part of the ROC bioconductor package[183]. Alternatively, to analyse the performance of the models built in the NanoString2 chapter the ROC function of the epi package[184] was used, this calculated the sensitivity and specificity as well as the AUC.

2.5 Clustering

2.5.1 Principal Component Analysis (PCA)

PCA allows the visualisation of the maximum variability of a data set in two-dimensions. For a simple explanation, imagine there were ten samples and five genes and a graph was drawn with five axes, with each of the ten samples placed at the point that represents their value along each axis. Then identify a line that goes through as many of the samples as possible with the highest variation, that imaginary line was the first principal component. The second principal component was the line with the second highest variation and so on. Therefore, the majority of the variation of the data was found in the first two principal components and a 2D plot of these was enough to identify the biggest differences in samples.

This unsupervised mathematical procedure aiming to reduce dimensions of data works using a coordination transformation from the original data space to “eigenspace” using eigenvectors and eigenvalues of a matrix[185]. The first step was to calculate a covariance matrix of the data, with the aim to reduce redundancy and maximise variance. From the covariance matrix, which was used to measure how much the dimensions vary from the

mean, the eigenvalues and eigenvectors can be determined. The covariance of two variables tells was a measure of how they vary together (Equation 2.12). Once the eigenvectors and the eigenvalues have been determined the eigenvalues can be sorted in descending size order.

Equation 2.12 PCA covariance equation

$$cov(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})\,(y_i - \bar{y})}{N}$$

2.5.2 Hierarchical Clustering

In this work, the commonly used UPGMA ("unweighted pair-group method using arithmetic averages") method of hierarchical clustering was used. The highest similarity (or smallest distance) was used to identify the next two clusters to be merged. The distance of each sample to members of a cluster were computed with equal weights and the similarity or distance matrix was produced. This was updated and reduced at each computation, as samples/clusters were combined, allowing clustering to proceed by agglomeration as the similarity criterion was relaxed[186].

2.5.2.1 Pvclust

Pvclust[187] was a bootstrapping method that calculates the *p*-value for each cluster in a hierarchical clustering dendogram object through the application for bootstrap resampling; clusters with significant AU *p*-values were shown with a red box.

2.5.3 *k*-means Clustering

k-means clustering aims to separate points into *k*-clusters so that the within clusters sum of squares was minimalized by seeking local optima so that moving of a point from one cluster to another will not reduce the sum of squares (Equation 2.13)[188]:

Equation 2.13 Optimal local within cluster sum of squares. $x_i^{(j)}$ was the data point and c_j was the centroid, where *i* was a data point in cluster *j*. *k* was the number of clusters and n_j was the number of samples in cluster j.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j \right\|^2$$

Initially, centroids were arbitrarily picked each point was assigned to the closest centroid in Euclidean distance. Then the centroid was adjusted to the new cluster mean, the samples were reassigned to the closest centroid and this was repeated until convergence was reached. Convergence was when no observations can change the clusters when added and centroids were subsequently redefined[189]. The advantages of *k*-means clustering include speed and simplicity, whilst disadvantages include differing results per run due to the random starting centroid points and an unknown input value for *k*[190]. In this project, the optimal number of clusters was determined using the Bioconductor function NbClust[191], which uses 30 metrics including the Silhouette, Dunn and Davies-Bouldin Indices (section 2.5.4).

2.5.4 Silhouette, Dunn and Davies-Bouldin Indices

Three of these main techniques used for comparing how well data was clustering were the Silhouette, Dunn and Davies-Bouldin Indices.

The silhouette index compares the mean distance of a point to the others in it's cluster and then other clusters. It provides an index value between -1 and 1, where 1 was an indication that the point belongs to the correct cluster and -1 means it does not.

The Dunn index was the minimum distance of points between two different clusters divided by the maximum distance of points within a cluster for each cluster. Here, a larger value was representative of good clustering.

The Davies-Bouldin index takes the mean distance of the points within a cluster from their Barycentre and then divides this by the distance between the Barycentres and so a smaller value was an indication of good clustering.

2.5.5 Latent Process Decomposition

Latent Process Decomposition (LPD)[192] is a hierarchical Bayesian (probabilistic) model that was designed for the clustering of microarray data and thus can also be used with other forms of count data. It estimates the most probable/optimal number of clusters, and determines the probability of a sample belonging to each cluster, rather than membership of a cluster being assigned. This was important as samples were often heterogeneous made up of cells from different clones of cancer. Also different biological processes often work together to influence expression levels.

LPD makes the assumption that a sample's expression was determined by a series of processes. Each process has an associated expression profile which was determined by the algorithm. A sample's expression profile was then de-convoluted in to these process expression profiles. For example, Gene A has expression of *n* genes similar to the expression of the genes that make up the signature of process 1, and expression of *m* genes was similar to the expression of genes that make up the signature of process 2. *n* genes were of a higher similarity to process 1 than *m* genes were to process 2 and so max likelihood was higher for process 1; Gene A has 0.78 for process 1 but still 0.22 for process 2, etc.. So it has some similarity through some genes to process 2 but has more similarity through more genes to process 1.

The first step of LPD was to estimate the most probable number of clusters or "processes" using the maximum likelihood solution and a uniform prior. A uniform prior was a probability assumption with limited knowledge. E.g. a ball under 3 cups A, B, C has probability prior of p(A) = p(B) = p(C) = 1/3, where changing the order of the probabilities of the cups makes no change to the prediction. In the final model, a prior was defined to avoid over fitting by penalizing over complex. The parameter (sigma) in this prior was estimated next through cross-validation.

After these parameters were defined, the final solution was obtained by iteratively updating various parameter values of the Dirichlet distribution (a collection of multivariate,

continuous probability distributions that was a more generalized version of beta distributions) modelling expression. Process mean values were initialized to the mean expression across the data set for each gene, whilst variances were set to the variance of their respective genes[192].

2.6 Model Optimisation

Predictive models were a supervised learning method, which has been applied to both the NanoString1 and NanoSting2 data. For the NanoString2 data, it was divided into training, and test sets for a more robust and accurate model evaluation. A number of different models and modelling techniques were applied.

2.6.1 GLM: Generalised Linear Model

There are two important aspects of GLM[193]: General and Linear. Linear because the underlying equation was that of a straight line: $Y = \beta_0 + \beta_1 X_1$. In this example Y was the predicted or response variable, whilst X was a single predictor or explanatory variable. β_0 was the y-intercept and was constant, whilst β_1 was the slope or weight of variable X_1. General because the equation was able to handle multiple explanatory (X) variables e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Any control variables may be included and if so, should precede the explanatory variable of interest within the equation, in general practice.

The explanatory variables may be numerical and continuous or binomial/factorial with levels. The GLM generalised linear regression by allowing the linear model to be related to the response variable via a link function allowing the modelling of binary response variables through logistic regression and ordinal variables through proportional odds models.

GLM was performed as an initial step in identifying probes that were significant for predicting clinical category (CB vs. Ca, CB vs HR-Ca, CB-L-I-H trend and CB-Ca-Advanced Ca trend) within all of the data (NanoString1) and within the training data (NanoString2). Significant probe lists were then shrunk and selected using techniques such

as Lasso, Step and Random Forest (section 2.6.2, section 2.6.4, section 2.6.3, respectively) and then these condensed lists were then used to build a final model. These models were then tested upon the test data (NanoString2). The R function glm from the stats package was used for logistic regression models and the polr function from the MASS package for proportional odds models. These were used with the R version 3.2.1.

2.6.2 Lasso

Least absolute shrinkage and selection operator (LASSO) was a regression method that was capable of performing selection and regularisation in order to improve both interpretability and prediction accuracy of statistical models, respectively. A constraint was applied to which the sum of the absolute value of the regression coefficients must be less than. This forces some of the coefficients to be set to zero, allowing these covariates to be disregarded from the optimal statistical model. Thus allowing both subset selection and shrinking large regression coefficients so as to reduce over fitting[194]. Over fitting of models can be problematic because these models tend to have poor predictability and can be over responsive to minor fluctuations within the test data set. Lasso can be easily applied to a variety of statistical models including generalised linear models and proportional hazard models, amongst others.

2.6.3 Random Forest

Random forest[195] (RF) was an ensemble method that was a combination of tree predictors (weak learners) such that each tree was built using a sample set constructed by random selection replacement (bootstrapping). Once built the result of the model was the combination of the results of all trees (votes for binary outcomes and mean for continuous outcomes). The random forest function (from the random forest package) was used for classification and for regression models.

Each decision tree was built by taking the bootstrap data and repeatedly separating it at nodes. At each node a small subset of, *m*, variables were chosen at random, and the combination that optimises the split, according to some objective function, was found. At the next node another *m* variables were chosen and the same method was performed. *m* was generally set at $\sqrt{p}$ or $\frac{p}{3}$, where p was the number of variables. As the number of trees increases, the generalisation error of the forest converges.

The importance of variables in the model were assessed in two ways. Internal out-of-bag (OOB) estimates were used to judge the quality of the model. OOB was the average error calculated for each variable from the trees that do not contain that specific variable in their respective bootstrap sample. The error was calculated using the misclassification rate of the subjects. These estimates were produced using a single run of a forest with 1,000 trees and no test set. Variables with large mean decrease in accuracy or OOB were more important for classification of the data. Additionally, a Gini coefficient was also used to assess importance. This was a measure of how each variable contributes to the homogeneity of the nodes and leaves in the RF. Each time a variable was used to split a node, the Gini Coefficient for the child nodes were calculated and compared to the original nodes coefficient. The coefficient can be between 0 (homogenous) and 1 (heterogeneous). These changes in Gini were summed and normalised for each variable. Again, variables that were more important have a higher mean decrease in Gini.

Random forest was applicable to regression. Mean squares error was usually used to determine error rate when using random forest with regression. MSE was the mean (divided by *n* (number of data points)) of the squares of the errors[196].

Random forest (from the random forest package) was used for classification and for regression.

2.6.4 Step for feature selection

StepAIC was a function of the MASS package in R[197]. It was an automated model selection technique that takes a model and inserts or removes each variable and assess the model quality using the AIC – Akaike Information Criteria (Equation 2.14). The model with a smallest AIC was selected as the optimal and then this model was fused in the next step, this was repeated until no further improvements in AIC were observed. StepAIC can be run forwards (where you begin with all variables and remove them), backwards (where you begin with a small number of variables and add them) or both (where variables were added or removed as required)[198].

Equation 2.14 shows how to calculate AIC. Where the model with the lowest AIC was deemed optimal. Where k was the number of parameters and L was the maximum value of the likelihood function for the model.

$$AIC = 2k - 2\,log\,(\hat{L})$$

2.7 Pathway Analysis

2.7.1 DAVID

DAVID was the Database for Annotation, Visualization and Integrated Discovery, it was a gene functional classification tool. It was a web-based tool whereby you submit a list of transcripts of interest and DAVID classifies the list into functional related gene groups, ranks the importance of the discovered gene groups (dgg) and summarises the major biology of the dgg[199]. DAVID was used to identify if there were any interesting biological functions of the transcripts identified as significant.

2.8 Survival Analysis Tools

Survival analysis was the analysis of data where the response variable was the time to an event, for example to death or as in our case time to failure. Individuals that fail after the end of the study at some point in the future were known to be censored. Survival analysis

tools were used to identify if there were transcripts capable of predicting relapse to hormone therapy (HT) prior to a two-year period. The non-failures were said to be censored as after the last follow up date you don't know if they have failed or not[200].

2.8.1 Kaplan Meier (KM) Curves

The KM survival distribution was a discrete stepped survivorship curve, which gains information as each event (failure) occurs. There were two variables at any time point on the KM (Equation 2.15); those that have failed, $d(t_i)$ and those at risk of failing, $r(t_i)$ and this produces a step at each failure.

Equation 2.15 The KM function.

$$\hat{S}_{KM} = \prod_{t_i<t} \frac{r(t_i) - d(t_i)}{r(t_i)}$$

Censored points were denoted by a + on KM plots. Kaplan Meier plots were created using the survfit() function, specifying type=("kaplan-meier") from the survival package[201] and ggsurv() of GGally package[202], on R version 3.2.1. Dichotomised high/low expression levels were determined for each probe using *k*-means clustering and *k*=2 (section 2.5.3).

2.8.2 Cox Proportional Hazard

Cox Proportional hazard model was the most commonly used regression model for survival data. It assumes the hazard was of the form $\lambda(t; Z_i) = \lambda_0(t) r_i(t)$, where $Z_i(t)$ was the set of explanatory variables for individual i at time t. The risk score for individual i was $r_i(t) = e^{\beta Z_i(t)}$, where β was a vector of parameters from the linear predictor $\lambda_0(t)$, which was an unspecified baseline hazard function that will cancel in due course. It guarantees that λ was positive for any regression model. Hazard was the instantaneous risk of failure, or instantaneous rate of change in the log number of survivors per unit time. Coxph was part of the survival package[201].

3

NanoString Data Analysis 1: The Pilot Study

3.1 Summary

The Movember GAP1 global PCa biomarker initiative has multiple collaborators working on the identification of urinary biomarkers for the risk-stratification of PCa. Our laboratory is specifically interested in the RNA expression changes in PCa that are detectable within urinary cell sediments and extracellular vesicles (EVs). The EV RNA expression pilot study described here had the following aims:

1. Identify if PCa specific transcripts can be detected in urinary EVs

2. Assess whether transcript levels within urinary EVs were able to i) identify PCa per se, ii) distinguish aggressive from indolent PCa

3. Identify if the NanoString system could be applied to Nugen Ovation amplified cDNA (Nanostring probes are strand specific and designed to be applied to mRNA).
4. Identify suitable methods for the analysis of the NanoString data

In the pilot study, expression levels of 57 transcripts were measured in 194 samples using NanoString technology (section 1.6.1). The NanoString technology was able to detect PCa specific markers (section 1.4.6), such as *TMPRSS2:ERG* which was detected in 58% of all PCa samples and in 19% of samples from men with no clinical evidence of PCa (CB). This result, confirmed by RTPCR demonstrated that i) NanoString technology was capable of capturing cDNA amplified by the Nugen Ovation kit; ii) EV mRNA contains PCa-specific transcripts, and iii) the methodology was sensitive enough to identify PCa in men with undiagnosed cancer or HGPIN.

Latent Process Decomposition unsupervised analysis (section 2.5.5), clustered the EV expression data into four groups: LPD groups 1 and 4 were saturated with high-risk and advanced cancers, whilst LPD groups 2 and 3 showed clinical diversity. The majority of the intermediate-risk samples resided within LPD group 2 and most of the CB were in LPD group 3 (section 3.5.5).

Supervised statistical approaches (Mann Whitney U test) determined nine probes significantly differently expressed between PCa (advanced, high-, intermediate- and low-risk) and non-PCa samples (Table 3.19), eleven probes significantly different between high-risk PCa and non-PCa samples (Table 3.20) and six probes between advanced PCa and non-PCa samples (Table 3.21).

Supervised modelling of the data (using generalised linear models (glm) and Lasso for shrinkage (section 2.6.2)) identified three models that distinguished; i) PCa vs. non-PCa with an AUC of 0.937, ii) aggressive PCa vs. non-aggressive PCa with an AUC of 0.852 and iii) advanced PCa vs. benign with an AUC of 0.983.

Twenty-three transcripts were significantly differentially expressed between PCa and non-PCa (Table 3.42), however, only seven were consistently differentially expressed

between the various data-analytic methods used (*DLX1*, *ERG*3', *TMPRSS2:ERG*, *HOXC4*, *ERG5*', *PCA3* and *HPN*). Four transcripts were consistently differentially expressed between aggressive PCa and non-aggressive PCa in various tests (Table 3.43) and two transcripts between advanced and non-PCa (Table 3.44).

These findings highlight that the transcript data collected from urinary EVs in PCa patients comes, at least in part, from the prostate and holds clinically relevant structure.

3.2 Introduction

3.2.1 The Research Gap

Risk stratification is currently based on PSA, Gleason score and T stage. MRI is being phased in, but has been shown to have only 41% specificity in a recent study of low risk patients[203]. Patient clinical pathways would benefit from additional information on their PCa diagnostic and prognostic status. We propose that urine EV mRNA data could provide useful clinical information that could help tailor patients to treatment pathways based on their genetic composition and potentially improve uncertainty over which treatment pathway each patient should be assigned to. The PCA3 test has shown to provide minor improvements to risk stratification but importantly shows the utility of urine in PCa diagnostics and prognostics.

3.2.2 The Pilot Study Aims

The pilot study used NanoString technology to investigate the RNA expression level changes of 57 target transcript sequences within EVs extracted from urinary samples collected at the NNUH as part of the Movember study. The aims of this pilot study were:

a) To identify if the transcript content of urinary EVs contained PCa derived material

b) To identify if transcript levels within urinary EVs are linked to PCa risk stratification

c) To identify if NanoString is a suitable method for obtaining transcript level data from our cDNA samples

d) To identify appropriate methods for analysing NanoString data.

3.2.3 The Probe Targets

The 57 target transcript sequences were selected for the following reasons: i) prostate specific transcripts, ii) transcripts overexpressed in advanced PCa tissue (literature search), iii) suspected housekeeping genes, iv) tissue-specific controls for kidney, bladder and blood.

3.2.4 Risk classification of prostate cancer patients

Patients were placed into clinical risk categories based on D'Amico and NICE criteria: Prostate Cancer Diagnosis and Treatment 2014 guidelines[40]. In addition the intermediate risk patients were subdivided on Gleason (G3+4 Vs. G4+3), as progression rates between these two groups are very different (Table 3.1). The median age and PSA at diagnosis for each clinical category have been recorded (Table 3.2). For some computational analyses, specific risk groups were combined (Table 3.3).

Table 3.1 Classification and Frequency of the sample types based on NICE criteria[40]. The quantity of samples for each clinical group can be seen as well as the clinical description of the group in terms of Gleason score, PSA level and T stage.

Classification: NICE Groupings		
Sample Class	***Description***	***Number of Samples***
Advanced (A)	***Metastatic , PSA>100, and G>8***	***17***
High-risk (H)	***G7 PSA>20***	***50***
Upper Intermediate-risk (UI)	***G4+3 PSA<20***	***19***
Intermediate-risk (I)	***G3+4 PSA<20 and IL= G6 PSA>10***	***53***
Low-risk (L)	***Low G6 PSA<10***	***10***
Abnormal (S)	***High PSA no Bx***	***4***
*CB<1**	***No evidence of Ca and PSA<1***	***18***
*CBn**	***No evidence of Ca and PSA normal to age***	***22***

M_19_5	***Removed for technical failure***	***1***
Total		***194***

*CBN and CB<1 were combined to CB (as there was no significant difference between their expression levels $p > 0.05$: Two sample *t*-test) and UI and I were combined to I (as there was no significant difference between their expression levels $p > 0.05$: Two sample *t*-test).

Table 3.2 Median age and PSA at diagnosis for each clinical category, of samples that are used in subsequent analysis.

Sample Class	Number of Samples	Median Age	Median PSA at Dx
Advanced (A)	17	78	110
High-risk (H)	50	73	27
Upper Intermediate-risk (UI)	19	74	9.55
Intermediate-risk (I)	53	67.5	8.35
Low-risk (L)	10	68	5.95
CBN and CB<1	40	68	1.1

Table 3.3 Sample numbers used in i) 'Cancer', ii) 'Aggression' and iii) 'Extreme' computational analyses.

Group	*Number of Samples*
Cancerous (A, H, I and L) and No Evidence of Cancer (CB)	***Cancerous =149 / CB =40***
Aggressive (A, H) and Non-Aggressive (I, L)	***Aggressive = 67/ Non-Aggressive = 82***
Extremes (A Vs. CB)	***A=17 / CB= 40***

3.3 Data Pre-processing and Technical Variation

3.3.1 Normalisation and Background correction

The NanoString analyses provided data for 57 test probes, and 14 non-human system control probes (6 positive-control probes and 8 negative-control probes) in 194 Nugen Ovation amplified cDNA samples. The 6 positive control probes detected spiked-in control sequences that were used to assess the overall NanoString assay transcript detection efficiency for each sample, and generated a normalisation factor (NF) in the following way: The average nCount for each samples' positive controls was calculated and this number was divided by the sum of all samples averages. Each nCount value

was then multiplied by the sample-specific NF. Background correction was not applied and background subtraction was found not to be appropriate for this data (data not shown).

3.3.2 NanoStringNorm – Quality of Data and its Normalisation

The quality of the normalisation was evaluated using the NanoStringNorm R package (section 2.3.2.1). Other than a few flagged samples (M_14_7, M_19_5, M_36_7, (Table 3.4)), and a few flagged probes (*KLK4*, *GAPDH* and *FOLH1,* (Table 3.5)), the data was of overall good quality. The three probes were flagged due to high mean and/or standard deviation or for *FOLH1* not following the Loess curve of best fit. For *GAPDH*, we predicted similar housekeeping properties as in cell RNA, however that is not what has been observed (section 3.3.5). For *KLK4*, it suggests high expression in the samples with a wide range of signals (considering we have samples across different clinical categories this is expected). For *FOLH1*, the Loess curve of best fit is a non-parametric regression derived curve that is similar to a line of best fit through all of the data. To not follow it simply suggests that this probe is expressed rather differently to the other probes in these samples (again could be due to the range of clinical categories used).

Some cartridges (each cartridge is loaded with twelve samples and then run on the NanoString machine) showed significantly different means and standard deviations in comparison to others in the raw data. The flagged outliers were considered with caution and reviewed further in subsequent analyses

.

Table 3.4 Three samples were flagged by NanoStringNorm.

Samples	*Issues*
M_14_7	***Low sample mean***
M_19_5	***Low sample mean***
M_36_7	***Low sample mean*** ***Normalisation factor flagged as influential outlier***

Table 3.5 Three probes were flagged by NanoStringNorm.

Probes	*Issues*
KLK4	***High mean and SD***
GAPDH	***High mean and SD***
FOLH1	***Doesn't follow Loess curve of best fit***

3.3.3 Experimental and Technical Investigation

3.3.3.1 NanoString Scanners and Cartridges

NanoStringNorm showed significant differences between the mean and standard deviation of the normalised data between cartridges; indicating there might be batch effects on the scanner and cartridge-dependent variables. Scanner and cartridge-dependent variations were therefore examined using Principal component analyses (PCA) (section 2.5.1). PCA did not detect any clustering based on technical artefacts (Figure 3.1A), and there was no significant association between mean expression per sample and either cartridge (Kruskal-Wallis rank sum test: $p = 0.17$, $\chi = 21.21$), or Scanner (Kruskal-Wallis rank sum test: $p = 0.71$, $\chi = 0.14$).

3.3.3.2 RNA Extraction and Amplification

At the beginning of the urine-collection study, protocol optimisation for RNA yield from RNA extractions was conducted (by Marcel Yazbek-Hanna and Rachel Hurst, section 2.1.3), which led to samples from multiple variant protocols being included in the pilot study set. PCA (section 2.5.1) was applied and no clustering was observed due to RNA extraction protocol (Figure 3.1B). There was no significant association between the median expression for each sample and the RNA extraction protocol used (Kruskal-Wallis rank sum test: $p = 0.16$, $\chi^2 = 6.5$).

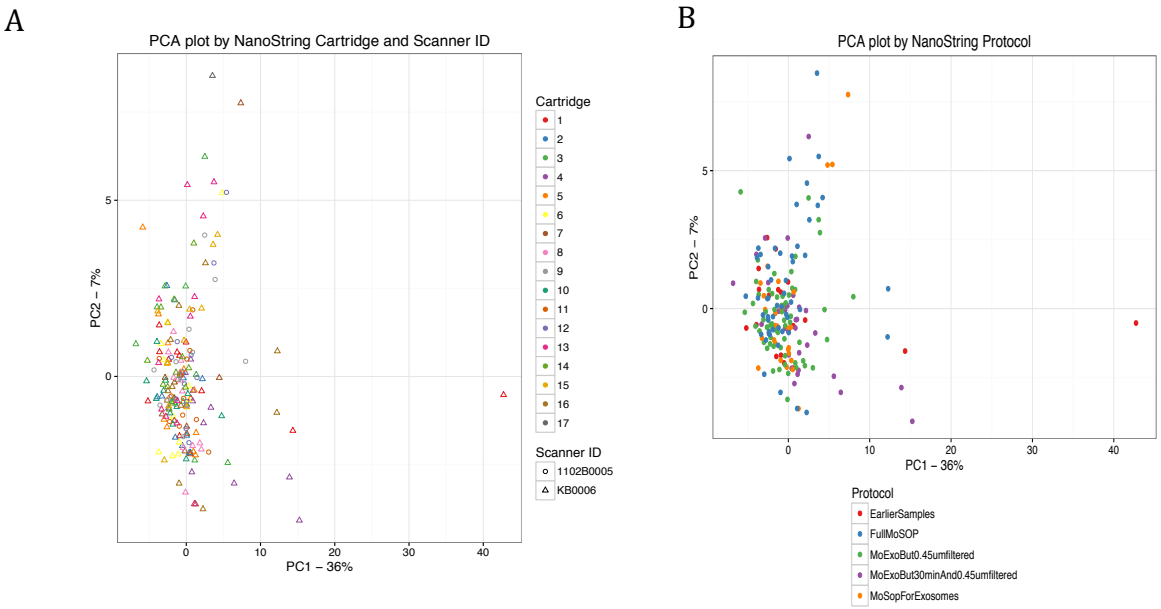
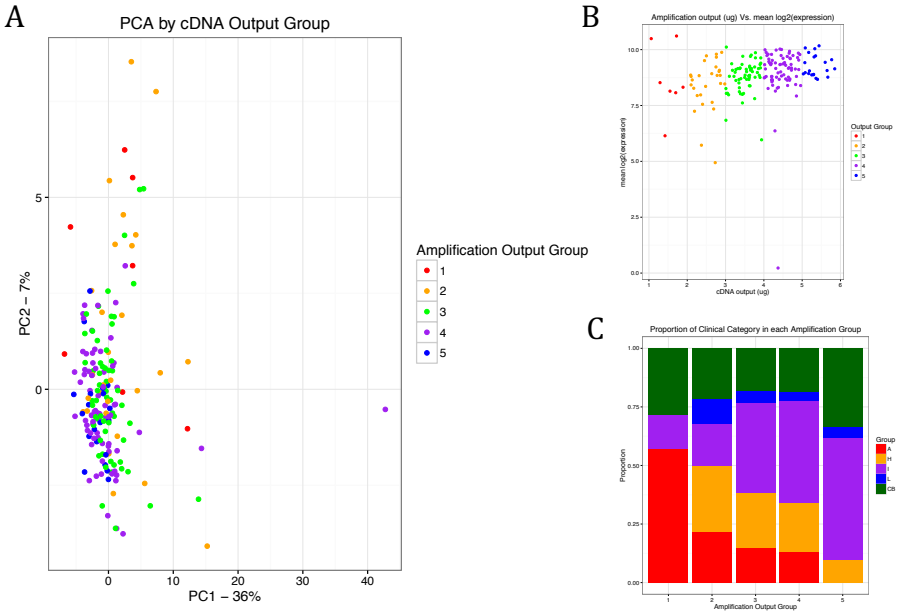


Figure 3.1 To ensure there were no batch issues PCA plots were produced of NanoString loading batches and RNA extraction protocol. A) PCA did not identify any clustering associated with NanoString cartridge or scanner used. Along with the Kruskal-Wallis rank sum results also (Cartridge: $p = 0.17$, Scanner $p = 0.71$), it was deemed there was no batch effect produced by NanoString loading. B) PCA does not identify any clustering associated with RNA extraction protocol used and the Kruskal-Wallis rank sum test was also insignificant ($p = 0.16$). Thus it was deemed that using no filter, a 45µm filter, and a 45µm filter with a 30-minute wait along side the Qiagen micro RNA RNeasy kit using manufactures' protocols made no difference.

Due to the limited amounts of EV RNA harvestable from urine, 15-20ng RNA from each sample was amplified using a Nugen Ovation WTA2 cDNA amplification kit. The amount of cDNA obtained after amplification (in µg) was investigated for clustering affects using PCA (section 2.5.1) and correlation (section 2.4.3). cDNA yields were split into groups; group 1 = 1-2µg, group 2 = 2-3µg, group 3 = 3-4µg, group 4 = 4-5µg, and group 5: >5µg. Mild clustering affects were observed; samples with lower Ovation output had a lot more spread than higher amounts of output (Figure 3.2A) but no significant correlation was found between cDNA yield and median $\log_2$ expression per sample ($p = 0.09$, $r = 0.12$, Pearson's correlation, Figure 3.2B). The distribution of clinical categories within each Amplification yield group was not statistically significant; ($\chi = 26.2$, $p > 0.05$, χ^2 test (section 2.4.4), Figure 3.2C).



Figur
e 3.2 A) Amplification cDNA yield shows mild clustering (cDNA yields were grouped: group 1 = 1-2µg, group 2 = 2-3µg, group 3 = 3-4µg, group 4 = 4-5µg, and >5µg in group 5). B) Amplification cDNA yield shows no influence on sample mean expression C) Amplification cDNA yield shows dependence on clinical category.

3.3.4 Transforming data to a normal distribution and the Shapiro-Wilk test

$\log_2$ and square root transformation (section 2.3.3) was applied to attempt to get the dataset closer to a normal distribution (Figure 3.3). Neither the $\log_2$-transformed, nor the square root transformed, nor the non-transformed data are normally distribution according to the Shapiro-Wilk test (section 2.4.11, Table 3.7, Table 3.8). However, for the majority of the samples (the first 70 and last 70), the *W* statistic is higher for the $\log_2$-transformed data, indicating that the data is closer to a normal distribution than for the other transformations (Table 3.6).

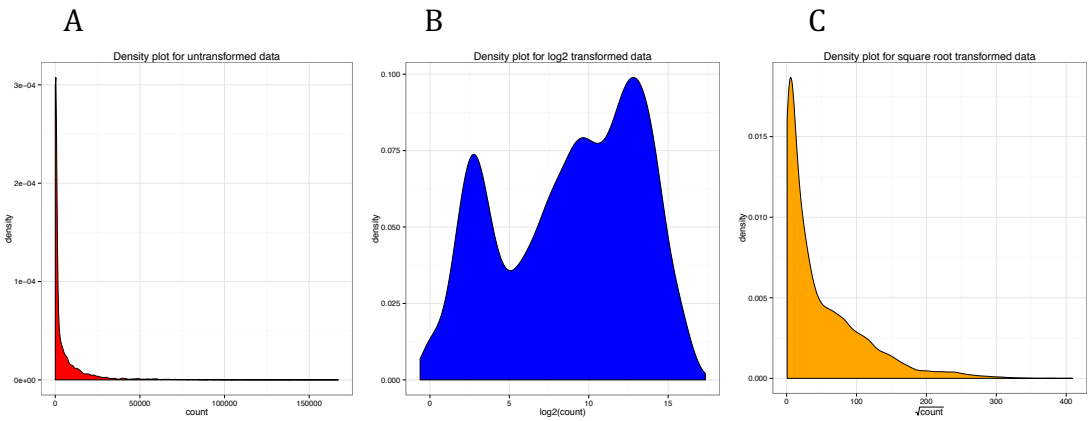


Figure 3.3 Density plots showing the distribution of a) the non-transformed data. B) the log2 transformed data. C) the square root transformed data.

The Shapiro-Wilk test was also applied to ten randomly selected probes in each of the datasets (un-transformed, $\log_2$ transformed and square root transformed) to see how the distribution of some individual probes varied; the majority were not normally distributed. The NanoStringNorm flagged probes (*KLK4*, *GAPDH* and *FOLH1*) had similar results to the other probes. These results led to the use of non-parametric tests wherever possible during analysis. A $\log_2$ transformation was applied so that probe data was closer to a normal distribution, as is standard practice for NanoString data[204].

Table 3.6 Shapiro-Wilk test results on the first 70 and last 70 samples (all probes) for the non-transformed, log_2 transformed and square root transformed datasets.

	Un-transformed			*Log_2 transformation*			*Square root transformation*		
	W	***p-value***	***Normally Distributed***	***W***	***p-value***	***Normally Distributed***	***W***	***p-value***	***Normally Distributed***
The first 70 samples (1-70)	***0.5143***	***$2.2x10^{-16}$***	***No***	***0.9479***	***$2.2x10^{-16}$***	***No***	***0.7937***	***$2.2x10^{-16}$***	***No***
The last 70 samples (124-194)	***0.525***	***$2.2x10^{-16}$***	***No***	***0.9496***	***$2.2x10^{-16}$***	***No***	***0.8105***	***$2.2x10^{-16}$***	***No***

Table 3.7 Shapiro-Wilk test results for 10 randomly selected probes for the non-transformed, log_2 transformed and square root transformed datasets.

	Un-transformed			*Log_2 transformation*			*Square root transformation*		
	W	***p-value***	***Normally Distributed***	***W***	***p-value***	***Normally Distributed***	***W***	***p-value***	***Normally Distributed***
Probe 24	***0.7704***	***$3.886x10^{-16}$***	***No***	***0.6369***	***$2.2x10^{-16}$***	***No***	***0.9421***	***$4.929x10^{-07}$***	***No***
Probe 2	***0.8414***	***$2.847x10^{-13}$***	***No***	***0.5119***	***$2.2x10^{-16}$***	***No***	***0.942***	***$4.883x10^{-07}$***	***No***
Probe 22	***0.7613***	***$2.2x10^{-16}$***	***No***	***0.8579***	***$1.762x10^{-12}$***	***No***	***0.9548***	***$7.663x10^{-06}$***	***No***
Probe 17	***0.9394***	***$2.906x10^{-07}$***	***No***	***0.7784***	***$7.562x10^{-16}$***	***No***	***0.9952***	***0.7955***	***Yes***
Probe 47	***0.9888***	***0.1301***	***Yes***	***0.6097***	***$2.2x10^{-16}$***	***No***	***0.9694***	***0.000306***	***No***
Probe 34	***0.8222***	***$4.011x10^{-14}$***	***No***	***0.7533***	***$2.2x10^{-16}$***	***No***	***0.9652***	***$9.965x10^{-05}$***	***No***
Probe 13	***0.6355***	***$2.2x10^{-16}$***	***No***	***0.9869***	***0.0708***	***Yes***	***0.8741***	***$1.227x10^{-11}$***	***No***
Probe 43	***0.1609***	***$2.2x10^{-16}$***	***No***	***0.885***	***$4.967x10^{-11}$***	***No***	***0.4477***	***$2.2x10^{-16}$***	***No***
Probe 29	***0.9918***	***0.3421***	***Yes***	***0.5435***	***$2.2x10^{-16}$***	***No***	***0.9626***	***$5.107x10^{-05}$***	***No***
Probe 26	***0.9817***	***0.01245***	***No***	***0.5934***	***$2.2x10^{-16}$***	***No***	***0.9573***	***$1.362x10^{-05}$***	***No***

Table 3.8 Shapiro-Wilk test results for the three probes identified by NanoStringNorm as having potential quality issues in the three datasets: non-transformed, log_2 transformed and square root transformed.

	Un-transformed	*Log_2 transformation*	*Square root transformation*

	W	***p-value***	***Normally Distributed***	***W***	***p-value***	***Normally Distributed***	***W***	***p-value***	***Normally Distributed***
KLK4	***0.9918***	***0.3421***	***Yes***	***0.5435***	***$2.2x10^{-16}$***	***No***	***0.9626***	***$5.107x10^{-05}$***	***No***
GAPDH	***0.9713***	***0.0005136***	***No***	***0.4575***	***$2.2x10^{-16}$***	***No***	***0.9635***	***$6.352x10^{-05}$***	***No***
FOLH1	***0.9394***	***$2.906x10^{-07}$***	***No***	***0.7784***	***$7.562x10^{-16}$***	***No***	***0.9952***	***0.7955***	***Yes***

3.3.5 Housekeeping Probes

Five probes (*ALAS1, B2M, GAPDH, HPRT* and *TBP*) were added to the NanoString project to identify housekeeping transcripts (transcripts that remain relatively consistent between samples of different clinical category). Housekeeping transcripts are added so that comparisons between the samples within an expression analysis may be performed accurately. The five transcripts are known housekeeping transcripts in cell mRNA, but there is very little known about EV RNA housekeeping transcripts at present.

There is very little correlation between the six clinical categories (Adv, H, I, L, S, CB) within each housekeeper expression profile (Tukey-ANOVA test, Table 3.1, Figure 3.4); the S clinical group (those with a high PSA but no Bx, $n = 4$) has the most significant differences compared to the other clinical categories; for *ALAS1* comparisons with all other clinical groups and the S group were significant. For *HPRT* two comparisons were significant, one between CB and S and the other between CB and Adv. For *TBP* only one comparison was significant, (between the S group and advanced group). However, there were only four samples in the S group and so the results of the significance test for this group were treated cautiously. Ignoring significant comparisons that included the S group, left only one significant comparison (for the *HPRT* probe between CB and Adv).

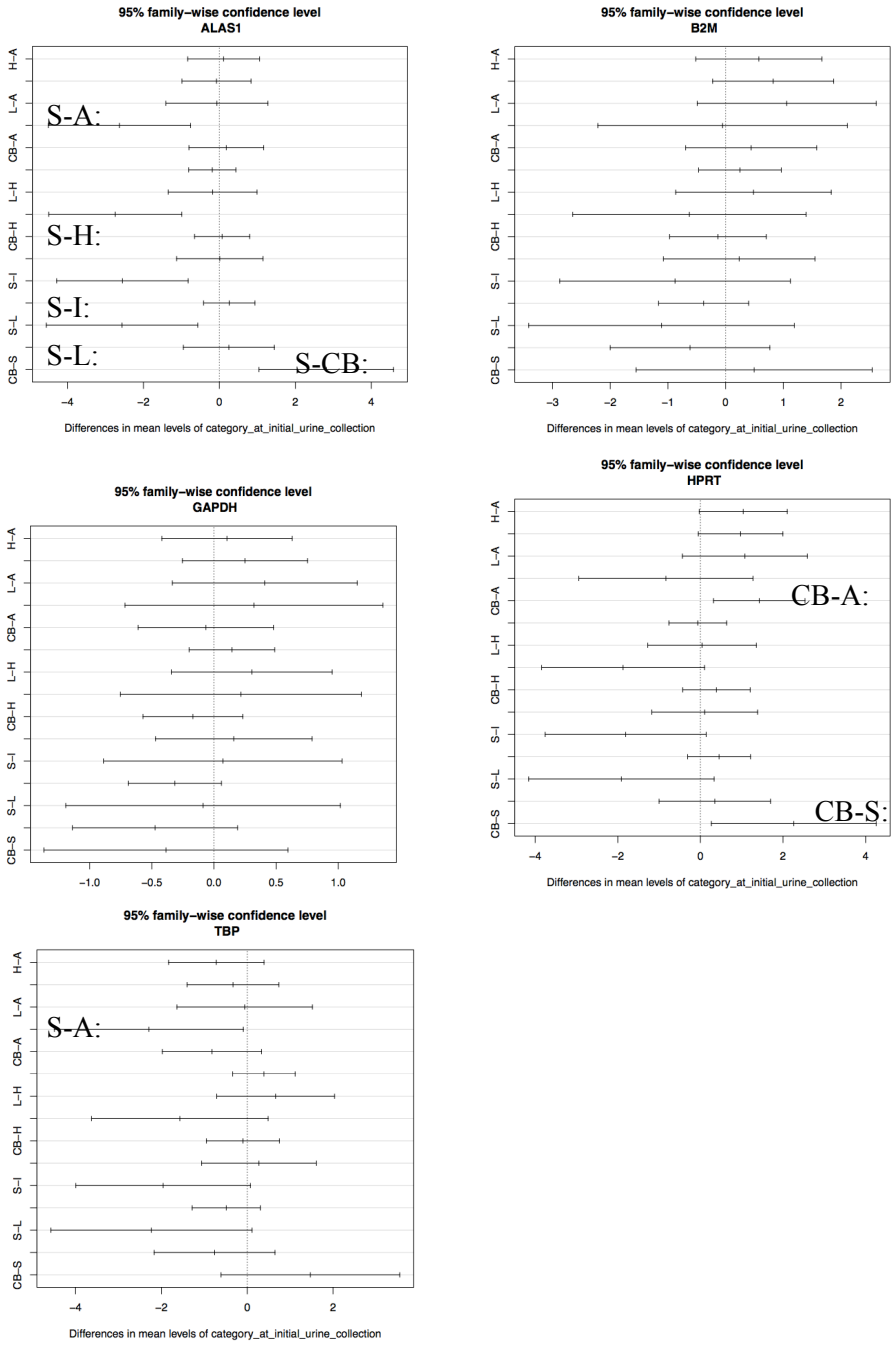


Figure 3.4 Tukey test comparisons of clinical category for housekeeping probes. When the bar does not cross the mid-point of the *x*-axis then the comparison is significant. The Tukey test takes each of the five probes (*ALAS1*, *B2M*, *GAPDH*, *HPRT*, and *TBP*) and detects significant expression differences between the six clinical categories. The significant comparisons with S (high PSA/negative Bx samples) is treated cautiously as there were only *n* = 4 samples within this group. This leaves only one group comparison (CB with Advanced samples in *HPRT*) that showed any significant difference. A good housekeeping probe would be expected to not differ between clinical categories.

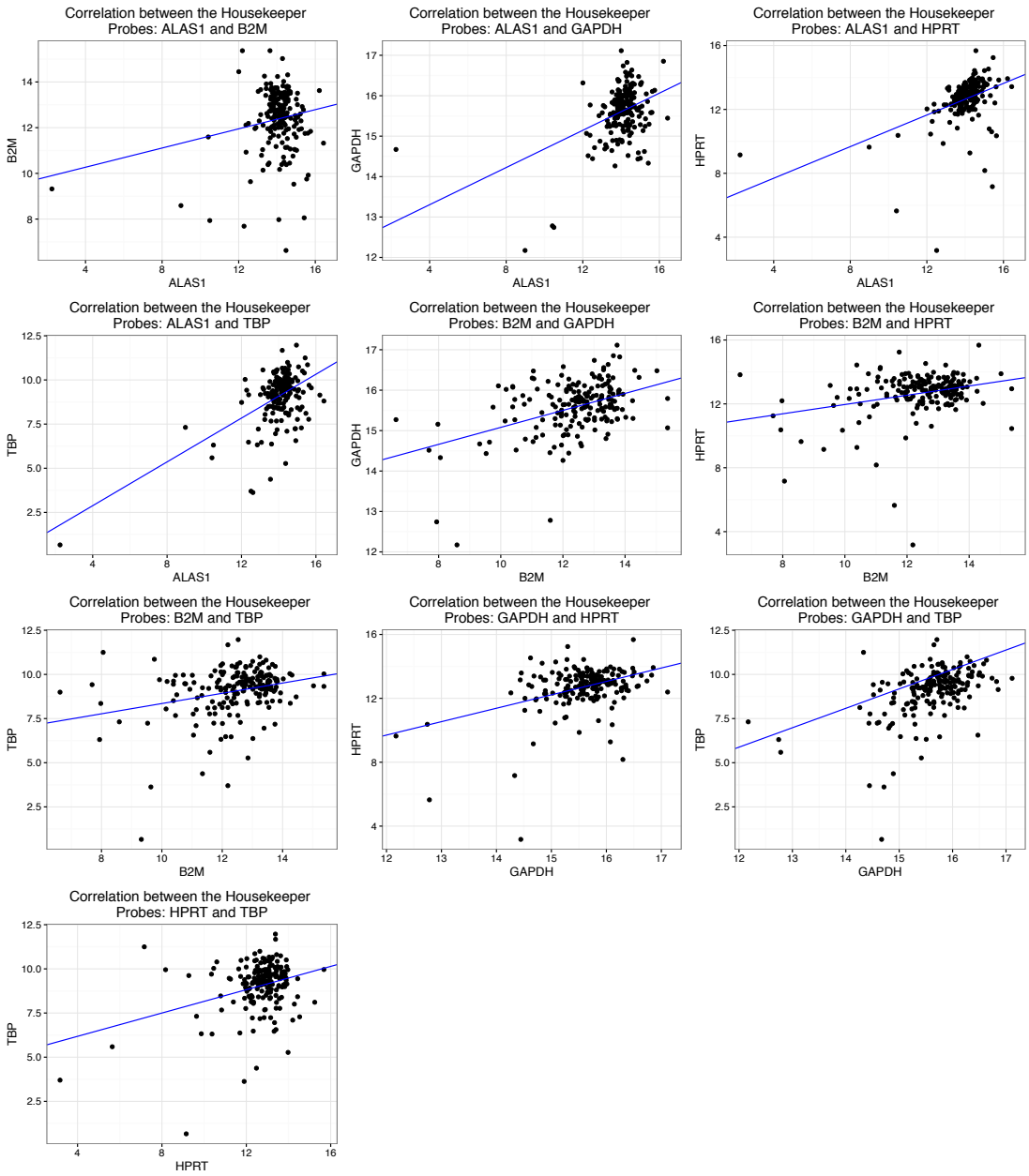


Figure 3.5 Correlation plots between the housekeeper transcripts: ALAS1, B2M, GAPDH, HPRT, and TBP.

Pearson's correlation coefficients (*R*) between housekeeping probes was below 0.5 in 9/10 comparisons (0.53 being the highest correlation), which suggests they are not well correlated (Table 3.9, Figure 3.5). This makes the choice of which housekeeping probes to normalise the data with difficult. So, for this reason it was decided to go ahead without using housekeeper style normalisation for these data.

Table 3.9 Housekeeper probe Pearson's correlation results, looking for correlating housekeeping probes.

R	*ALAS1*	*B2M*	*GAPDH*	*HPRT*	*TBP*
ALAS1	**-**	***0.19 ($p = 0.008$)***	***0.43 ($p = 7.8x10^{-10}$)***	***0.44 ($p = 1.8x10^{-10}$)***	***0.53 ($p = 1.6x10^{-15}$)***
B2M		**-**	***0.44 ($p = 2.5x10^{-10}$)***	***0.29 ($p = 5.2x10^{-05}$)***	***0.28 ($p = 7.7x10^{-05}$)***
GAPDH			**-**	***0.4 ($p = 1.1x10^{-08}$)***	***0.49 ($p = 7.8x10^{-13}$)***
HPRT				**-**	***0.32 ($p = 4.8x10^{-06}$)***
TBP					**-**

An alternative to using housekeeping transcripts could be to use a similar method to the PCA3 test, which uses *KLK3* (PSA) to enhance the expression of other probes in the data. *KLK3* adjusted data was produced but the resulted data showed much weaker, plateaued, signal strength and therefore, was not used for any subsequent analysis (data not shown).

3.3.6 Removal of Outliers

M_19_5 was identified via PCA (Figure 3.6) and NanoStringNorm (Table 3.4) as being an outlier that may hinder further analyses. Further investigation into this sample highlighted that 44 out of 57 probes for sample M_19_5 failed; in the (positive control normalised, $\log_2$ transformed) data all 44 probes had a value of "-0.07400058", indicating that they were undetectable. The other samples of this cartridge and scanner appear to have worked. Therefore, this sample alone will be removed for all subsequent analyses.

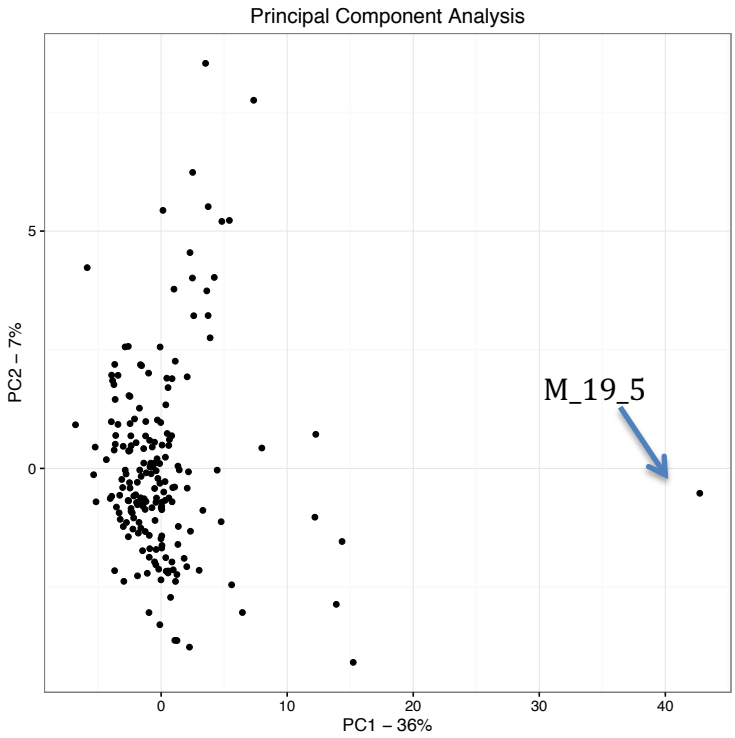


Figure 3.6 PCA plot of all log2 normalised data identifies an outlier samples M_19_5.

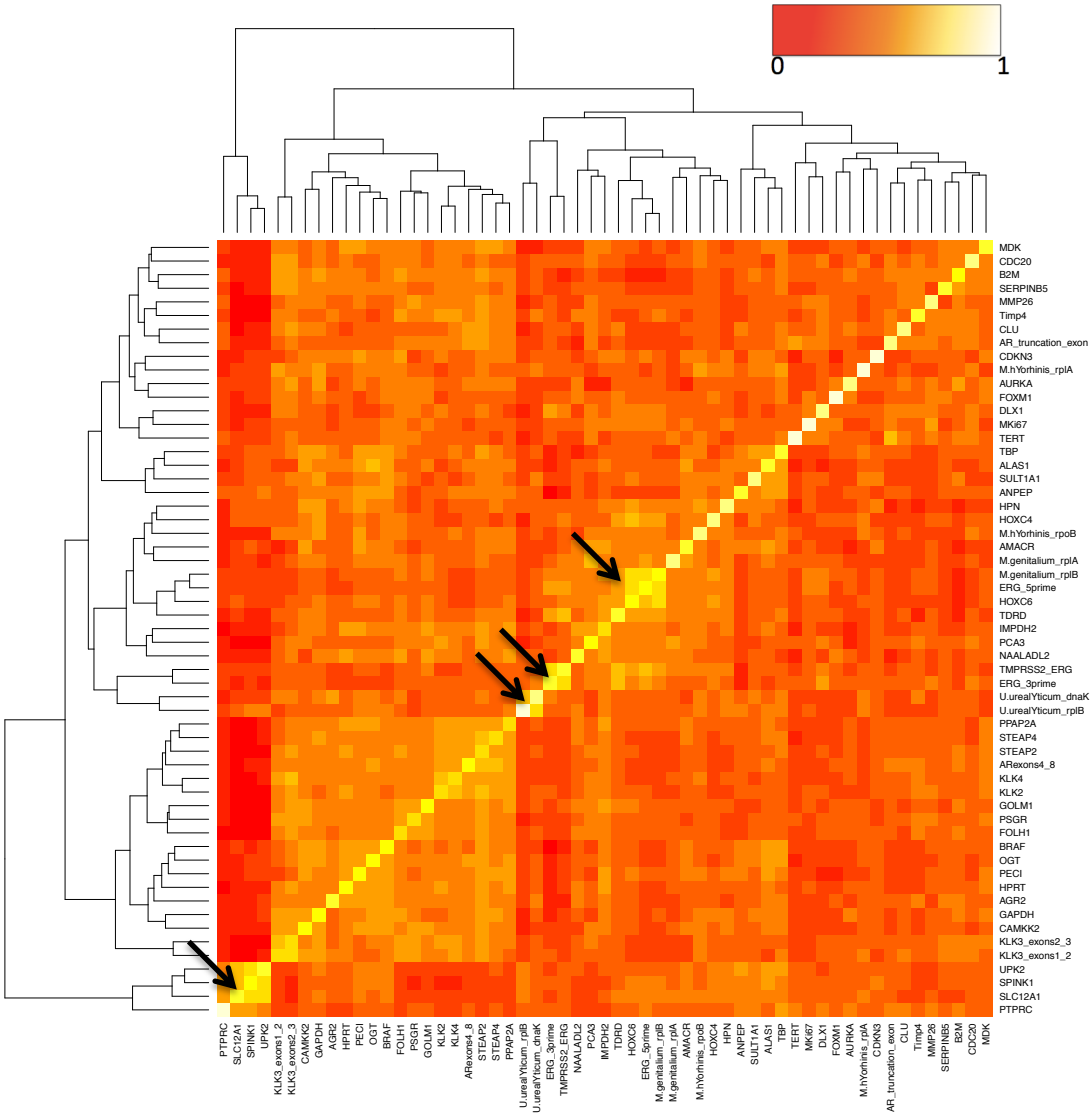
3.3.7 Correlating Gene Probes

Pearson's correlation (section 2.4.3) between data from all of the probes identifies four clusters of probes that showed strong inter-correlation (Figure 3.7, Table 3.10)**.** Cluster 1: probes for *ERG* 3' and *TMPRSS2:ERG*; Cluster 2: the two probes for the bacteria *U.urealYticum*; Cluster 3: two *M.genitalium* probes, *HOXC6* and *ERG* 5' and Cluster 4: *SLC12A1*, *SPINK1* and *UPK2*.

The data for probes in Clusters 1 and 2 were biologically expected to correlate, as were Cluster 3's two bacterial probes (*M.genitalium* RplA and RplB: Pearson's correlation: $p = 1.23\text{x}10^{-05}$, $R = 0.31$). However, Cluster 3's other correlations were not expected and were even more pronounced *i) M.genialium RplB* and *HOXC6* (Pearson's correlation: $p < 2.26\text{x}10^{-16}$, $R = 0.88$) ii) *M.genitalium RplB* and *ERG* 5' (Pearson's correlation: $p < 2.26\text{x}10^{-16}$, $R = 0.83$) and iii) between *HOXC6* and *ERG* 5' (Pearson's correlation: $p < 2.26\text{x}10^{-16}$, $R = 0.73$) (Figure 3.8). Also in Cluster 3, the two *M.genitalium* probes would be expected to have similar signal strength, which is not the case (Figure 3.8).

M.genitalium RplA had signal range ~-1 to 4, whilst *RplB* had a signal range of 0-12 with most samples above 6. *M.genitalium* RplB signal strength was actually more similar to *HOXC6* (~5-16) and *ERG3*' (~2-12).

Figure 3.7 Heatmap showing correlation between all NanoString probe data. The colours reflect the



***R* value of the correlation, where 1 is perfect correlation (represented by yellow) and 0 is uncorrelated (represented by red), with orange in between.**

Table 3.10 Four Clusters of probes that correlate with each other (Pearson's correlation).

	Correlating probes r values		
	ERG 3'		
TMPRSS2:ERG	$p < 2.2x10^{=16}$, **R = 0.74**		
	U.urealYticum dnaK		
U.urealYticum RplB	$p < 2.2x10^{=16}$, **R = 0.56**		
	M.genitalium RplB	***HOXC6***	***ERG 5'***
M.genitalium RplA	$p = 1.23x10^{-05}$, **R = 0.31**	$p = 4.08x10^{-07}$, **R = 0.36**	$p = 3.65x10^{-08}$, **R = 0.38**
M.genitalium RplB		$p < 2.26x10^{-16}$, **R = 0.88**	$p < 2.26x10^{-16}$, **R = 0.83**
HOXC6			$p < 2.26x10^{-16}$, **R = 0.73**
	SLC12A1	***UPK2***	
SPINK1	$p < 2.26x10^{-16}$, **R = 0.64**	$p < 2.26x10^{-16}$, **R = 0.78**	
SLC12A1		$p < 2.26x10^{-16}$, **R = 0.62**	

Needleman-Wunsch alignment of the capture and reporter probes for *HOXC6*, *ERG* 5', *M.genitalium* RplA and *M.genitalium* RplB gave low percentage alignments and scores with each other. These scores were similar to alignments with three randomly selected NanoString probes (which were selected for a control comparison) that showed no expression correlation; *ALAS1*, *KLK2* and *KLK3*. BLAT analysis detected some homology between *M.genitalium* RplA reporter probe sequence and non-coding sequences on human ChrX, whilst *M.genitalium* RplB capture probe hits non-coding sequence on Chr10. Both *HOXC6* and *ERG* 5' capture and reporter probes only had sequence homologies with their own encoding gene sequences and nowhere else in the genome. These analyses suggest that cross-hybridisation is not likely to be the cause of their correlation.

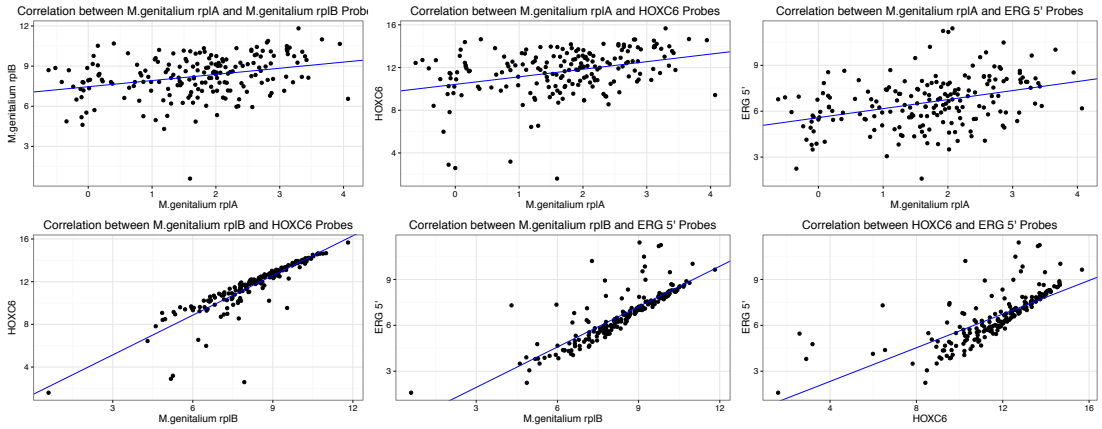


Figure 3.8 Correlation plots between data for probes: *M.genitalium* RplA, *M.genitalium* RplB, *HOXC6* and *ERG* 5'.

In Cluster 4, one transcript (Spink1) is known to be associated with PCa while the other two are tissue specific controls; *UPK2* is a bladder specific marker and *SLC12A1* is a kidney specific marker. It is understandable to see some correlation between the non-prostate tissue specific markers, as the proportion of these would result from the proportion of cells that are not from the prostate. The correlation between *UPK2* and *SLC12A1* data, whilst significant is not strong enough to suggest that they are cross hybridising ($p < 2.26\text{x}10^{-16}$, $R = 0.62$) (Figure 3.9). *UPK2* and *SPINK1* correlate strongly ($p < 2.26\text{x}10^{-16}$, $R = 0.78$), whereas *SLC12A1* correlation with *SPINK1* is weaker ($p < 2.26\text{x}10^{-16}$, $R = 0.64$) (Figure 3.9). All three probes have similar signal strength also, ranging ~0 to ~15 (Figure 3.9). Needleman-Wunsch alignment of the capture and reporter probes for *SPINK1*, *SLC12A1* and *UPK2* gave low percentage alignments and scores with each other. These scores were similar to those of three randomly selected NanoString control probes that showed no expression correlation; *ALAS1*, *KLK2* and *KLK3*. Furthermore, BLAT analysis detected no other sites of homology in the human genome for *SPINK1* probe sequences, whilst both *UPK2* and *SLC12A1* reporter probes had one partial match each: *CTNNA3* (Chr 10) and *FLRT2* (Chr 14), respectively. The capture probes for *UPK2* and *SLC12A1* also had no alternative sites of homology in the human genome. This suggests that the probes are not cross-hybridising to each others target probes.

It is possible that some of these probes in the two clusters are cross-hybridising (and then it is of course possible that at least one is a true representation for that probe) or that there is a clinical reasoning for their correlation. For this reason, I have included all of these probes in the subsequent analyses but any identification of their significance in clinical comparisons or clustering should be taken with caution.

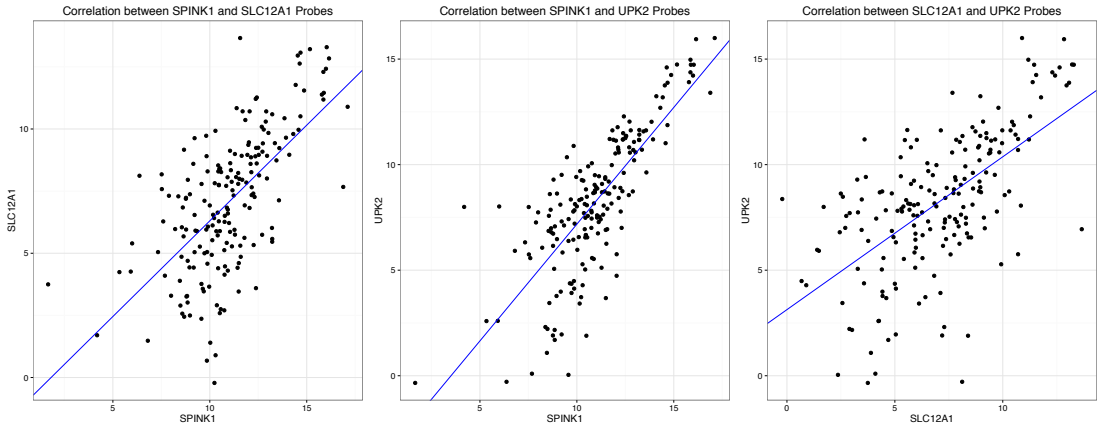


Figure 3.9 Correlation plots for a second group of probes that correlate: *SPINK1*, *SLC12A1* and *UPK2*. All correlate with $p < 2.26 \times 10^{-16}$ and $R < 0.6$.

3.4 Identification of Prostate and Cancer Specific Transcripts and DRE relevance

3.4.1 Kallikrein identification

NanoString median signals for the *KLK2*, *KLK3* exons 1-2, *KLK3* exons 2-3 and *KLK4* probes were at significantly higher levels than those for the control tissue probes for blood, kidney and bladder (*PTPRC*, *SLC12A1* and *UPK2* respectively) (Mann Whitney U test: $p < 2.2 \times 10^{-16}$ in each case, Table 3.11, Figure 3.10).

Table 3.11 Median expression values for kallikreins (prostate specific transcripts) and other tissue markers.

Probe	*Tissue*	*Log2 Median expression*
KLK2	***Prostate***	***13.49***
KLK3 exons 1-2	***Prostate***	***14.35***
KLK3 exons 2-3	***Prostate***	***14.02***
KLK4	***Prostate***	***15.59***
PTPRC	***Blood***	***4.08***
SLC12A1	***Kidney***	***7.24***
UPK2	***Bladder***	***8.15***

The kallikreins are prostate specific transcripts[205]; identification of *KLK2*, *KLK3* and *KLK4* at higher levels in the blood, kidney and bladder specific markers along with the RNA yield of post radical prostatectomy samples (section 3.4.4) suggest that a good proportion of the material captured is in fact from the prostate. Additionally, both the *KLK3* probes (exons 1-2 and exons 2-3) have a strong correlation (p<2.2x10^{-16}, R = 0.89, Figure 3.10B).

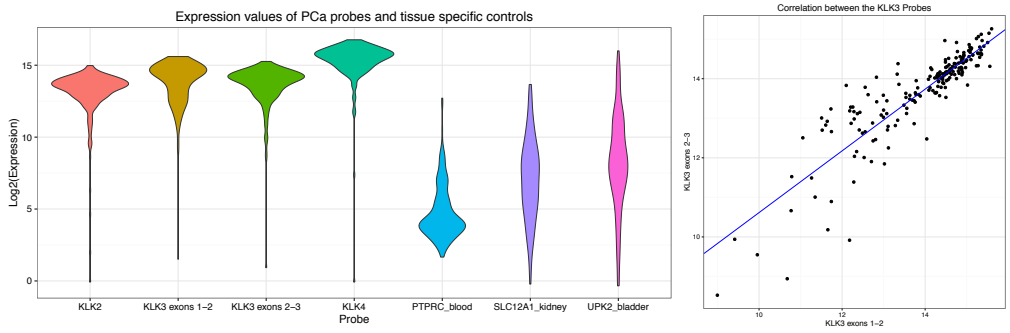


Figure 3.10 A) Kallikreins are observed at higher expression levels than the blood, kidney and bladder specific markers in the NanoString data. B) Correlation between the two KLK3 probes is strong.

3.4.2 *TMPRSS2:ERG* Identification

TMPRSS2:ERG fusions, and alleviated *ERG* 3' and *ERG* 5' expression are found in PCa, and this is observed in the Nanostring data where a significant difference is

observed between cancer vs CB (Mann Whitney U; *TMPRSS2:ERG*: $p < 2.2\text{x}10^{-16}$; $W = 6179$, *ERG* 3': $p < 2.2\text{x}10^{-16}$; $W = 6105$, and *ERG* 5': $p < 2.2\text{x}10^{-16}$; W = 6253; Error! R eference source not found.). The density plots for *TMPRSS2:ERG* and *ERG3'* have two peaks which would be compatible with an on/off pattern of a gene fusion (Error! R eference source not found.). Approximately 50% of the samples from men with cancer have detectable *TMPRSS2:ERG* fusions which is in agreement with the literature (section 1.4.6). The *ERG*5' probe, which is not part of the *TMPRSS2:ERG* fusion transcript, does not follow this pattern. The *ERG* 5' probe was also identified as having potential cross hybridisation (section 3.3.7).

When dicotomised (using the optimal threshold 4.93 identified by the Brent method), *TMPRSS2:ERG* expression had a significant association with clinical category (chi-square test, $\chi^2 = 37.82$, $p = 4.1\text{x}10^{-07}$).

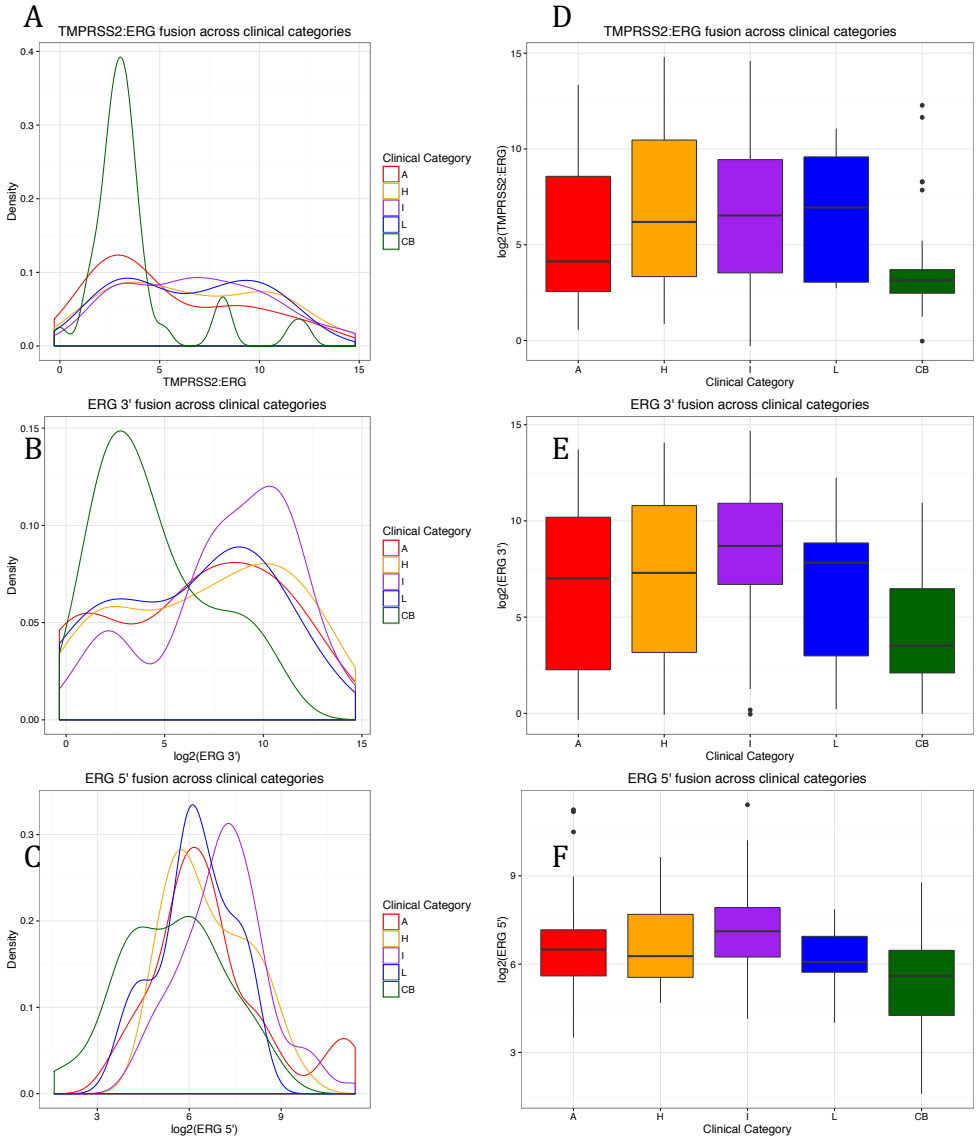


Figure 3.11 A) Density plot for *TMPRSS2:ERG* expression coloured by clinical category. Generally, two peaks are seen suggesting an on/off pattern of expression. B) Density plot for *ERG* 3' expression coloured by clinical category. Again, two bumps are generally seen suggesting an on/off pattern. C) Density plot for *ERG* 5' expression coloured by clinical category. No observable on/off pattern can be seen. D) Box plot showing spread of *TMPRSS2:ERG* expression across clinical categories. Higher expression is observed in cancer than benign. E and F] Box plots showing expression of *ERG* 3' and *ERG*5' respectively across clinical categories. Median expression is Higher in cancer than benign.

TMPRSS2:ERG fusions were identified by NanoString and nested RT-PCR (section 2.1.5 and section 2.1.6) (Figure 3.12). PCR was capable of identifying not only T1/E4 fusion transcripts but also fusions involving other *TMPRSS2* and *ERG* exons (section 1.4.6), including T1/E5, T1/E6 and T2/E2 amongst others. *TMPRSS2:ERG* PCR data were divided into four groups: i) T1/E4 fusions ("T1/E4"), ii) those with T1/E4 plus other fusion types ("T1/E4 plus"), iii) those with only non-T1/E4 products ("other") and iv) those where no fusions were identified ("negative").

The minimum curve threshold was calculated from NanoString expression density plots. A cut off of 6.78 for the *TMPRSS2:ERG* probe, showed 97% correlation for the PCR negatives (95/98 are classed as negative in both), 79% accuracy for T1/E4 only fusions (41/52 are classed as positive), 88% accuracy for T1/E4 plus other fusions (21/24 are classed as positive), and 42% accuracy for other fusions (8/19 are classed as positive). The NanoString *TMPRSS2:ERG* probe had been designed to specifically pick up the T1/E4 fusion, and so the poor accuracy for detecting other fusions was expected. Using the optimal threshold identified by the Brent method for the *TMPRSS2:ERG* probe (4.93), the on/off pattern compared with the PCR results, showed an improved and significant association (chi-square test $\chi^2 = 131.6$, $p < 2.2\text{x}10^{-16}$).

The *ERG*3' NanoString signal correlates well to the *TMPRSS2:ERG* PCR positive samples for both T1/E4 fusion and non-T1/E4 PCR products. However there are a proportion of the PCR negative samples that also have high *ERG*3' NanoString signals; this would appear to indicate that *ERG*3' has been overexpressed via an alternate mechanism (section 1.4.6).

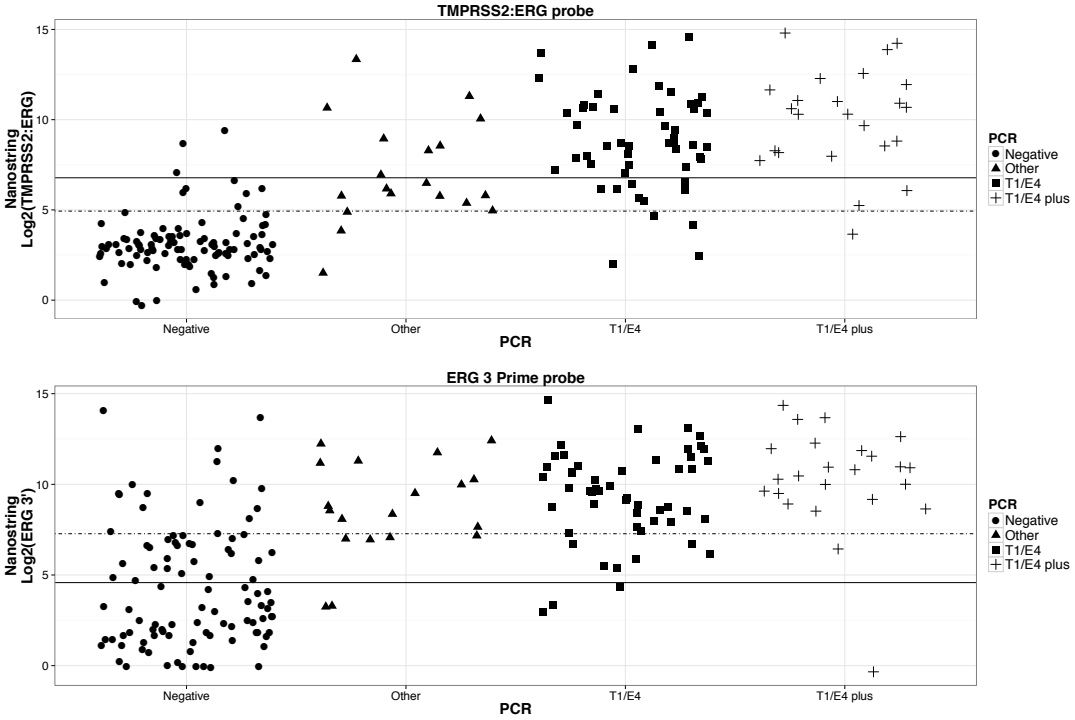


Figure 3.12 Detection of *TMPRSS2:ERG* by NanoString probes for *TMPRSS2:ERG* (upper) and *ERG*3' (lower) versus PCR detection of *TMPRSS2:ERG* transcripts. T1/E4 indicates a *TMPRSS2* ex1/*ERG*ex4 fusion transcript, 'Other' indicates a different fusion transcript, 'Plus' indicates a mixture of T1/E4 and other transcripts. The dotted lines are the optimal thresholds (4.93 for *TMPRSS2:ERG* and 7.28 for *ERG*3') calculated using the Brent method, similarly the solid line is the min curve of a density plot (6.78 and 4.58 for *TMPRSS2:ERG* and *ERG*3' respectively) containing all of the *TMPRSS2:ERG* and *ERG* data.

These results suggest that a) NanoString is a sensitive and flexible method for detecting transcripts and b) that a proportion of the genetic material identified is coming from prostate cancer or HG-PIN.

3.4.3 PCA3 Test

The *PCA3* test (section 1.4.2) is the ratio of *PCA3* expression with *KLK3* expression in whole urine, and is approved clinically to predict whether a second biopsy will be cancer positive after an initial negative biopsy. The *PCA3* score calculated from the NanoString data shows a significantly increased expression in PCa compared with non-PCa samples (Mann-Whitney U test: $p < 2.2\text{x}10^{-16}$, Figure 3.13), but was no evidence for a significantly difference between different clinical categories of PCa ($p < 0.05$; Kruskal-Wallis rank sum).

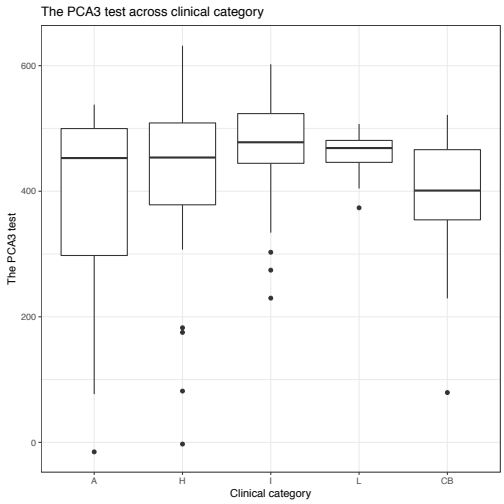


Figure 3.13 Nanostring PCA3 score calculation (*PCA3* divided by *KLK3* multiplied by 1000 as per the usual PCA3 score (section 1.4.2). The PCA3 score is significantly increased in PCa samples compared to those with no clinical evidence of PCa (CB). However, there is no significant difference between the intra-clinical categories of PCa. The uPM3^{TM} assay has shown to be able to detect PCa from non-PCa samples. The NanoString probes have shown to follow this same pattern.

3.4.4 RNA yield, clinical group and DRE

Digital rectal examination (DRE, section 1.3.3) has proven to increase the efficacy of the *PCA3* test (section 1.4.2). It is hypothesised that digital compression on the prostate encourages secreted biomarkers in the gland to flow towards the urethra. Four patient

pairs for pre- and post-DRE urine samples were added to NanoString to see how the transcript levels varied within patients (Figure 3.15). First-void urine post-DRE had higher median RNA yields than non-DRE samples (Figure 3.14). RNA yield is significantly higher in post-DRE collection of localised PCa samples compared to pre-DRE samples ($p = 0.04$, Mann Whitney U test) and prostatectomy samples ($p = 0.01$, Mann Whitney U test). As seen previously there were also increased numbers of prostate derived transcripts (section 3.4.1) and PCa derived transcripts (section 3.4.2) on post-DRE samples. Overall, the post-DRE samples had 0.178 $\log_2$ fold increased expression of all transcripts compared to the pre-DRE collected samples ($p = 1.854 \times 10^{-10}$, paired Man Whitney U test). The median of the sample pairs individually varied with the pre- or post-DRE, however, the post-DRE sample always showed a lower IQR (Figure 3.15).

The urine taken from three patients who had previously undergone radical prostatectomy (post-RP) had very low amounts of RNA collected (0.8-2ng) from their urine samples. This suggests that the majority of the EV RNA is likely to have originated in the prostate (Figure 3.14).

The median RNA yields for advanced PCa patients are not significantly lower than for localised-PCa patients ($p < 0.05$, Mann Whitney U test, Figure 3.14). The RNA yields for benign samples are observably (Figure 3.14) and significantly lower compared to localised PCa patient samples ($p = 0.02$, Mann Whitney U test).

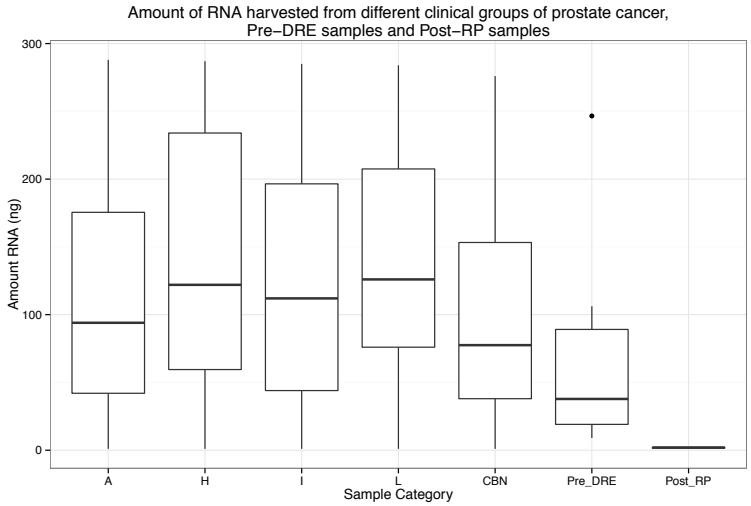


Figure 3.14 Most of the transcripts detected are from the prostate; DRE boosts transcript level detection and post radical prostatectomy patients offer very low signals in their samples. Samples *n* = 389. The advanced (A), high-risk (H), intermediate risk (I), low-risk (L) and no evidence of clinical PCa (CB) samples were taken post-DRE. Pre-DRE and post-RP urine samples have been taken without DRE.

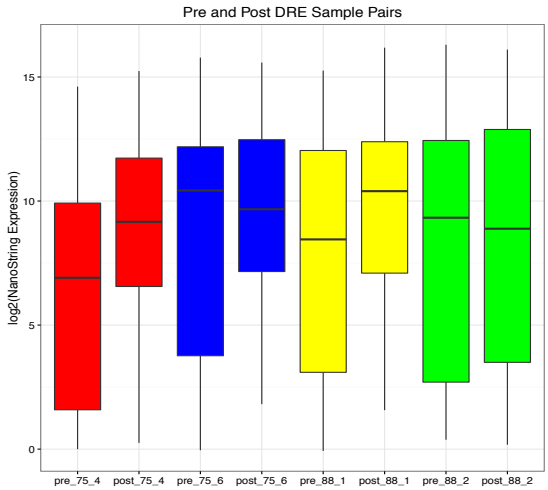


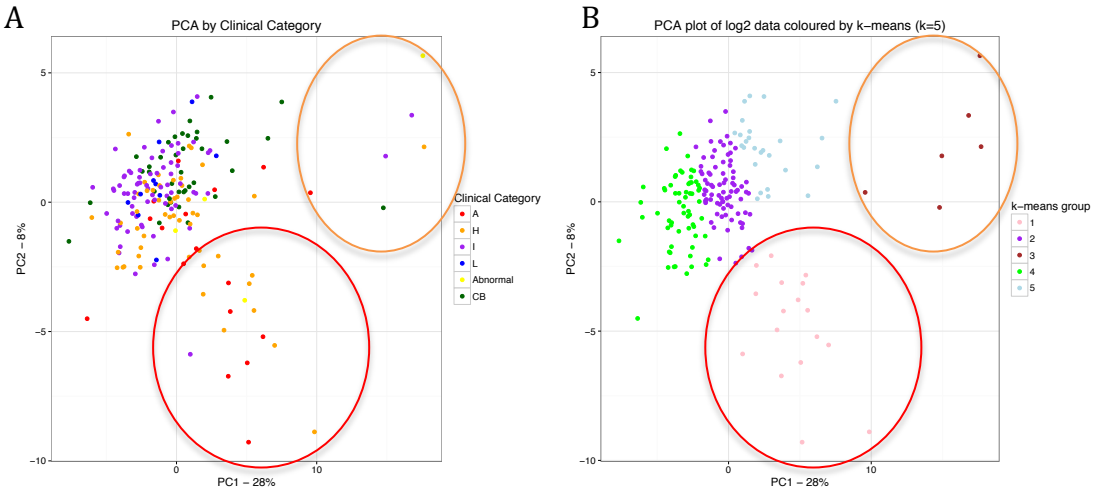
Figure 3.15 The NanoString probe expression distribution of four patient paired samples (pre- and post-DRE).

3.5 Clustering

3.5.1 Principal Component Analysis (PCA) and *k*-means Clustering

PCA (section 2.5.1) can be utilised to visualise groups in the data; colouring data by clinical category can allow clusters of biological interest to be identified (Figure 3.16A). PCA analysis identified two outlying clusters, A and B, (Figure 3.16). Cluster A had 17 samples consisting mostly of advanced and higher risk samples (6 advanced, 9 high-risk, 1 intermediate risk and 1 abnormal sample). In contrast Cluster B consisted of 6 samples of varying clinical groups (1 advanced, 1 high risk, 2 intermediate risk, 1 abnormal and 1 CB).

Figure 3.16 PCA plots coloured by A) clinical category and B) k-means to identify cluster cut-offs.



Cluster A shown by red circle. Cluster B shown by orange circle.

3.5.2 Hierarchical Clustering

Hierarchical clustering was performed with an agglomerative approach (section 2.5.2). This showed that samples in Clusters A and B belonged to separate trees to the majority of other samples (Figure 3.17A). Fifteen of the samples belonging to Cluster A form a separate tree, whilst 5 of the 6 samples belonging to Cluster B also form a separate tree with 2 other samples. There was one significant cluster identified by Pvclust (section

2.5.2.1), which contains the bulk of the samples, but does not include the majority of Cluster A or Cluster B samples (Figure 3.17B).

3.5.3 Cluster A

Cluster A (identified by PCA and *k*-means clustering (Section 3.5.1) and supported by hierarchical clustering (Section 3.5.2) is predominantly made up by advanced and high-risk samples (6/17 and 9/17, respectively). It has significant over-representation of advanced and high-risk samples (Table 3.12) and there are twenty-three significant differentially expressed transcripts between cluster A and all other samples (Table 3.13). Analysis of the differential expressed transcript list with DAVID (section 2.7.1) identified PCa as an over represented KEGG pathway. This was due to the significantly lowered expression of *AR* and *KLK3* in Cluster A, however the over-representation was not significant at a 95% confidence level ($p = 8.5x10^{-02}$). Ten Gene Ontology (GO) biological processes were associated with the Cluster A defining transcripts (Table 3.14). As expected due to probe selection for involvement in PCa these biological processes were associated with cancer. However, different GO biological processes were identified using all of the transcripts applied to NanoString (Table 3.15). Thus suggesting there is a difference in biological processes involved specifically within cluster A.

RNA amount (ng) extracted is significantly lower in Cluster A compared to all other samples (not including Cluster B), (Table 3.12). Cluster A also has a significantly lower amplification yield, as well as a lower median probe value (Table 3.12). The cartridge number is also significant between members of Cluster A (Table 3.12). However, Scanner ID is not significant (Table 3.12).

Further investigation into the cartridges involved, showed there was no significant differences between the median probe values of these cartridges compared to others, or between the Cluster A samples and non-Cluster A samples on these cartridges (Table 3.12). This suggests that the cartridge is not a factor to why Cluster A may be presenting itself. However, RNA extraction amount, amplification yield and median probe value all seem important in the clustering. Especially as 21/23 significant

differentially expressed transcripts between Cluster A and all other samples (Table 3.13) have lower expression in Cluster A.

Lower RNA yields are observed in a fraction of advanced patients' samples, This is hypothesised to be due to a reduction in tumour microvesicles being harvested in the urine due to: a) efficiency of DRE: the surface of a normal prostate can be depressed by 1cm, however prostates containing advanced/higher grade tumours are commonly firm and not depressible. Samples from patients with advanced tumours are therefore more akin to non-DRE samples. b) Advanced tumours can also have fused glands, poorly formed lumen, and blind-ended lumen that no longer drain into the urethra[206, 207]. The position of the advanced tumour within the prostate may also block access of tumour biomarkers from less advanced PCa foci from entering the urine. Thus, the percentage of the tumour that is advanced and its positioning within the prostate can affect the amount of RNA extracted, and the amounts of PCa associated transcripts identified.

Table 3.12 Testing for Cluster A association to clinical and technical variables.

Variable	*Test and metric*	*p - value*
Clinical Category	***Chi square: χ = 20.29***	***p = 6.67x10^{-06}***
Amount RNA extracted (ng)	***Mann-Whitney U test: R = 2443***	***p = 4.9x10^{-07}***
Amplification yield (µg)	***Mann-Whitney U test: R = 2410***	***p = 3.1x10^{-06}***
Median probe value	***Mann-Whitney U test: R = 1904***	***p = 0.02***
Cartridge	***Chi-square test: χ^2 = 31.9***	***p = 0.01***
Scanner ID	***Chi-square test: χ^2 = 0.03***	***p = 0.9***
Median probe value of Cluster A samples on cartridge 13 (n = 4) compared to other samples on cartridge 13 (n = 8)	***Mann-Whitney U test: R = 16***	***p = 1***
Median probe value of samples on cartridge 13 (n = 12) compared to samples on all other cartridges (n = 168)	***Mann-Whitney U test: R = 1237***	***p = 0.3***
Median probe value of Cluster A samples on cartridge 15 (n = 3) compared to other samples on cartridge 15 (n = 9)	***Mann-Whitney U test: R = 4***	***p = 0.1***
Median probe value of samples on cartridge 15 (n = 12) compared to samples on all other cartridges (n = 168)	***Mann-Whitney U test: R = 1074***	***p = 0.8***

It should also be remembered that the vast majority of the NanoString probes were selected due to overexpression in tumour tissue. Thus, it is significant that the expression patterns for Cluster A are more than a general loss of tumour biomarkers as my analyses mark them as a group distinct from the other prostate samples. The only two probes not showing a significant up-regulation in the Cluster A samples are the kidney and bladder controls (Table 3.13).

It is hypothesised that the factors identified as technical issues (RNA amount extracted, amplification yield and median probe value) associated with Cluster A are due to these biological reasons and thus it is important to keep Cluster A's samples within future analyses.

Table 3.13 Transcripts significantly associated ($p < 0.05$) with Cluster A via Mann-Whitney U test after using Hochberg multiple testing correction.

Transcript	*p-value*	*Adjusted p-value*	*Log2 Fold Change*
DLX1	$6.6x10^{-04}$	$2.3x10^{-02}$	**-1.504**
Timp4	$1.2x10^{-04}$	$4.7x10^{-03}$	**-1.292**
AR exon 9	$2.3x10^{-06}$	$3.2x10^{-04}$	**-1.263**
MMP26	$7.4x10^{-06}$	$3.2x10^{-04}$	**-1,126**
CLU	$6.4x10^{-05}$	$2.6x10^{-03}$	**-1.017**
UPK2	$5.1x10^{-10}$	$2.8x10^{-08}$	**0.798**
SLC12A1	$2.5x10^{-09}$	$1.4x10^{-07}$	**0.736**
PSGR	$2.6x10^{-08}$	$1.3x10^{-06}$	**-0.555**
CDC20	$3x10^{-04}$	$1.1x10^{-02}$	**-0.543**
SPINK1	$1.5x10^{-10}$	$8.3x10^{-09}$	**-0.497**
GOLM1	$1.4x10^{-05}$	$6x10^{-04}$	**-0.485**
PCA3	$1.8x10^{-04}$	$6.9x10^{-03}$	**-0.456**
SERPINB5	$3x10^{-04}$	$1.1x10^{-02}$	**-0.287**
KLK3 exons 2-3	$4.2x10^{-10}$	$2.4x10^{-08}$	**-0.251**
KLK3 exons 1-2	$4.9x10^{-08}$	$2.4x10^{-06}$	**-0.235**
FOLH1	$6x10^{-08}$	$2.9x10^{-06}$	**-0.214**
B2M	$1.1x10^{-04}$	$4.2x0^{-03}$	**-0.207**
AR exons 4-8	$7.9x10^{-07}$	$3.6x10^{-05}$	**-0.186**
STEAP2	$1.2x10^{-08}$	$6.2x10^{-07}$	**-0.183**
KLK2	$1.2x10^{-08}$	$6.2x10^{-07}$	**-0.174**
KLK4	$6.1x10^{-09}$	$3.2x10^{-07}$	**-0.132**
STEAP4	$8.5x10^{-06}$	$3.6x10^{-04}$	**-0.129**
PPAP2A	$3.5x10^{-07}$	$1.7x10^{-05}$	**-0.128**

Table 3.14 Gene Ontology (GO) over-represented biological processes in Cluster A's significantly associated transcript list via DAVID.

Term	*Count (%)*	*Transcripts*	*p - value*	*Adjusted p - value*
Proteolysis	***7 (3.9)***	***FOLH1, KLK2, KLK3, KLK4, CLU, MMP26, CDC20***	$8.9x10^{-04}$	$2.5x10^{-01}$
Iron ion transport	***2 (1.1)***	***STEAP4, STEAP2***	$3.4x10^{-02}$	***1***
Androgen receptor signalling pathway	***2 (1.1)***	***AR, PPAP2A***	$4.2x10^{-02}$	$9.9x10^{-01}$
Response to organic substance	***4 (2.2)***	***AR, TIMP4, STEAP2, B2M***	$5.0x10^{-02}$	$9.8x10^{-01}$
Steroid hormone receptor signalling pathway	***2 (1.1)***	***AR, PPAP2A***	$6.6x10^{-02}$	$9.9x10^{-01}$
Response to hormone stimulus	***3 (1.7)***	***AR, TIMP4, STEAP2***	$6.9x10^{-02}$	$9.8x10^{-01}$
Transition metal ion transport	***2 (1.1)***	***STEAP4, STEAP2***	$8.1x10^{-02}$	$9.8x10^{-01}$
Response to endogenous stimulus	***3 (1.7)***	***AR, TIMP4, STEAP2***	$8.1x10^{-02}$	$9.7x10^{-01}$
Intracellular receptor-mediated signalling pathway	***2 (1.1)***	***AR, PPAP2A***	$8.5x10^{-02}$	$9.6x10^{-01}$
Response to molecule of bacterial origin	***2 (1.1)***	***TIMP4, B2M***	$9.7x10^{-02}$	$9.6x10^{-01}$

Table 3.15 Gene Ontology (GO) over-represented biological processes in all of the transcripts used on NanoString via DAVID.

Term	*Count (%)*	*Transcripts*	*p - value*	*Adjusted p - value*

Proteolysis	***7 (0.1)***	***ANPEP, HPN, KLK2, KLK3, KLK4, MMP26, TMPRSS2***	***$9.6x10^{-04}$***	***$3.5x10^{-01}$***
Iron ion transport	***3 (0)***	***STEAP2, STEAP4, B2M***	***$2.2x10^{-03}$***	***$3.9x10^{-01}$***
Negative regulation of neuron apoptotic process	***4 (0.1)***	***BRAF, DLX1, MDK, TERT***	***$3.5x10^{-03}$***	***$4.1x10^{-01}$***
Ferric iron import into cell	***2 (0)***	***STEAP2, STEAP4***	***$1.2x10^{-02}$***	***$7.3x10^{-01}$***
Cell cycle	***4 (0.1)***	***AURKA, CDC20, CDKN3, FOXM1***	***$1.4x10^{-02}$***	***$7.1x10^{-01}$***
Copper ion import	***2 (0)***	***STEAP2, STEAP4***	***$1.6x10^{-02}$***	***$7x10^{-01}$***
Positive regulation of gene expression	***4 (0.1)***	***BRAF, AR, AGR2, HPN***	***$2.3x10^{-02}$***	***$7.7x10^{-01}$***
Positive regulation of stem cell proliferation	***2 (0)***	***PTPRC, TERT***	***$2.8x10^{-02}$***	***$7.9x10^{-01}$***
Positive regulation of transcription, DNA-templated	***5 (0.1)***	***TBP, AR, CAMKK2, FOXM1, MDK***	***$3.1x10^{-02}$***	***$7.9x10^{-01}$***
Response to cadmium ion	***2 (0)***	***B2M, TERT***	***$5.7x10^{-02}$***	***$9.2x10^{-01}$***
Negative regulation of endothelial cell apoptotic process	***2 (0)***	***BRAF, TERT***	***$6.3x10^{-02}$***	***$9.3x10^{-01}$***
Embryonic skeletal system development	***2 (0)***	***DLX1, HOXC6***	***$6.7x10^{-02}$***	***$9.2x10^{-01}$***
Protein phosphorylation	***4 (0.1)***	***BRAF, ERG, AURKA, CAMKK2***	***$8.9x10^{-02}$***	***$9.6x10^{-01}$***
Cell differentiation	***4 (0.1)***	***ERG, ANPEP, AGR2, MDK***	***$9.1x10^{-02}$***	***$9.5x10^{-01}$***
Response to peptide hormone	***2 (0)***	***BRAF, Timp4***	***$9.7x10^{-02}$***	***$9.5x10^{-01}$***

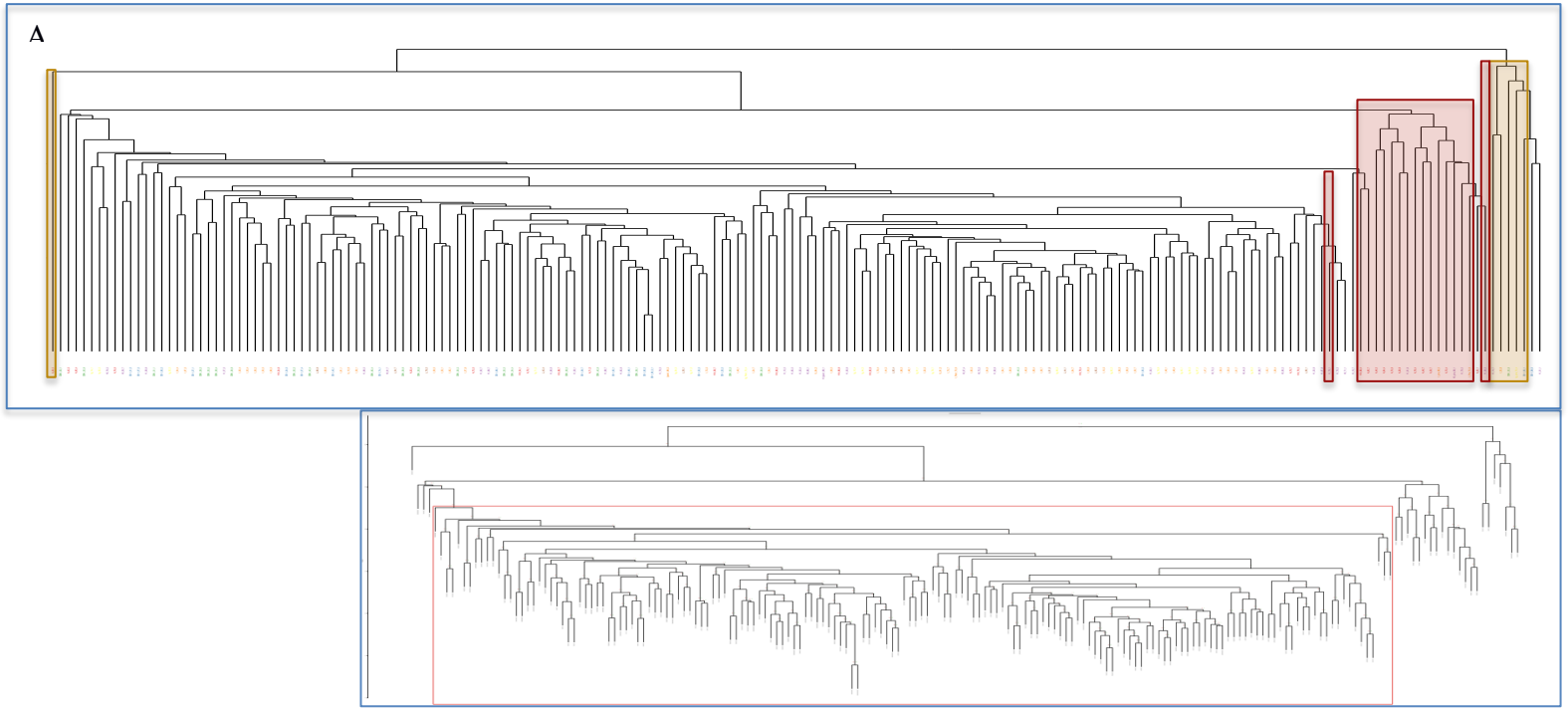


Figure 3.17 Hierarchical clustering provides further evidence for Cluster A and B identification. A) Samples belonging to Cluster A and B are shown in red and yellow boxes, respectively. B) Clusters with significant AU *p*-values are encapsulated within a red box. Both Cluster A and B are not included within this main.

3.5.4 Cluster B

Cluster B (identified by PCA and *k*-means clustering section 3.5.1 and supported by hierarchical clustering section 3.5.2) contains six samples and are not associated with clinical factors (Chi square: χ^2 = 9.7, p = 0.08), suggesting that Cluster B could be associated with technical artefacts: The amount of RNA extracted was lower in Cluster B (Mann-Whitney U test: R = 811, p = 0.008); the total amount of cDNA from amplification was also lower (Mann-Whitney U test: R = 808, p = 0.01) and thus was the median probe value (Mann-Whitney U test: R = 14196, $p < 2.2\text{x}10^{-16}$). Cartridge and Scanner ID were both not significantly associated with Cluster B (Chi square: χ^2 = 0.67, p = 0.88, and χ^2 = 2.67, p = 0.1, respectively). It is therefore, unlikely that there is biological reasoning to this cluster.

3.5.5 Latent Process Decomposition (LPD)

LPD (section 2.5.5) was performed on 187 of the samples (with M_19_5, LNCAP, and the five samples in cluster B removed) and 51 of the transcripts (with *FOXM1 and* the six bacterial genes removed) to identify the optimal number of groups and an assign a probability of membership for each group for each sample.

The modelling and estimation stage suggested that there were four clusters, with a sigma parameter of -1. LPD analysis was performed 100 times with these parameters and samples were associated with a probability to each group (Table 3.16, Figure 3.18A, Figure 3.18B). LPD 1 consisted mainly of high-risk samples (χ = 16.5, p = 0.01), whilst LPD 4 consisted mostly of advanced and high-risk samples (χ = 29.44, $p = 5\text{x}10^{-05}$). Both LPD 2 and 3 contain a mixed representation of clinical category. LPD group 2 consists mostly of the intermediate risk samples (χ = 29.44, $p = 5\text{x}10^{-05}$), it holds 66% of the intermediate and low-risk samples. LPD group 3 holds 57% of the

benign samples though they are not significantly represented within the group (χ = 11.82, p = 0.07).

Clinical category is significantly associated with LPD process, (χ = 65.47, p = $2.83x10^{-08}$, Table 3.16). PSA level is significantly associated with LPD process, with higher values in LPD groups 1 and 4 (ANOVA, p = $7.53x10^{-06}$, Figure 3.18C), as well as Gleason score (χ = 85.38 and p = $9.98\ x10^{-10}$); a higher Gleason score appears to be associated with LPD process 4, whilst processes 2 and 3 have much lower (Figure 3.18D). Age is also significantly associated with LPD process (ANOVA, p = 0.002), with a higher age present in LPD process 4 (Figure 3.18E).

Alternative analysis was performed using NbClust (section 2.5.3), which identified three clusters as the optimal number of clusters in the data, and *k*-means with PCA (section 2.5.3) was used to identify which samples belonged to which cluster (Figure 3.18E). These clusters showed high overlap with the four clusters identified by LPD (Figure 3.18F), providing further evidence that this clustering is reliable.

Table 3.16 Composition of sample type in each LPD cluster (Cluster B samples and bacterial probes removed). Chi-square test: $p = 2.8x10^{-08}$, $X = 65.47$.

	Total Number of Samples	*Number of aggressive samples (A and H risk)*	*Number of lower-risk cancer samples (L and I risk)*	*Number of Abnormal samples (S)*	*Number of CB samples*
LPD1	***8***	***6***	***1***	***0***	***1***
LPD2	***79***	***20***	***53***	***0***	***6***
LPD3	***55***	***14***	***18***	***2***	***21***
LPD4	***17***	***13***	***1***	***1***	***2***
LPD NA	***26***	***12***	*7*	***0***	*7*

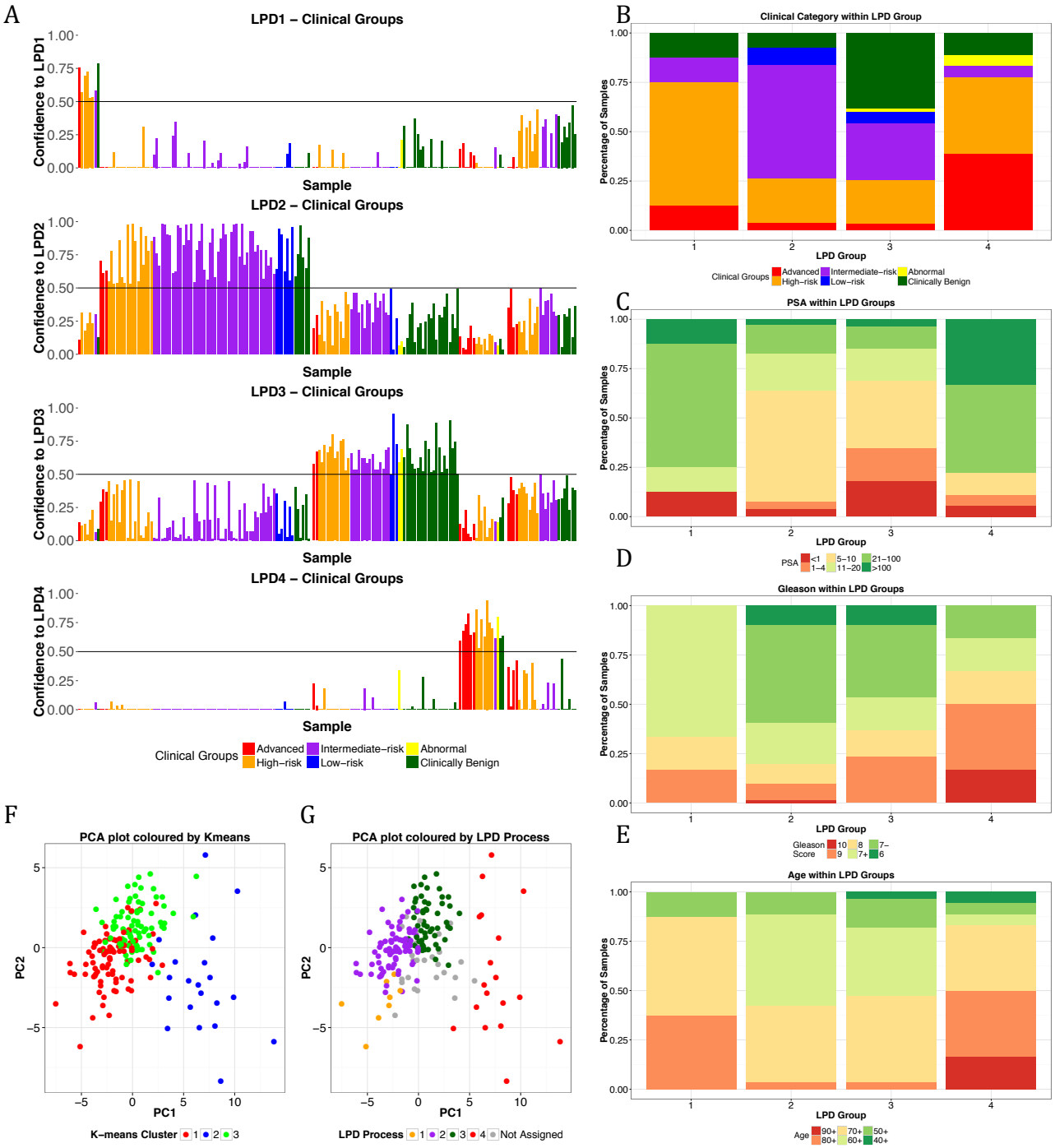


Figure 3.18 A) LPD group bar charts B,C,D,E) Clinical distribution, PSA, Gleason score and age without LPD group, respectively. F,G) PCA plots for *k*-means and LPD clustering comparison.

Table 3.17 Transcripts significantly different between each LPD group members and those that are not. These are the transcripts that define each LPD cluster.

LPD Process 1			*LPD Process 2*			*LPD Process 3*			*LPD Process 4*		
Gene	***p-value***	***Fold Change***	***Gene***	***p-value***	***Fold Change***	***Gene***	***p-value***	***Fold Change***	***Gene***	***p-value***	***Fold Change***
HOXC6	***0.0003***	***0.29***	***KLK3_PSA_exons2_3_***	***9.71E-14***	***0.08***	***TMPRSS2_ERG***	***4.40E-07***	***-1.19***	***KLK2***	***4.60E-09***	***-0.31***
AMACR	***0.001***	***0.27***	***KLK3_PSA_exons1_2_***	***5.49E-11***	***0.11***	***TDRD***	***9.02E-07***	***-0.83***	***KLK3_PSA_exons2_3_***	***4.65E-09***	***-0.25***
ERG_5prime	***0.002***	***0.44***	***STEAP2***	***7.09E-11***	***0.06***	***ERG_5prime***	***9.30E-07***	***-0.28***	***KLK4***	***8.32E-09***	***-0.14***
NAALADL2	***0.005***	***0.12***	***CAMKK2***	***6.02E-09***	***0.15***	***HOXC6***	***2.37E-05***	***-0.15***	***STEAP2***	***1.13E-08***	***-0.19***
TDRD	***0.012***	***0.86***	***MMP26***	***4.07E-08***	***0.69***	***AMACR***	***2.59E-05***	***-0.14***	***PPAP2A***	***4.37E-08***	***-0.15***
PECI	***0.012***	***0.12***	***KLK4***	***2.28E-07***	***0.05***	***ERG_3prime***	***8.38E-05***	***-1.28***	***FOLH1_PSMA***	***9.65E-07***	***-0.21***
FOLH1_PSMA	***0.016***	***0.14***	***GAPDH***	***2.94E-07***	***0.03***	***HOXC4***	***7.48E-04***	***-0.70***	***ARexons4_8***	***1.55E-06***	***-0.19***
IMPDH2	***0.018***	***0.10***	***FOLH1_PSMA***	***3.17E-07***	***0.09***	***CAMKK2***	***2.28E-03***	***-0.09***	***STEAP4***	***2.82E-06***	***-0.20***
DLX1	***0.020***	***0.91***	***OR52A2_PSGR***	***2.03E-06***	***0.13***	***DLX1***	***5.33E-03***	***-1.21***	***OR52A2_PSGR***	***2.97E-06***	***-0.48***
			ARexons4_8	***3.70E-06***	***0.07***	***GAPDH***	***5.33E-03***	***-0.03***	***KLK3_PSA_exons1_2_***	***4.03E-06***	***-0.24***
			KLK2	***6.96E-06***	***0.07***	***CDKN3***	***0.007***	***-0.42***	***MMP26***	***3.23E-05***	***-1.22***
			CDC20	***1.05E-05***	***0.50***	***PECI***	***0.009***	***-0.04***	***PCA3***	***4.85E-05***	***-0.57***
			TBP	***3.98E-05***	***0.08***	***MMP26***	***0.010***	***-0.46***	***AR_truncation_exon***	***5.61E-05***	***-1.38***
			CLU	***5.89E-***	***0.42***	***TERT***	***0.010***	***-0.17***	***UPK2***	***1.62E-***	***0.75***

		05					*04*	
	SERPINB5_Maspin	*7.52E-05*	*0.18*	*HPN*	*0.042*	*-0.08*	*SLC12A1*	*2.10E-04*	*0.67*
	GOLM1	*8.48E-05*	*0.16*	*BRAF*	*0.048*	*-0.03*	*SPINK1*	*3.20E-04*	*0.44*
	DLX1	*1.89E-04*	*1.24*				*HPRT*	*3.25E-04*	*-0.16*
	BRAF	*1.96E-04*	*0.06*				*GOLM1*	*4.14E-04*	*-0.65*
	TERT	*2.09E-04*	*0.25*				*Timp4*	*5.49E-04*	*-1.56*
	TDRD	*2.57E-04*	*0.61*				*CLU*	*6.21E-04*	*-1.07*
	TMPRSS2_ERG	*4.27E-04*	*1.11*				*CDC20*	*1.20E-03*	*-0.85*
	ERG_3prime	*6.23E-04*	*0.83*				*DLX1*	*6.39E-03*	*-1.58*
	PCA3	*6.23E-04*	*0.16*				*B2M*	*0.009*	*-0.09*
	AR_truncation_exon	*1.71E-03*	*0.71*				*IMPDH2*	*0.010*	*-0.10*
	MDK	*0.003*	*0.05*				*AGR2*	*0.021*	*-0.48*
	MKi67	*0.003*	*0.72*				*MDK*	*0.048*	*-0.21*
	B2M	*0.003*	*0.08*				*NAALADL2*	*0.048*	*-0.10*
	PPAP2A	*0.006*	*0.04*						
	Timp4	*0.006*	*0.21*						
	UPK2	*0.011*	*-0.23*						
	OGT	*0.011*	*0.04*						
	CDKN3	*0.012*	*0.35*						
	SPINK1	*0.012*	*-0.11*						
	ERG_5prime	*0.016*	*0.17*						
	AURKA	*0.036*	*0.15*						

3.6 Significantly varying genes

Expression distribution of each transcript was fairly even between clinical categories for most probes (Figure 3.19), with only 16 of the 57 probes found to be significantly different between the clinical categories (Kruskal-Wallis rank sum test, adjusted $p < 0.05$, Table 3.18).

Mann Whitney U tests (section 2.4.1) were applied to three separate data comparisons; i) cancer vs. non-cancer, ii) aggressive cancer (Advanced and high risk) vs. Non-aggressive cancer (I, L) and iii) the two extremes (Advanced vs. CB), (Table 3.3, page115). Nine probes were significantly differentially expressed ($p < 0.05$, Mann-Whitney U test) between cancer and non-cancer samples after multiple testing correction via the Hochberg method (Table 3.19). All of these transcripts were up-regulated in the cancer (Figure 3.20) and included many well established PCa-associated transcripts such as *ERG*, *TMPRSS2:ERG* and *PCA3*.

Table 3.18 Kruskal-Wallis identified 16 probes that significantly differ across clinical category.

Probe	*p-value*	*Adjusted p-value*	χ
SPINK1	3.9×10^{-09}	2.2×10^{-07}	47.79
SLC12A1	2.5×10^{-08}	1.4×10^{-06}	43.78
KLK3 exons 2-3	1.8×10^{-06}	9.9×10^{-05}	34.60
KLK3 exons 1-2	2.3×10^{-06}	1.3×10^{-04}	34.04
TMPRSS2:ERG	1.7×10^{-05}	8.8×10^{-04}	29.69
UPK2	1.7×10^{-05}	8.8×10^{-04}	29.70
ERG 3'	2.2×10^{-05}	0.001	29.09
STEAP2	2.9×10^{-05}	0.001	28.52
DLX1	3.1×10^{-05}	0.002	28.35
KLK4	3.6×10^{-05}	0.002	28.00
HPN	8.5×10^{-05}	0.004	26.12
ERG 5'	1×10^{-04}	0.005	25.73
PSGR	1.3×10^{-04}	0.01	25.08
PCA3	3.6×10^{-04}	0.02	22.84
KLK2	4.1×10^{-04}	0.02	22.56
CAMKK2	6.5×10^{-04}	0.03	21.49

Table 3.19 Transcripts differentially expressed between cancer (A, H, I, L) and non-cancer samples (Mann Whitney U test).

Transcript	*p - value*	*Adjusted p - value*	*Log2 Fold Change*
DLX1	3.2×10^{-06}	**0.0002**	**1.33**
ERG 3'	4.25×10^{-05}	**0.002**	**1.25**
TMPRSS2:ERG	1.19×10^{-04}	**0.006**	**0.93**
HOXC4	2.6×10^{-04}	**0.013**	**0.635**
ERG 5'	1.73×10^{-05}	**0.001**	**0.281**
HOXC6	4.97×10^{-05}	**0.002**	**0.242**
PCA3	2.02×10^{-04}	**0.01**	**0.225**
M.genitalium RplB	4.48×10^{-04}	**0.022**	**0.144**
HPN	9.02×10^{-06}	**0.0005**	**0.127**

Eleven transcripts were significantly differentially expressed between aggressive and non-aggressive cancers ($p < 0.05$, Mann-Whitney U test, Table 3.20). Three of these transcripts were up regulated in the aggressive cancer; *SLC12A1*, *UPK2* and *SPINK1* (Figure 3.21). *SLC12A1* and *UPK2* are tissue specific controls for kidney and bladder, respectively. Advanced tumours often become more solidified and firm which might cause the release of cells and EVs from these prostates to be inhibited. This would cause a relative increase in detection of transcripts from other sources such as the kidney and bladder. Note that *SLC12A1*, *UPK2* and *SPINK1* were heavily correlated across all of the samples (section 3.3.7) and so this result should be taken with some caution. Eight transcripts were down-regulated in the aggressive cancers, again I hypothesise that this is due to a decreased level of cells and EVs emerging from the prostate and it's cancer via DRE, as these transcripts are mostly either prostate or cancer specific.

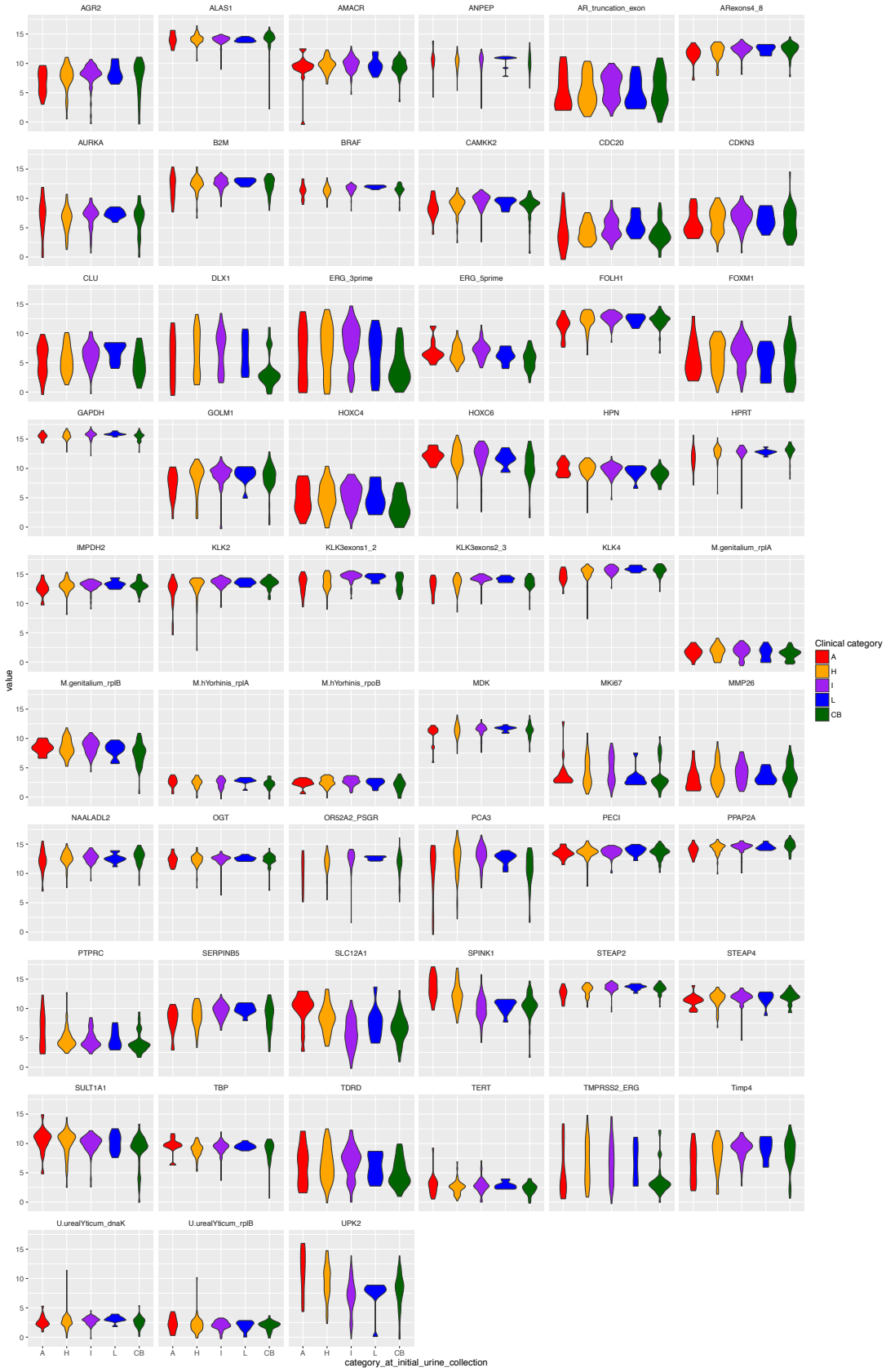


Figure 3.19 Violin plots showing distribution of each probe across clinical category.

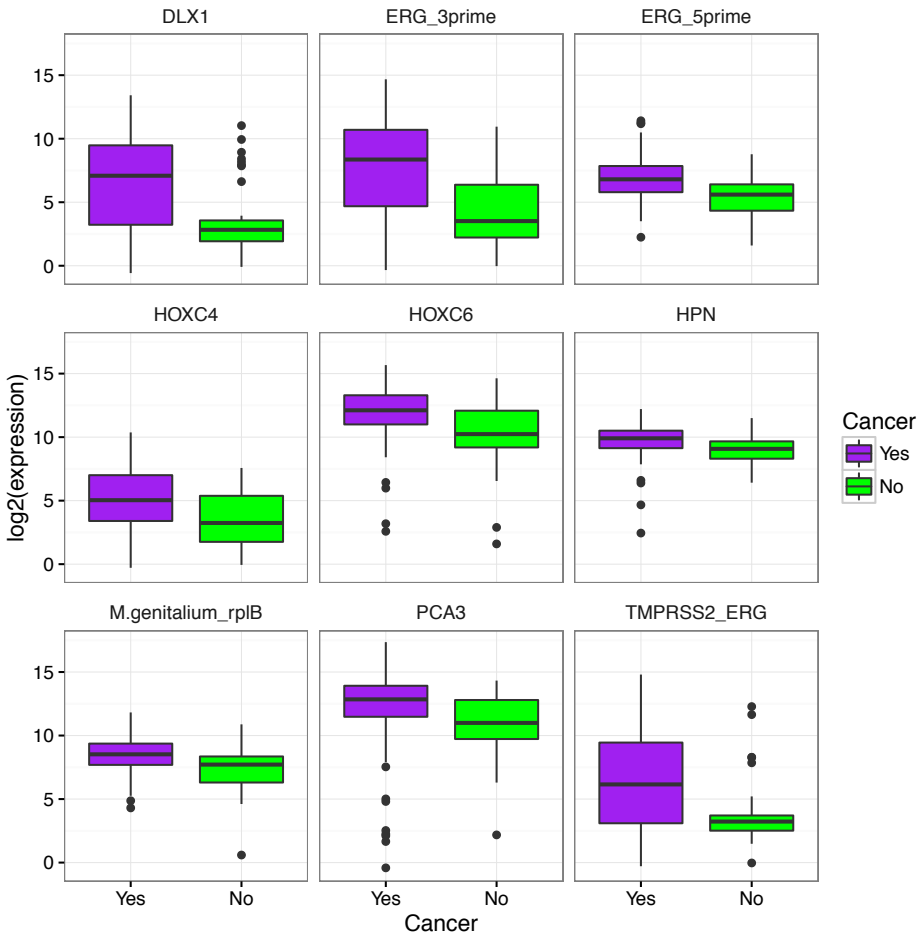


Figure 3.20 Boxplots showing the expression levels in significantly differentially expressed genes between cancer and non-cancer samples found by Mann Whitney U test.

Six transcripts were significantly differentially expressed between Advanced and CB ($p < 0.05$, Mann-Whitney U test, after multiple testing correction, Table 3.21). *SLC12A1* and *SPINK1* are up-regulated as has been previously discussed. The other four transcripts were down-regulated in the advanced samples, and again these include prostate specific transcripts such as *KLK4* and cancer related transcripts such as *PPAP2A* and *STEAP2* (Figure 3.22).

Table 3.20 Transcripts differentially expressed between aggressive cancer and non-aggressive samples (Mann Whitney U test).

Transcript	*p - value*	*Adjusted p - value*	*Log2 Fold Change*
SLC12A1	$2.86x10^{-08}$	$1.6x10^{-06}$	**0.59**
UPK2	$2.29x10^{-07}$	$1.26x10^{-05}$	**0.52**
SPINK1	$3.94x10^{-10}$	$2.25x10^{-08}$	**0.29**
SERPINB5	$2.78x10^{-04}$	$1.36X10^{-02}$	**-0.13**
CAMKK2	$4.44x10^{-04}$	$2.13x10^{-02}$	**-0.13**
PSGR	$9.55x10^{-05}$	$4.77x10^{-03}$	**-0.11**
KLK3 exons 1-2	$6.18x10^{-07}$	$3.34x10^{-05}$	**-0.1**
KLK3 exons 2-3	$9.15x10^{-07}$	$4.85x10^{-05}$	**-0.07**
KLK2	$4.77x10^{-04}$	$2.24x10^{-02}$	**-0.05**
STEAP2	$3.38x10^{-05}$	$1.76x10^{-03}$	**-0.05**
KLK4	$8.02x10^{-05}$	$4.09x10^{-03}$	**-0.04**

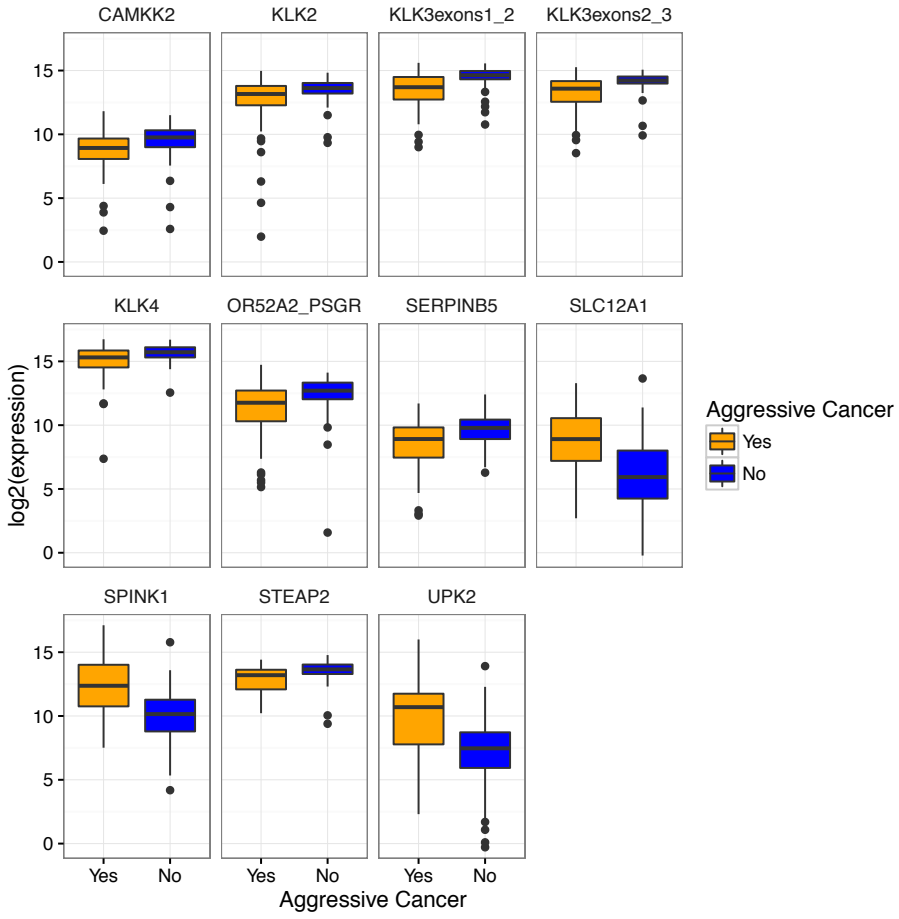


Figure 3.21 Boxplots showing differential expression between aggressive cancer and not aggressive PCa samples for those deemed significant by Mann Whitney U test.

Table 3.21 differentially expressed transcripts when comparing advanced samples with benign (no evidence of cancer) samples (Mann Whitney U test).

Transcript	*p - value*	*Adjusted p - value*	*Log2 Fold Change*
SLC12A1	$6.24x10^{-06}$	$3.49x10^{-04}$	*0.68*
SPINK1	$1.08x10^{-06}$	$6.14x10^{-05}$	*0.35*
HPRT	$1.29x10^{-04}$	$7.1x10^{-03}$	*-0.17*
KLK4	$1.53x10^{-04}$	$8.29x10^{-03}$	*-0.12*
STEAP2	$6.52x10^{-04}$	$3.39x10^{-02}$	*-0.09*
PPAP2A	$5.6x10^{-04}$	$2.97x10^{-02}$	*-0.07*

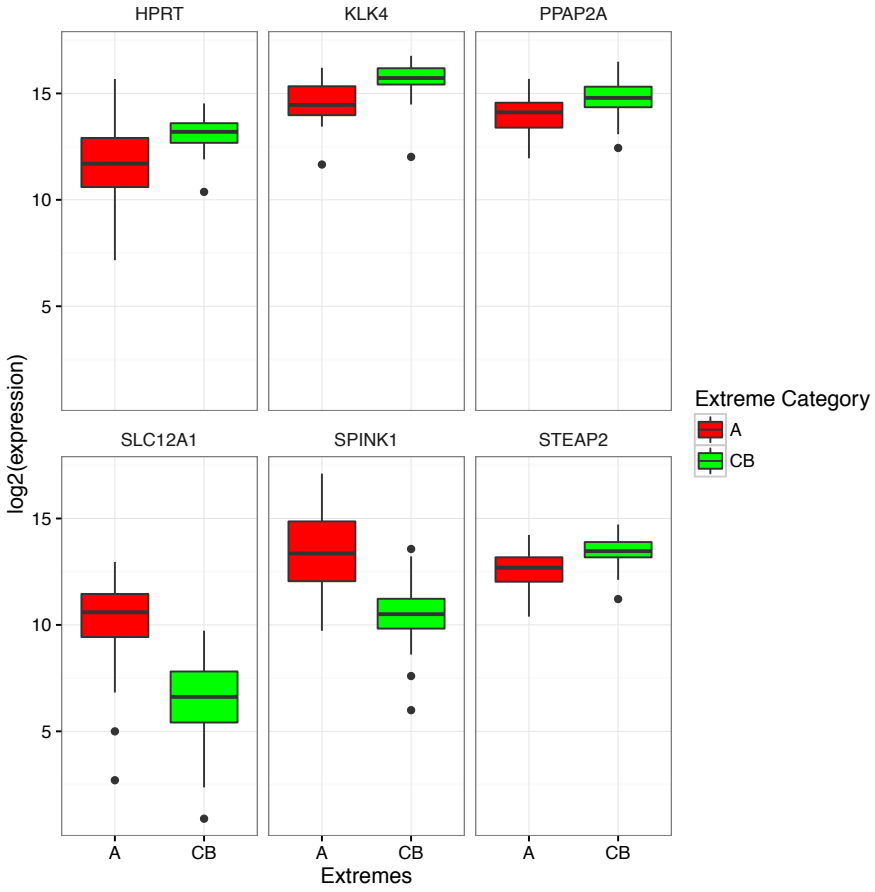


Figure 3.22 Boxplots showing differential expression between advanced cancer and non-cancer samples for those deemed significant via Mann Whitney U testing.

3.7 Low-risk, intermediate-risk and high-risk trend

Five probes showed significant increasing or decreasing expression trend with increasing D'Amico risk category (Spearman's correlation, $p < 0.05$ after multiple testing correction, Table 3.22). Three of these probes were identified to be highly correlated in general (*SPINK1*, *UPK2* and *SLC12A1*: section 3.3.7). The other two probes are from the same transcript (*KLK3*). The decrease

in *KLK3* with increasing cancer risk has been reported in previous prostate tissue studies (section 1.4.1) and urine.

Table 3.22 Spearman's correlation results comparing expression with ordered clinical categories: Low-, Intermediate- and High-risk.

Transcript	*p – value*	*Adjusted p - value*	*R*
SPINK1	$1.27x10^{-06}$	$7.24x10^{-05}$	**0.41**
UPK2	$2.74x10^{-05}$	$1.53x10^{-03}$	**0.36**
SLC12A1	$1.08x10^{-04}$	$5.96x10^{-03}$	**0.33**
KLK3 exons 2-3	$1.12x10^{-04}$	$6.04x10^{-03}$	**-0.33**
KLK3 exons 1-2	$1.14x10^{-04}$	$6.04x10^{-03}$	**-0.33**

3.8 Clinical Prediction models

To test the ability of NanoString data derived from urine EVs to predict the presence of cancer and/or it's aggressiveness various models were produced to distinguish between a) PCa and benign samples, b) aggressive PCa and non-aggressive PCa and c) advanced and benign samples (Table 3.3). All samples were used in the training set due to the pilot nature of this work. The modelling techniques applied here are logistic regression models using step wise variable selection (section 2.6.4), Lasso logistic regression models for shrinkage and variable selection (section 2.6.2), and random forest (section 2.6.3).

3.8.1 Logistic regression models using step wise variable selection

The optimal output cancer vs. non-cancer model contained 33 transcripts and had an AIC score of 68 (Table 3.23), the optimal aggressive cancer vs. non-aggressive cancer model contained 37 transcripts and had an AIC score of 76 (

Table 3.25), and the optimal model for distinguishing Advanced cancer from CB contained 9 transcripts and had an AIC score of 18 (Table 3.27). In each model the sample category was predicted with 100% sensitivity, 100% specificity, and 100% PPV (Table 3.24, Table 3.26, Table 3.28). This may mean the models are over-fitting the data and caution should be taken.

Table 3.23 Transcripts in the Step derived model for comparing cancer to non-cancer.

Transcript	*p - value*	*Coefficient*
CDKN3	***0.97***	***374.2***
FOLH1	***0.97***	***-838.3***
FOXM1	***0.97***	***138.8***
HPN	***0.97***	***743.9***
IMPDH2	***0.97***	***691.2***
KLK3 exons 2-3	***0.97***	***1844***
M.genitalium RplB	***0.97***	***-491.5***
NAALADL2	***0.97***	***-549.9***
AURKA	***0.98***	***-236.7***
BRAF	***0.98***	***562.1***
KLK2	***0.98***	***-1534.6***
M.hyorhinis RplA	***0.98***	***-1415.5***
PSGR	***0.98***	***364.4***
SULT1A1	***0.98***	***686***
TMPRSS2:ERG	***0.98***	***283.5***
ANPEP	***0.99***	***876.6***
AR truncation exon	***0.99***	***-232.3***
AR exons 4-8	***0.99***	***-462.4***
B2M	***0.99***	***1219.1***
CAMKK2	***0.99***	***-432.1***
DLX1	***0.99***	***220.4***
ERG 3'	***0.99***	***240.6***
ERG 5'	***0.99***	***985.7***
KLK4	***0.99***	***-1187.3***
MDK	***0.99***	***-1265.6***
MMP26	***0.99***	***-730***
OGT	***0.99***	***-1185.7***
PCA3	***0.99***	***399.8***
SERPINB5	***0.99***	***-234.1***
TBP	***0.99***	***497***
U.urealyticum dnaK	***0.99***	***-919.6***
U.urealyticum RplB	***0.99***	***837.9***
UPK2	***0.99***	***-132.5***

Table 3.24 Category predictions using the cancer vs. non-cancer step model.

Test	*Actual Category*	
	Disease Present	***No evidence of disease***
Positive	***148***	***0***
Negative	***0***	***40***

Table 3.25 Transcripts in the Step derived model for comparing aggressive cancers (A, H) to non-aggressive cancers (I, L).

Transcript	*p - value*	*Coefficient*
AGR2	***0.98***	***-233.03***
AMACR	***0.98***	***-1066.97***
AURKA	***0.98***	***-395.93***
BRAF	***0.98***	***-1106.78***
ERG 5'	***0.98***	***164.37***
FOXM1	***0.98***	***149.23***
HPRT	***0.98***	***317.38***
IMPDH2	***0.98***	***913.69***
KLK3 exons 1-2	***0.98***	***-607.92***
M.genitalium RplA	***0.98***	***-369.55***
M.hyorinis RplA	***0.98***	***1236.48***
MKi67	***0.98***	***-200.18***
NAALADL2	***0.98***	***204.06***
PSGR	***0.98***	***-826.31***
PCA3	***0.98***	***190.98***
PPAP2A	***0.98***	***546.96***
SERPINB5	***0.98***	***-357.41***
SPINK1	***0.98***	***708.99***
STEAP4	***0.98***	***585.87***
SULT1A1	***0.98***	***144.98***
TMPRSS2:ERG	***0.98***	***-109.31***
Timp4	***0.98***	***-304.36***
U.urealyticum dnaK	***0.98***	***390.71***
U.urealyticum RplB	***0.98***	***-456.65***
AR exons 4-8	***0.99***	***148.66***
CDC20	***0.99***	***-180.92***
DLX1	***0.99***	***47.8***
FOLH1	***0.99***	***492.55***
GAPDH	***0.99***	***-366.81***
GOLM1	***0.99***	***499.32***
M.hyorinis rpoB	***0.99***	***218.58***
PTPRC	***0.99***	***-126.41***
TBP	***0.99***	***-224.34***
TDRD	***0.99***	***-160.03***
TERT	***0.99***	***156.96***
UPK2	***0.99***	***74.34***
AGR2	***0.98***	***-233.03***

Table 3.26 Category Predictions when using the aggressive cancer model derived from Step.

	Aggressive Disease Present	**Aggressive undetectable**
Positive	**68**	**0**
Negative	**0**	**80**

Table 3.27 Transcripts in the extreme model (A Vs. CB) derived from Step.

Transcript	*p - value*	*Coefficient*
ALAS1	***0.995***	***-600.84***
KLK4	***0.995***	***-465.6***
KLK3 exons 2-3	***0.995***	***330.63***
BRAF	***0.995***	***422.54***
M.genitalium RplB	***0.995***	***302.06***
HPN	***0.995***	***236.54***
Timp4	***0.995***	***-158.99***
AR truncation exon	***0.995***	***31.34***
ALAS1	***0.995***	***-600.84***

Table 3.28 Category predictions using the extreme model derived from Step.

Test	*Actual Category*	
	Advanced Cancer Present	***CB***
Positive	***17***	***0***
Negative	***0***	***40***

3.8.2 Lasso logistic regression models

The cancer vs. non-cancer model had 16 transcripts (Table 3.29), an AUC of 0.937 and 99.32% sensitivity, 52.5% specificity and 88.55% PPV (Table 3.30). The aggressive cancer (A) vs. non-aggressive cancer model had four transcripts (Table 3.29), an AUC of 0.852, and 61.76% sensitivity, 86.25% specificity and 79.25% PPV (Table 3.31). The extreme Lasso model (A Vs. CB samples) had 12 transcripts (Table 3.29), an AUC of 0.983 and 82.35% sensitivity, 100% specificity and 100% PPV (Table 3.32).

Table 3.29 Lasso coefficients for three models A) cancer Vs. Non-cancer B) Aggressive cancer Vs. Non-aggressive cancer C) extreme model (A Vs. CB)

Cancer Model		*Aggressive Model*		*Extreme Model*	
Gene name	***Coefficient***	***Gene name***	***Coefficient***	***Gene name***	***Coefficient***
ALAS1	***-0.042***	***KLK3 exons 1-2***	***-0.268***	***AURKA***	***0.035***
AR exons4-8	***-0.278***	***SERPINB5***	***-0.087***	***DLX1***	***0.043***
CLU	***0.037***	***SLC12A1***	***0.069***	***ERG 5'***	***0.235***
DLX1	***0.164***	***SPINK1***	***0.278***	***HOXC4***	***0.003***
ERG 3'	***0.082***			***HPN***	***0.188***
ERG 5'	***0.169***			***HPRT***	***-0.554***
HOXC4	***0.111***			***PPAP2A***	***-0.223***
HPN	***0.283***			***SLC12A1***	***0.104***
IMPDH2	***-0.131***			***SPINK1***	***0.321***
KLK2	***-0.046***			***STEAP2***	***-0.297***
KLK3 exons 1-2	***0.05***			***STEAP4***	***-0.113***
M.hyorhinis rpoB	***0.399***			***SULT1A1***	***0.118***
MMP26	***-0.085***				
NAALADL2	***-0.003***				
PCA3	***0.101***				
PPAP2A	***-0.62***				
SULT1A1	***0.057***				
TMPRSS2:ERG	***0.037***				
Timp4	***-0.033***				

Table 3.30 Category predictions using the Lasso cancer Vs. non-cancer model

Test	*Actual Category*	
	Disease Present	***Disease Absent***
Positive	***147***	***19***
Negative	***1***	***21***

Table 3.31 Category predictions using the Lasso aggressive cancer Vs. non-aggressive cancer model

Test	*Actual Category*	
	Aggressive Disease Present	***Aggressive Disease Absent***
Positive	***42***	***11***
Negative	***26***	***69***

Table 3.32 Category prediction for the Lasso extreme model (A Vs. CB)

Test	*Actual Category*	
	Advanced Cancer Present	***CB***
Positive	***14***	***0***
Negative	***3***	***40***

3.8.3 Random Forest

The cancer vs. non-cancer model had an OOB error estimate of 18.52%, with 87.25% sensitivity, 60% specificity and 89.04% PPV (Table 3.33, the ranked transcript importance provided in Table 3.34). The aggressive vs. non-aggressive cancer model had an OOB error estimate of 22.82%, with 71.64% sensitivity, 81.71% specificity and 76.19% PPV (Table 3.35, Table 3.36). The extremes model had an OOB error estimate of 15.79%, with 70.59% sensitivity, 90% specificity and 75% PPV (

Table 3.37, Table 3.38).

Table 3.33 Confusion matrix for random forest modelling samples on cancer vs. non-cancer. OOB error estimate of 18.52%.

	Cancer not predicted	*Cancer predicted*	*Class error*	*Sum*
CB	***24***	***16***	***0.4***	***40***
Cancer	***19***	***130***	***0.13***	***149***

Table 3.34 Gini values for the random forest model to categorise the samples into cancer and non-cancer.

Transcript	*Gini*	*Rank*	*Transcript*	*Gini*	*Rank*	*Transcript*	*Gini*	*Rank*
DLX1	***2.27***	***1***	***CLU***	***0.61***	***20***	***UPK2***	***0.46***	***39***
ERG 3'	***2.08***	***2***	***SPINK1***	***0.61***	***21***	***GAPDH***	***0.45***	***40***
TMPRSS2:ERG	***2.06***	***3***	***B2M***	***0.60***	***22***	***Timp4***	***0.44***	***41***
HOXC6	***1.86***	***4***	***FOXM1***	***0.58***	***23***	***KLK3 exons 2-3***	***0.44***	***42***
HPN	***1.83***	***5***	***AR exons 4-8***	***0.58***	***24***	***SERPINB5***	***0.44***	***43***
PCA3	***1.21***	***6***	***CAMKK2***	***0.56***	***25***	***CDC20***	***0.42***	***44***
PPAP2A	***1.16***	***7***	***SULT1A1***	***0.56***	***26***	***PECI***	***0.42***	***45***
ERG 5'	***1.14***	***8***	***M.genitalium RplA***	***0.54***	***27***	***AR truncation exon***	***0.41***	***46***
HOXC4	***1.12***	***9***	***STEAP2***	***0.54***	***28***	***U.urealYticum RplB***	***0.41***	***47***
M.hYorhini rpoB	***1.04***	***10***	***MKi67***	***0.54***	***29***	***GOLM1***	***0.41***	***48***
ALAS1	***0.85***	***11***	***MMP26***	***0.53***	***30***	***AURKA***	***0.40***	***49***
PTPRC	***0.77***	***12***	***KLK2***	***0.52***	***31***	***ANPEP***	***0.40***	***50***
M.genitalium RplB	***0.77***	***13***	***AGR2***	***0.50***	***32***	***MDK***	***0.39***	***51***
SLC12A1	***0.75***	***14***	***AMACR***	***0.49***	***33***	***U.urealYticum dnaK***	***0.38***	***52***
HPRT	***0.75***	***15***	***FOLH1***	***0.49***	***34***	***CDKN3***	***0.37***	***53***
KLK3 exons 1-2	***0.71***	***16***	***TBP***	***0.48***	***35***	***BRAF***	***0.36***	***54***
NAALADL2	***0.67***	***17***	***TERT***	***0.48***	***36***	***KLK4***	***0.36***	***55***
TDRD	***0.63***	***18***	***IMPDH2***	***0.47***	***37***	***PSGR***	***0.33***	***56***
STEAP4	***0.61***	***19***	***M.hYorhinis RplA***	***0.46***	***38***	***OGT***	***0.30***	***57***

Table 3.35 Confusion matrix for random forest modelling samples on aggressive cancer vs. non-aggressive cancer. OOB error estimation of 22.82%.

	Aggressive Cancer not predicted	*Aggressive Cancer predicted*	*Class error*	*Sum*
Non-aggressive cancer	***67***	***15***	***0.18***	***82***

Aggressive Cancer	***19***	***48***	***0.28***	***67***

Table 3.36 Gini values for the random forest model to categorise the samples into aggressive cancer and non-aggressive cancer.

Transcript	*Gini score*	*Rank*	*Transcript*	*Gini score*	*Rank*	*Transcript*	*Gini score*	*Rank*
SPINK1	***5.21***	***1***	***ERG 3'***	***1.00***	***20***	***HOXC6***	***0.72***	***39***
KLK3 exons 1-2	***4.07***	***2***	***MDK***	***0.99***	***21***	***TERT***	***0.72***	***40***
KLK3 exons 2-3	***3.63***	***3***	***CAMKK2***	***0.97***	***22***	***NAALADL2***	***0.71***	***41***
UPK2	***3.55***	***4***	***CLU***	***0.95***	***23***	***ERG 5'***	***0.69***	***42***
SLC12A1	***3.48***	***5***	***GAPDH***	***0.93***	***24***	***AR exons 4-8***	***0.68***	***43***
SERPINB5	***1.89***	***6***	***AGR2***	***0.87***	***25***	***M.genitalium RplA***	***0.66***	***44***
SULT1A1	***1.88***	***7***	***TMPRSS2:ERG***	***0.86***	***26***	***PCA3***	***0.63***	***45***
KLK4	***1.86***	***8***	***CDKN3***	***0.86***	***27***	***DLX1***	***0.62***	***46***
BRAF	***1.57***	***9***	***B2M***	***0.85***	***28***	***TBP***	***0.62***	***47***
PSGR	***1.49***	***10***	***TDRD***	***0.83***	***29***	***PECI***	***0.62***	***48***
HPN	***1.39***	***11***	***ALAS1***	***0.82***	***30***	***MMP26***	***0.59***	***49***
U.urealYticum dnaK	***1.28***	***12***	***AR truncation exon***	***0.82***	***31***	***M.hYorhinis rpoB***	***0.59***	***50***
Timp4	***1.25***	***13***	***IMPDH2***	***0.81***	***32***	***OGT***	***0.59***	***51***
STEAP2	***1.24***	***14***	***GOLM1***	***0.79***	***33***	***HOXC4***	***0.58***	***52***
U.urealYticum RplB	***1.09***	***15***	***FOLH1***	***0.79***	***34***	***AMACR***	***0.57***	***53***
M.hYorhinis RplA	***1.08***	***16***	***FOXM1***	***0.79***	***35***	***ANPEP***	***0.56***	***54***
KLK2	***1.08***	***17***	***MKi67***	***0.77***	***36***	***HPRT***	***0.54***	***55***
PPAP2A	***1.07***	***18***	***M.genitalium RplB***	***0.73***	***37***	***PTPRC***	***0.52***	***56***
CDC20	***1.01***	***19***	***AURKA***	***0.73***	***38***	***STEAP4***	***0.50***	***57***

Table 3.37 Confusion matrix for random forest modelling the samples belonging to the extreme clinical categories (A vs. CB). OOB error estimate of 15.79%.

	CB predicted	*Advanced predicted*	*Class error*	*Sum*
CB	***36***	***4***	***0.1***	***40***
Advanced	***5***	***12***	***0.29***	***17***

Table 3.38 Gini values for the random forest model to categorise the extreme samples (A vs. CB).

Transcript	*Gini score*	*Rank*	*Transcript*	*Gini score*	*Rank*	*Transcript*	*Gini score*	*Rank*
SPINK1	***5.21***	***1***	***ERG 3'***	***1.00***	***20***	***HOXC6***	***0.72***	***39***
KLK3 exons 1-2	***4.07***	***2***	***MDK***	***0.99***	***21***	***TERT***	***0.72***	***40***
KLK3 exons 2-3	***3.63***	***3***	***CAMKK2***	***0.97***	***22***	***NAALADL2***	***0.71***	***41***
UPK2	***3.55***	***4***	***CLU***	***0.95***	***23***	***ERG 5'***	***0.69***	***42***
SLC12A1	***3.48***	***5***	***GAPDH***	***0.93***	***24***	***AR exons 4-8***	***0.68***	***43***
SERPINB5	***1.89***	***6***	***AGR2***	***0.87***	***25***	***M.genitalium RplA***	***0.66***	***44***
SULT1A1	***1.88***	***7***	***TMPRSS2:ERG***	***0.86***	***26***	***PCA3***	***0.63***	***45***
KLK4	***1.86***	***8***	***CDKN3***	***0.86***	***27***	***DLX1***	***0.62***	***46***
BRAF	***1.57***	***9***	***B2M***	***0.85***	***28***	***TBP***	***0.62***	***47***
PSGR	***1.49***	***10***	***TDRD***	***0.83***	***29***	***PECI***	***0.62***	***48***
HPN	***1.39***	***11***	***ALAS1***	***0.82***	***30***	***MMP26***	***0.59***	***49***
U.urealYticum dnaK	***1.28***	***12***	***AR truncation exon***	***0.82***	***31***	***M.hYorhinis rpoB***	***0.59***	***50***
Timp4	***1.25***	***13***	***IMPDH2***	***0.81***	***32***	***OGT***	***0.59***	***51***
STEAP2	***1.24***	***14***	***GOLM1***	***0.79***	***33***	***HOXC4***	***0.58***	***52***
U.urealYticum RplB	***1.09***	***15***	***FOLH1***	***0.79***	***34***	***AMACR***	***0.57***	***53***
M.hYorhinis RplA	***1.08***	***16***	***FOXM1***	***0.79***	***35***	***ANPEP***	***0.56***	***54***
KLK2	***1.08***	***17***	***MKi67***	***0.77***	***36***	***HPRT***	***0.54***	***55***
PPAP2A	***1.07***	***18***	***M.genitalium RplB***	***0.73***	***37***	***PTPRC***	***0.52***	***56***
CDC20	***1.01***	***19***	***AURKA***	***0.73***	***38***	***STEAP4***	***0.50***	***57***

3.8.4 Random Forest applied to all clinical categories

A random forest model was also constructed to classify the samples into their five main types of category (advanced, high-, intermediate-, low-risk and clinically benign), but the results were poor (OOB error estimate of 45.5%, Table 3.39, Table 3.40). The OOB was only modestly improved when the data was adjusted to have equal numbers of samples per category in each tree (53.44%,

Table 3.41). This poor performance could be down to the low number of samples per category using the current methods.

Table 3.39 Confusion matrix for random forest on all 5 clinical categories. OOB error estimate of 45.5%.

	A	*H*	*I*	*L*	*CB*	*Class error*	*Sum*
A	***1***	***12***	***4***	***0***	***0***	***0.94***	***17***
H	***1***	***20***	***21***	***0***	***8***	***0.6***	***50***
I	***1***	***8***	***58***	***0***	***5***	***0.19***	***72***
L	***0***	***0***	***6***	***0***	***4***	***1***	***10***
CB	***1***	***6***	***9***	***0***	***24***	***0.4***	***40***

Table 3.40 Sensitivity, Specificity and PPV for each category after categorising samples into five clinical categories using random forest.

Advanced (A)			
	True A	***True Not A***	***Sensitivity: 25%***
Outcome A	***1***	***16***	***Specificity: 91.35%***
Outcome Not A	***3***	***169***	***PPV: 5.88%***

High Risk (H)			
	True H	***True Not H***	***Sensitivity: 40%***
Outcome H	***20***	***26***	***Specificity: 81.29%***
Outcome Not H	***30***	***113***	***PPV: 43.48%***

Intermediate Risk (I)			
	True I	***True Not I***	***Sensitivity: 80.56%***
Outcome I	***58***	***40***	***Specificity: 65.81%***
Outcome Not I	***14***	*77*	***PPV: 59.18%***

Low Risk (L)			
	True L	***True Not L***	***Sensitivity: 0%***
Outcome L	***0***	***0***	***Specificity: 100%***
Outcome Not L	***10***	***179***	***PPV: *%***

Clinically Benign (CB)			
	True CBN	***True Not CBN***	***Sensitivity: 60%***
Outcome CBN	***24***	*17*	***Specificity: 88.59%***
Outcome Not CBN	***16***	***132***	***PPV: 58.54%***

Table 3.41 Confusion matrix for random forest on all 5 categories with random sampling to equalise categorical sample sizes. OOB error estimate of 53.44%.

	A	*H*	*I*	*L*	*CB*	*Class error*	*Sum*
A	***10***	***4***	***2***	***0***	***1***	***0.41***	***17***
H	***14***	***7***	***15***	***5***	***9***	***0.86***	***50***
I	***3***	***8***	***46***	***4***	***11***	***0.36***	***72***
L	***1***	***0***	***4***	***1***	***4***	***0.9***	***10***

CB	***2***	***5***	***3***	***6***	***24***	***0.4***	***40***

3.8.5 Comparing the Models

The OOB error was found to be lowest for modelling the extremes (CB v Aggressive PCa), this was expected as they are the samples that should be least alike in their expression and so should be the easiest categories to separate. The model does give good sensitivity and specificity; however, this error is still fairly high at 15.79%, meaning there could still be improvements. Similarly, the Lasso model (high AUC) and Step model for the extremes comparison both had high sensitivity and specificity, though the step models are likely to be over fitted. From the top fifteen most important transcripts via random forest, five were in common (four uniquely) with the Lasso selected transcripts and five were in common (four uniquely) with the Step selected transcripts. The only transcript common to all three models was HPN, which interestingly only appeared to have mid level importance in each model.

The OOB error for comparing cancer vs. non-cancer was also fairly high, even though it was the second lowest (18.52%). This model showed high sensitivity but was not so specific to identifying cancer in the samples. The Lasso model had a good AUC (0.937) and also showed high sensitivity but not so good specificity to detecting cancer, unlike the Step model, which showed high sensitivity and specificity. However, Step is the least robust of the methods for modelling data. From the top fifteen most important transcripts via random forest, ten were common in the Lasso selected transcripts and seven were common with the Step selected transcripts. There were six transcripts common to all three models: *DLX1, ERG* 3', *TMPRSS2:ERG, HPN, PCA3* and *ERG* 5', all of which are transcripts known to be involved or associated to PCa.

The OOB error for comparing aggressive cancers to non-aggressive cancers was higher at 22.82%, though this model had good sensitivity and specificity ratios for the random forest model, 72% and 82%, respectively. From the top fifteen most important transcripts selected via random forest, all four Lasso identified transcripts were in common and ten Step selected transcripts were common. Three of the fours Lasso identified transcripts were common to all models: *SPINK1*, *KLK3* exons 1-2 and *SERPINB5*.

This highlights that there is structure in the data that could likely be further improved with data from more samples and more probes.

3.9 Transcripts that show high-importance

Seven transcripts were identified by three different methods (Table 3.19, Table 3.29, Table 3.34) to be differentially expressed between cancer and non-cancer samples: *DLX1*, *ERG* 3', *TMPRSS2:ERG*, *HOXC4*, *ERG* 5', *PCA3* and *HPN* (Table 3.40). These transcripts all have published associations with PCa. Interestingly, *ERG* 5', *HOXC6* and *M.genitalium* RplB were significant in the Mann Whitney U test and had been ranked highly by random forest, but were not present in the Lasso model. This is likely due to the inter-correlation of their NanoString signals, as Lasso penalises correlating variables and keeps those it deems to hold the most information (section 3.3.7).

Table 3.42 Transcripts identified to distinguish between PCa and non-cancer using Mann Whitney U and Lasso. Random Forest rank is also shown.

Transcript	*Mann Whitney U*	*Lasso*	*Random Forest rank*
DLX1	***Y***	***Y***	***1***
ERG 3'	***Y***	***Y***	***2***
TMPRSS2:ERG	***Y***	***Y***	***3***
HOXC4	***Y***	***Y***	***9***
ERG 5'	***Y***	***Y***	***8***
HOXC6	***Y***	***N***	***4***
PCA3	***Y***	***Y***	***6***
M.genitalium RplB	***Y***	***N***	***13***
HPN	***Y***	***Y***	***5***
PPAP2A	***N***	***Y***	***7***
M.hYorhini rpoB	***N***	***Y***	***10***
ALAS1	***N***	***Y***	***11***
KLK3 exons 1-2	***N***	***Y***	***16***
NAALADL2	***N***	***Y***	***17***
CLU	***N***	***Y***	***20***
SULTA1	***N***	***Y***	***26***
MMP26	***N***	***Y***	***30***
KLK2	***N***	***Y***	***31***
IMPDH2	***N***	***Y***	***37***
Timp4	***N***	***Y***	***41***
PTPRC	***N***	***N***	***12***
SLC12A1	***N***	***N***	***14***
HPRT	***N***	***N***	***15***

Four transcripts identified by three different methods (Table 3.20, Table 3.31, Table 3.39) were differentially expressed between aggressive cancer and non-aggressive cancer samples: *SLC12A1, SPINK1, SERPINB5* and *KLK3* exons 1-2.

Table 3.43 Transcripts repeatedly shown to be differentially expressed between aggressive PCa and non-aggressive PCa.

Transcript	*Mann Whitney U*	*Lasso*	*Random Forest rank*
SLC12A1	***Y***	***Y***	***5***
UPK2	***Y***	***N***	***4***
SPINK1	***Y***	***Y***	***1***
SERPINB5	***Y***	***Y***	***6***
CAMKK2	***Y***	***N***	***22***
PSGR	***Y***	***N***	***10***
KLK3 exons 1-2	***Y***	***Y***	***2***
KLK3 exons 2-3	***Y***	***N***	***3***
KLK2	***Y***	***N***	***17***
STEAP2	***Y***	***N***	***14***
KLK4	***Y***	***N***	***8***
SULT1A1	***N***	***N***	***7***
BRAF	***N***	***N***	***9***

Two of these transcripts were also identified by three different methods (Table 3.21, Table 3.32, Table 3.38) to be differentially expressed between advanced cancer and clinically benign cancer samples: *SLC12A1* and *SPINK1*. It is perplexing that less transcripts are selected for this extreme comparison, but this may be due to a lack of material coming from the solid advanced cancers.

Table 3.44 Transcripts commonly found to be differentially expressed by the Mann Whitney U test, GLM and Lasso and Random Forest between advanced and benign samples.

Transcript	*Mann Whitney U*	*Lasso*	*Random Forest rank*
SLC12A1	***Y***	***Y***	***5***
SPINK1	***Y***	***Y***	***1***
HPRT	***Y***	***Y***	***55***
KLK4	***Y***	***N***	***8***
STEAP2	***Y***	***N***	***14***
PPAP2A	***Y***	***N***	***18***
DLX1	***N***	***Y***	***46***
ERG 5'	***N***	***Y***	***42***
HOXC4	***N***	***Y***	***52***
HPN	***N***	***Y***	***11***
STEAP4	***N***	***Y***	***57***
SULT1A1	***N***	***Y***	***7***

3.10 Conclusions

Detection of prostate-specific (*KLK2* and *KLK3*) and PCa-specific (*TMPRSS2:ERG*) transcripts demonstrates that these are present in urine EVs harvested and analysed by our methods. RNA yields post-radical prostatectomy suggests that the vast majority of the urinary RNA originates in the prostate. The identification of differential transcripts between non-aggressive and aggressive cancers demonstrates NanoString's potential ability to distinguish these clinical categories using transcripts from urinary EVs.

It is vital to emphasise that the clinical categories in this study are based on current, and not perfect clinical tests. Hence the current need for novel biomarkers to distinguish accurately between them. Particularly true of CB samples, where it is expected that ~20% of the men that show no clinical evidence of cancer will in fact have PCa. Therefore, it is notable that 12% of CB samples were found to have a *TMPRSS2:ERG* fusion in this study. As *TMPRSS2:ERG* is expected to be in ~50% of PCa, this would suggest clinically undetected PCa in 24% of our CB samples. In LPD, 21 of the 37 CB samples are clustered together, leaving 16 spread amongst the other groups, five (14%) of which are associated to a group where overall *TMPRSS2:ERG* is significantly up-regulated. Seven of the CB samples are left un-grouped, showing no distinct underlying signature.

The detection of *TMPRSS2:ERG* by NanoString and confirmation of this find by RTPCR demonstrated the sensitivity of our methods for detection of PCa.

Some Nanostring probes performed much better than others in models throughout the analyses (section 3.9). However, transcripts were identified that were differentially expressed in samples from different clinical categories (PCa present, increasing PCa aggressiveness etc). Due to the nature of the probe set, most of these were known PCa markers, but some were not. The latter demonstrating that it can be difficult to predict what type of transcripts should be targeted in our analyses. As a result of probe selection for advanced PCa associated transcripts, it was expected that we observe unusual distribution and medians for our data.

Housekeeping probes were selected, as they are known to be useful when investigating PCa tissue; their use in urine and urinary EVs has not been studied in detail. Therefore, it may be of interest to investigate further options for urinary microvesicle house keeping transcripts, as the probes selected did not show great correlation.

The analyses have revealed that there is structure in the data, as demonstrated for example, by the detection of differentially expressed transcripts, LPD groups and linear model analysis. Lasso logistic regression predictive models were able to categorise cancer from non-cancer samples and aggressive from non-aggressive cancer samples fairly robustly (AUCs of 0.94 and 0.85, respectively). However, sensitivity and specificity, even on the training set could be improved. For this it could be suggested that we are not using the optimal starting probes and thus more probes should be included to identify the clearest signature available. Another complication is the complexity of cancer within individual prostates, with multifocal tumours being detectable in the majority of cancerous prostates, each with the potential to have a different path and rate of progression. The Mann Whitney U test and random forest results were similar to each other and that of the Lasso models, suggesting that these results are accurate, but also highlighting that the methods were suitable for analysing the NanoString data. The LPD showed some clinical separation of the samples, though again a better selection of probes could provide further discrimination between the lower and intermediate samples with the benign samples. The inclusion of known PCa transcripts in our differential expression and predictive model for cancer results and the inclusion of known prognostic PCa transcripts in our differential expression and predictive model for aggressive cancer results provide evidence of accuracy.

These analyses form the ground work for expansion of the urine biomarker study to include a larger number of probes, and samples which should provide much improved power to dissect the complexities of this disease within individual prostates. The probes that provided no information were determined. These probes were reviewed to unveil if they should be replaced or removed in the larger study (for example the *PCA3* probe didn't work very well and was redesigned for the

large study). *FOXM1* showed no clinical association and was not identified in any clustering or prediction models and so was removed from subsequent studies.

4

Response to Hormone Therapy

4.1 Summary

Stratification of treatment by gene expression levels has shown benefits in many cancers, such as breast cancer[208] and lung cancer[209,210], but it is yet to be utilised successfully in prostate cancer (PCa) treatment. Areas where stratification could benefit PCa patients include: deciding between treatment vs. active surveillance, identifying which radical prostatectomy (RP) and radiotherapy patients will succumb to biochemical recurrence (BCR), and which hormone therapy (HT) patients would benefit from additional treatment (i.e. those patients that are predicted to progress early to castration resistant prostate cancer (CRPC)). In this chapter we focus on men in our

cohort treated by hormone therapy and examine whether expression profiles of urinary microvesicles can be used to predict time to CRPC. Unfortunately, the cohort does not have a long enough follow up at this time to examine time to BCR after RP or radiotherapy.

In the normalised NanoString 1 dataset, a signature of seven transcripts was identified that could optimally predict progression of patients on hormone therapy (section 1.3.4.2.1) (cox-regression model; $p = 2.3x10^{-05}$; HR = 0.04288). The transcripts in the predictor are *AGR2*, *DLX1*, *KLK2*, *NAALADL2*, *AR* exons 4-8, *PPAP2A* and *AMACR*. This model was an independent predictor of progression when established clinical variables initial PSA, age, Gleason score and initial bone scan result were taken into account (cox-regression model; $p = 0.003$; HR = 0.03).

When the data was adjusted to *KLK2* levels, similar to *KLK3* adjustment used in the PCA3 test, an optimal model of three transcripts (*CAMKK2*, *PSGR* and *UPK)* was identified (cox-regression model; $p = 0.007$, *HR* = 1.0028). This model was not a significant predictor independent of established clinical factors (cox-regression model; $p = 0.14$; HR = 1.009).

Both of these models were applied to the second NanoString dataset but were not validated.

4.2 Introduction

4.2.1 The Research Gap

Hormone Therapy (HT) is the primary treatment of men with advanced prostate cancer, that is those diagnosed with a PSA > 100 or with evidence of metastatic spread (generally via bone scan). Response to the treatment is highly variable with some men failing to respond at all, whereas others take years to progress. All men will eventually progress to CRPC (section 1.3.4.2.2). Identification of men that are likely to relapse early could lead to more aggressive first line treatment being used, such as full

androgen blockage (section 1.3.4.2.1), combination with chemotherapy, or novel strategies such as combination with Aberiterone. Currently, there is no clinically available test to stratify advanced patients into those who will do well on HT and those that will quickly require further or alternative treatments.

4.2.2 Aim

I am to use the NanoString 1 data set (Chapter 3) from advanced patients (n = 32), to see if a significant predictor of early progression in patients on HT can be built and whether this predictor improves on current clinical information collected (e.g. PSA, Gleason score and bone scan). I will also attempt to validate these signatures in a second independent cohort (using the second NanoString data).

4.2.3 Summary of the HT patient cohort

The breakdown of the clinical data for the 32 patients on HT can be seen in Table 4.1. Many of the advanced patients are diagnosed as being advanced by a PSA > 100 and no biopsy is performed in these circumstances. Other patients with lower PSAs are determined to be advanced, by either a biopsy or a positive bone scan.

Table 4.1 Clinical summary of the hormone therapy cohort (n=32).

Clinical Variable	*Number of patients*
Gleason Score	
10	***0***
9	***8***
8	***4***
7 (4+3)	***4***
No Biopsy: Advanced	***16***
Bone Scan	
Positive	***18***
Negative	***13***
Unknown	***1***
PSA	***Median 98.7 (range: 9.6 – 2508)***
Age	***Median 78 (range: 55 - 98)***

4.3 Hormone Therapy Predictor constructed using Nanostring 1 data

4.3.1 Differentially expressed genes based on initial response, 12 month relapse and 24 month relapse

Five transcripts were significantly up regulated in those that had an initial response to HT ($n = 28$) compared to those that did not ($n = 4$) ($p < 0.05$; not adjusted for multiple testing; Mann-Whitney U test; Table 4.2). Three of these five transcripts, were capable of distinguishing patients that relapsed within 12 months ($n = 6$) (Table 4.2; $p < 0.05$; not adjusted for multiple testing): *STEAP4, AMACR, BRAF*. By 24 months, 14 patients were still responding to HT and 18 had progressed. Four different transcripts were significantly up regulated in patients still responding to treatment (Table 4.2; $p < 0.05$; not adjusted for multiple testing). These results need to be treated with caution due to the low numbers and the lack of significance after multiple testing correction.

Table 4.2 Mann-Whitney U test results for comparing samples that respond to HT and those that don't at different time points.

Initial Response			
Transcript	***p - value***	***Adjusted p - value***	***Log2 Fold change***
AGR2	***0.047***	***1***	***-0.57***
STEAP4	***0.024***	***1***	***-0.26***
HPRT	***0.029***	***1***	***-0.21***
AMACR	***0.034***	***1***	***-0.14***
BRAF	***0.04***	***1***	***-0.13***
12 Month Response			
Transcript	***p - value***	***Adjusted p - value***	***Log2 Fold change***
STEAP4	***0.019***	***0.98***	***-0.18***
AMACR	***0.01***	***0.57***	***-0.14***
BRAF	***0.033***	***0.98***	***-0.08***
24 Month Response			
Transcript	***p - value***	***Adjusted p - value***	***Log2 Fold change***
DLX1	***0.045***	***1***	***-1.28***
AR (truncation) exon 9	***0.025***	***1***	***-1.16***
AR exons 4-8	***0.018***	***1***	***-0.09***
TBP	***0.034***	***1***	***-0.08***

4.3.2 Survival analyses of time to progression after HT

Using the Mann-Whitney U test as described above lacks statistical power as time to progression is not taken into account, therefore I applied Cox's proportional hazards model and other survival analysis tools (section 2.8). Twelve probes were significant predictors of progression individually (Table 4.3; Cox regression model; $p < 0.05$; multiple testing correction not applied). There were no significant probes after multiple testing correction.

Expression for each gene was divided into two groups, low expression and high expression, using *k*-means to determine the threshold (section 2.5.3). Using these grouped data, ten transcripts were identified as having significant different times to progression ($p < 0.05$; log-rank test; Table 4.4), of which only one was significant after multiple testing correction: *NAALADL2*.

Table 4.3 Cox results for relapse to hormone therapy

Transcript	*p - value*	*Adjusted p - value*	*Hazard ratio*
KLK2	***0.011***	***0.62***	***0.74***
AMACR	***0.011***	***0.62***	***0.68***
DLX1	***0.011***	***0.62***	***0.87***
PPAP2A	***0.014***	***0.76***	***0.51***
STEAP4	***0.017***	***0.88***	***0.63***
PCA3	***0.034***	***1.00***	***0.87***
CDC20	***0.037***	***1.00***	***0.81***
KLK4	***0.039***	***1.00***	***0.63***
TDRD	***0.042***	***1.00***	***0.86***
STEAP2	***0.043***	***1.00***	***0.66***
NAALADL2	***0.045***	***1.00***	***0.79***
Timp4	***0.049***	***1.00***	***0.86***

Table 4.4 Significant probes using log rank test applied to data separated by *k*-means.

Transcript	*p - value*	*Adjusted p - value*	*Coefficient*
NAALADL2	***0.0004***	***0.03***	***12.36***
PPAP2A	***0.005***	***0.27***	***7.97***
KLK2	***0.006***	***0.31***	***7.66***
STEAP4	***0.007***	***0.4***	***7.19***
DLX1	***0.01***	***0.56***	***6.55***
AGR2	***0.01***	***0.64***	***6.28***
PCA3	***0.02***	***0.93***	***5.57***
IMPDH2	***0.03***	***0.98***	***4.68***
STEAP2	***0.03***	***0.98***	***4.63***

FOLH1	***0.04***	***0.98***	***4.09***

Twenty transcripts have been identified as candidate predictors of progression after HT (Table 4.3 and Table 4.4). For the majority of probes a clear difference in time to progression is seen (Figure 4.1).

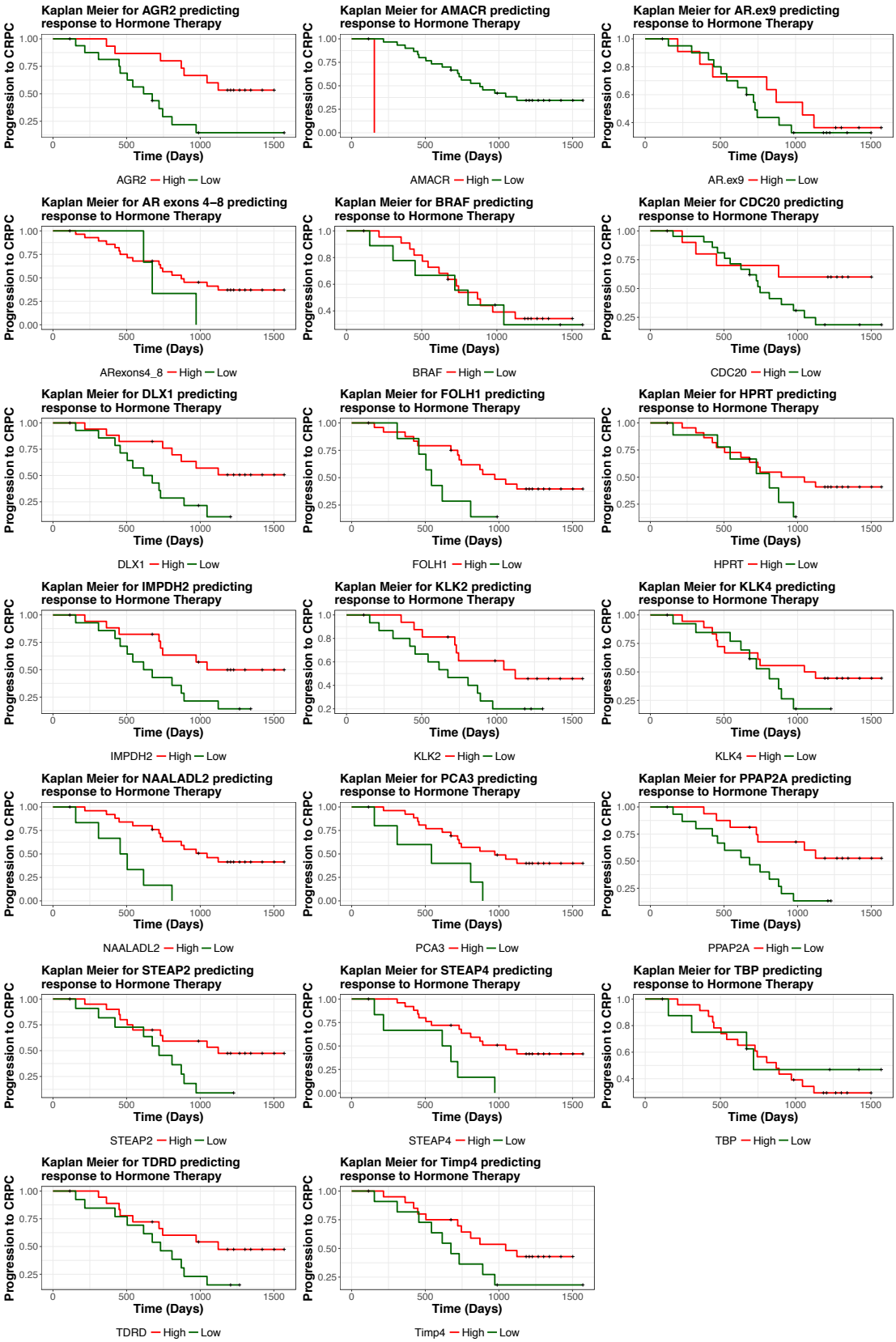


Figure 4.1 Kaplan Meier plots for each of the candidate probes (section 4.3.1). Expression for each probe is grouped into high and low expression using K-means clustering.

4.3.3 Determining the optimal predictor of progression after HT

The optimal model to detect time to progression after HT is likely to be formed from a combination of the expression from multiple probes. There are various methods for identifying the best combination of probes (variable selection) and here I will investigate three i.e. LASSO (section 2.6.2), stepwise regression (section 2.6.4) and random forest (section 2.6.3). Different starting sets of probes will be used based on results from the previous section.

4.3.3.1 Model built using differentially expressed transcripts based on initial response, 12 month relapse and 24 month relapse

Gene selection and three proposed optimal models were produced based on the nine transcripts identified as differentially expressed at initial response, 12 month relapse or 24 month relapse (Table 4.2): a Cox general linear model with shrinkage and variable selection using LASSO (section 2.6.2) (Table 4.5), Stepwise regression on a Cox model (Table 4.6), and a Random Forest model (section 2.6.3) (Table 4.7). The five transcripts selected by LASSO and step are identical, showing reliability in these results. Four out of these five transcripts most important in the random forest model are also similar (*DLX1*, *AR* exons 4-8, *AMACR* and *AGR2* have high importance), though *STEAP4* appears to have increased importance and *BRAF* has lost importance in the Random Forest model.

Table 4.5 The probes included the in the glm after LASSO shrinkage and variable selection, (of the Mann-Whitney U selected probes) with the corresponding beta coefficients

Transcript	*Beta Coefficient*
BRAF	***0.27***
DLX1	***-0.14***
AGR2	***-0.19***
AR exons 4-8	***-0.26***
AMACR	***-0.38***

Table 4.6 The probes included in the cox model after step variable selection (of the Mann-Whitney U selected probes) with the hazard values and *p*-values. The overall performance of the model to predict progression on HT is $p = 0.00024$.

Transcript	*HR*	*p-value*
DLX1	***0.83***	***0.008***
AMACR	***0.56***	***0.013***
AGR2	***0.76***	***0.038***
AR exons 4-8	***0.72***	***0.047***
BRAF	***1.67***	***0.054***

Table 4.7 The importance of each probe in the random forest predictor for HT relapse (of the Mann-Whitney U selected probes).

Transcript	*Importance*	*Relative Importance*
DLX1	***0.042***	***1***
AR exons 4-8	***0.015***	***0.35***
STEAP4	***0.013***	***0.32***
AMACR	***0.011***	***0.27***
AGR2	***0.009***	***0.21***
HPRT	***0.003***	***0.08***
BRAF	***-0.001***	***-0.01***
TBP	***-0.003***	***-0.08***
AR exon 9	***-0.006***	***-0.13***

Using the selected gene sets determined above, a single score was derived for each gene set and a Cox regression model was constructed (Table 4.8). The top four important transcripts of the random forest model performed best (HR=0.0573; $p = 3.29\text{x}10^{-05}$) but all of the models were highly significant ($p < 1.0\text{x}10^{-03}$) in discriminating samples from patient's that progressed on HT and those that did not.

Table 4.8 Overall performance of the models (created from the probes originally identified by Mann-Whitney U) tested by cox.

Model	*p-value*	*HR (95% confidence intervals)*
LASSO genes – (DLX1, AGR2, BRAF, AR exon 9 /AMACR)	***7.5x10^{-04}***	***0.106 (0.01761 - 0.6394)***
Step genes – (DLX1, AGR2, BRAF, AR exons 4-8 /AMACR)	***5.31x10^{-05}***	***0.0786 (0.01299 - 0.4752)***
LASSO and Step genes (DLX1, AGR2, BRAF, AR exons 4-8, AR exon 9 /AMACR)	***6.79x10^{-04}***	***0.0983 (0.01738 - 0.5567)***
Random Forest top 5 genes (DLX1, AGR2, AR exon 4-8, STEAP4 /AMACR)	***3.29x10^{-05}***	***0.0573 (0.008039 - 0.4079)***

4.3.3.2 Model built using Cox selected transcripts

Using the twelve transcripts identified using the Cox regression model on individual probes (Table 4.3), variable selection was performed. LASSO identified seven transcripts (Table 4.9), stepwise regression identified six transcripts (Table 4.10), and Random Forest identified the relative importance (Table 4.11). The transcripts selected by LASSO and stepwise regression have three common transcripts (*KLK2*, *CDC20* and *STEAP2*) but the importance of these probes was not high in the random forest model.

Table 4.9 The probes included the in the glm after LASSO shrinkage and variable selection, (of the cox selected probes) with the corresponding beta coefficients

Transcript	*Beta Coefficient*
KLK2	***-0.009***
CDC20	***-0.012***
PPAP2A	***-0.025***
STEAP4	***-0.032***
DLX1	***-0.059***
NAALADL2	***-0.072***
STEAP2	***-0.076***

Table 4.10 The probes included in the cox model after step variable selection (of the cox selected probes) with the hazard values and *p*-values. The overall performance of the model to predict progression on HT is $p = 0.00323$.

Transcript	*HR*	*p-value*
AMACR	***0.493***	***0.003***
KLK2	***0.496***	***0.008***
STEAP2	***0.446***	***0.057***
PCA3	***1.395***	***0.065***
KLK4	***1.863***	***0.094***
CDC20	***0.799***	***0.097***

Table 4.11 The importance of each probe in the random forest predictor for HT relapse (of the cox selected probes).

Transcript	*Importance*	*Relative Importance*
NAALADL2	***0.0215***	***1***
AMACR	***0.0199***	***0.928***
DLX1	***0.0178***	***0.829***
STEAP4	***0.0067***	***0.313***
PPAP2A	***0.0051***	***0.239***
STEAP2	***0.001***	***0.046***
TDRD	***0.001***	***0.046***
Timp4	***0.0008***	***0.039***
PCA3	***0.0002***	***0.008***
CDC20	***-0.0013***	***-0062***
KLK2	***-0.0033***	***-0.154***
KLK4	***-0.0049***	***-0.226***

The combined score Cox regressions (Table 4.12) showed that the top four important transcripts selected by the random forest model performed best (HR = 0.103; $p = 7.97 \times 10^{-05}$) but all were statistically significant in predicting patient's that progressed on HT.

Table 4.12 Overall performance of the models (created from the probes originally identified by cox) tested by cox.

Model	*p-value*	*HR (95% confidence intervals)*
LASSO genes – (KLK2, DLX1, NAALADL2, PPAP2A, STEAP2, STEAP4, CDC20)	***7.8x10^{-04}***	***0.00038 (1.168x10^{-06} - 0.1239)***
Step genes – (KLK2, STEAP2, PCA3, STEAP4, CDC20 /AMACR)	***0.01***	***0.048 (0.004 - 0.641)***
LASSO and Step genes (DLX1, NAALADL2, STEAP4, KLK2, STEAP2, PPAP2A, CDC20, PCA3, KLK4 /AMACR)	***1.22x10^{-03}***	***0.001 (5.4x10^{-06} - 0.192)***
Common to LASSO and Step genes (KLK2, CDC20, STEAP2)	***8.22x10^{-03}***	***0.026 (0.001 - 0.536)***
Random Forest top 4 genes (NAALADL2, DLX1, STEAP4 /AMACR)	***7.97x10^{-05}***	***0.103 (0.021 - 0.504)***

4.3.3.3 Model built from Log-rank selected transcripts

Variable selection was performed using the ten significant probes identified by the Log rank test (expression divided into two groups using *k*-means) (Table 4.4). LASSO identified seven transcripts (Table 4.13), stepwise regression identified five transcripts (

Table 4.14), and Random forest identified the relative importance of each of the ten transcripts (Table 4.15). The transcripts selected by LASSO and step have four in common (*KLK2*, *AGR2, DLX1* and *STEAP4*). These four transcripts are also the most important in the random forest model.

Table 4.13 The probes included the in the glm after LASSO shrinkage and variable selection, (of the log-rank selected probes) with the corresponding beta coefficients

Transcript	*Beta Coefficient*
KLK2	***-0.003***
AGR2	***-0.02***
PPAP2A	***-0.03***
STEAP4	***-0.03***
DLX1	***-0.06***
NAALADL2	***-0.07***
STEAP2	***-0.08***

Table 4.14 The probes included in the cox model after step variable selection (of the log-rank selected probes) with the hazard values and *p*-values. The overall performance of the model to predict progression on HT is $p = 0.0012$.

Transcript	*HR*	*p-value*
DLX1	***0.78***	***0.001***
NAALADL2	***0.66***	***0.01***
FOLH1	***1.44***	***0.03***
AGR2	***0.75***	***0.04***
STEAP4	***0.69***	***0.08***

Table 4.15 The importance of each probe in the random forest predictor for HT relapse (of the log-rank selected probes).

Transcript	*Importance*	*Relative Importance*
NAALADL2	***0.0258***	***1***
DLX1	***0.0181***	***0.701***
AGR2	***0.0151***	***0.583***
STEAP4	***0.0138***	***0.536***
PPAP2A	***0.0088***	***0.34***
KLK2	***0.0064***	***0.246***
STEAP2	***0.0031***	***0.119***
FOLH1	***-0.0039***	***-0.153***
PCA3	***0.0052***	***-0.202***
IMPDH2	***-0.096***	***-0.373***

In the combined score Cox regressions (Table 4.12) showed that the LASSO selected transcripts performed marginally better (HR = 0.01; $p = 4.7\text{x}10^{-04}$) but all were statistically significant in predicting patient's that progressed on HT.

Table 4.16 Overall performance of the models (created from the probes originally identified by log rank) tested by cox.

Model	*p-value*	*HR (95% confidence intervals)*
LASSO – (KLK, DLX1, NAALADL2, PPAP2A, STEAP2, STEAP4, AGR2)	***4.7x10-04***	***0.01 (0.0004 - 0.2654)***
Step – (DLX1, NAALADL2, STEAP4, AGR2, FOLH1)	***7x10-04***	***0.0212 (0.0013 - 0.3362)***
LASSO and Step (KLK2, DLX1, NAALADL2, PPAP2A, STEAP2, STEAP4, AGR2, FOLH1)	***6x10-04***	***0.0072 (0.0002- 0.2865)***
Common to LASSO, Step and Random Forest (KLK2, AGR2, DLX1, STEAP4)	***7x10-04***	***0.0275 (0.0023 - 0.326)***

4.3.3.4 Model built using combining probe selection lists to produce final model

Combining the candidate probe lists identified using the different gene selection model may produce a better predictor of HT progression. Using each combination of candidate probe lists (Table 4.2, Table 4.3, & Table 4.4), LASSO was applied for variable selection (as it is clear and is designed to avoid over-fitting) and a linear combination score with Cox regression model was produced (Table 4.17). Initiating the variable selection with a combination of the candidate probes identified as differentially expressed at initial response, 12 month relapse or 24 month relapse (Mann Whitney U) and by the log rank test produced the best model (p =1.3x10^{-04}; HR = 0.0091). This model includes *AGR2*, *AR* exons 4-8, *DLX1*, *KLK2*, *NAALADL2*, *PPAP2A* and *AMACR*. It has an AUC of 0.783 (Figure 4.1). This is the best performing model constructed using the Nanostring 1 data. The Kaplan Meier plot for the seven-transcripts combined was also produced (Figure 4.2).

Table 4.17 Comparing the Cox regression models of various linear combination scores producing from combining gene selection lists. Mann-Whitney U = candidate probes identified as differentially expressed at initial response, 12 month relapse or 24 month relapse; cox = candidate probes identified by step applied to cox regression models; Log rank = candidate probes identified by the log rank test on expression dichotomised into low and high expression.

Combination	*Method for variable selection*	*Resulting probes in model*	*(cox) p-value*	*(cox) HR*
Mann-Whitney U and cox	***LASSO***	***AGR2, AR exons 4-8, DLX1, KLK2, NAALADL2, TDRD/AMACR***	***$1.3x10^{-04}$***	***0.0091 (0.0004 - 0.214)***
Mann-Whitney U and Log rank	***LASSO***	***AGR2, AR exons 4-8, DLX1, KLK2, NAALADL2, PPAP2A/AMACR***	***$2.3x10^{-05}$***	***0.04288 (0.005 - 0.345)***
Cox and Log rank	***LASSO***	***DLX1***	***0.01***	***0.871 (0.781 – 0.972)***

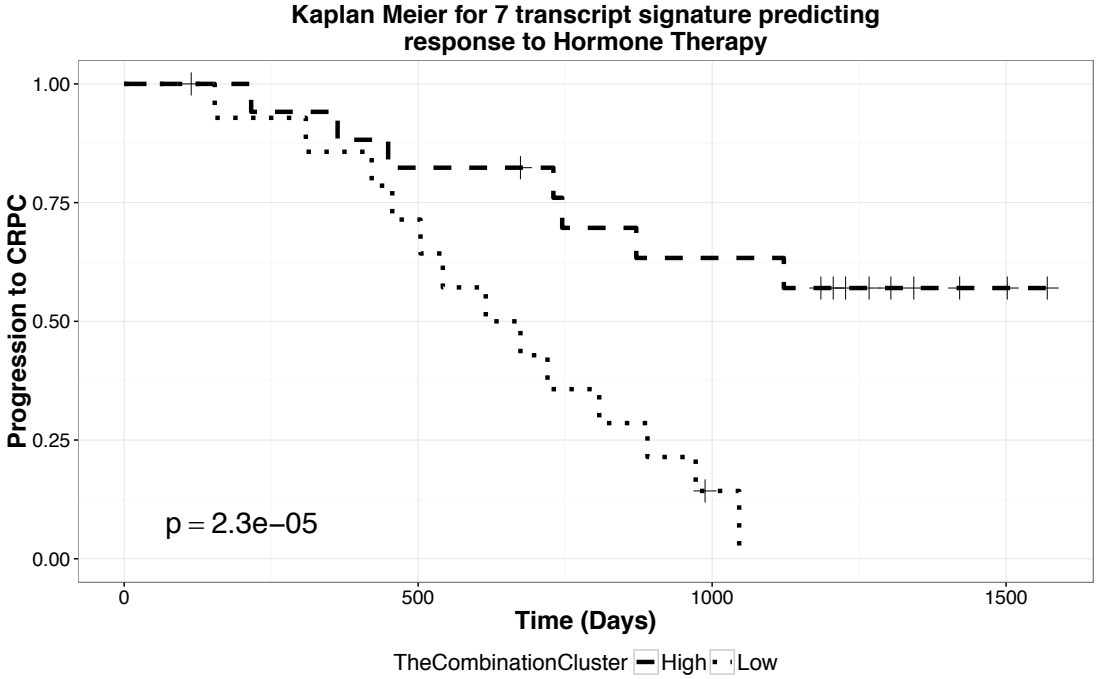


Figure 4.2 Kaplan Meier showing the seven-transcript signature (*AR* exons 4-8 * *AGR2* * *DLX1* * *KLK2* * *NAALADL2* * *PPAP2A* / *AMACR*) separated into low and high expression using *k*-means. The significance was measured using the cox model (Table 4.17), $p = 2.3x10^{-05}$.

The seven-transcript signature was a better predictor of HT relapse than other clinical variables (including bone scan outcome, Gleason score, initial PSA value and age) when treated individually (Table 4.18). LPD group (identified from chapter 3, section 3.5.5) was also tested. The seven-transcript score was a statistically significant independent predictor of HT progression when combined with covariate clinical variables (p = 0.003; HR = 0.03; Table 4.19).

Table 4.18 Univariate cox models showing the significance of clinical variables, LPD group and the seven-transcript signature on predicting HT relapse.

UNIVARIATE MODELS		
Model	***p-value***	***HR (95% CI)***
Age	***0.84***	***1.006 (0.949 - 1.067)***
PSA	***0.05***	***1.001 (1 - 1.002)***
Gleason Scores	***0.27***	
Gleason 7		
Gleason 8	***0.26***	***0.252 (0.022 - 2.844)***
Gleason 9	***0.91***	***1.101 (0.22 - 5.54)***
Gleason Category		
Gleason 7+8		
Gleason 9+NA	***0.15***	***2.291(0.666-7.883)***
Bone Scan	***0.19***	
Negative		
Positive	***0.2***	***1.854 (0.719 - 4.785)***
LPD group	***0.09***	
LPD1		
LPD2	***0.78***	***1.413 (0.128 - 15.59)***
LPD3	***0.24***	***3.716 (0.414 - 33.36)***
LPD4	***0.07***	***7.043 (0.844 - 58.77)***
LPD NA	***0.12***	***5.589 (0.637 - 49.04)***
*DLX1 * AGR2 * KLK2 * NAALADL2 * AR exons 4-8 * PPAP2A / AMACR*	$2.3x10^{-05}$	***0.04288 (0.005 - 0.345)***

Table 4.19 Multivariate cox model for predicting early relapse on HT.

MULTIVARIATE MODEL with the seven transcript signature, p-value = $7.7x10^{-04}$		
Variable	***p-value***	***HR (95% CI)***
*DLX1 * AGR2 * KLK2 * NAALADL2 * AR exons 4-8 * PPAP2A / AMACR*	***0.003***	***0.03 (0.003 – 0.313)***
Age	***0.996***	***1 (0.949 - 1.054)***
PSA	***0.176***	***1 (0.997 – 1.001)***
Gleason Category *Gleason 7 + 8* *Gleason 9 + NA*	***0.167***	***2.61 (0.67 – 10.143)***
Bone Scan *Negative* *Positive*	***0.276***	***1.85 (0.612 – 5.59)***

4.3.4 Validation of the seven-transcript signature using NanoString 2 data

The second set of NanoString data had 43 patients on HT (chapter 5), of which 27 samples were unique to NanoString 2 (Table 4.20).

Table 4.20 Clinical breakdown of the 27 HT patients unique to NanoString 2.

Clinical Variable	*Number of patients*
Gleason Score	
10	***1***
9	***13***
8	***3***
7 (4+3)	***0***
No Biopsy: Advanced	***10***
Bone Scan	
Positive	***10***
Negative	***14***
Unknown	***3***
PSA	***Median 63 (7.6 - 9604)***
Age	***Median 77 (61 - 93)***

The seven-transcript signature was not detected as a significant predictor of progression in Nanostring 2 (Cox-regression model; $p = 0.612$, $HR = 0.640$ (95% CI: 0.1118 - 3.669). This is confirmed by looking at Kaplan Meier plots of the combined signature (Figure 4.3) and the individual probes (Figure 4.4).

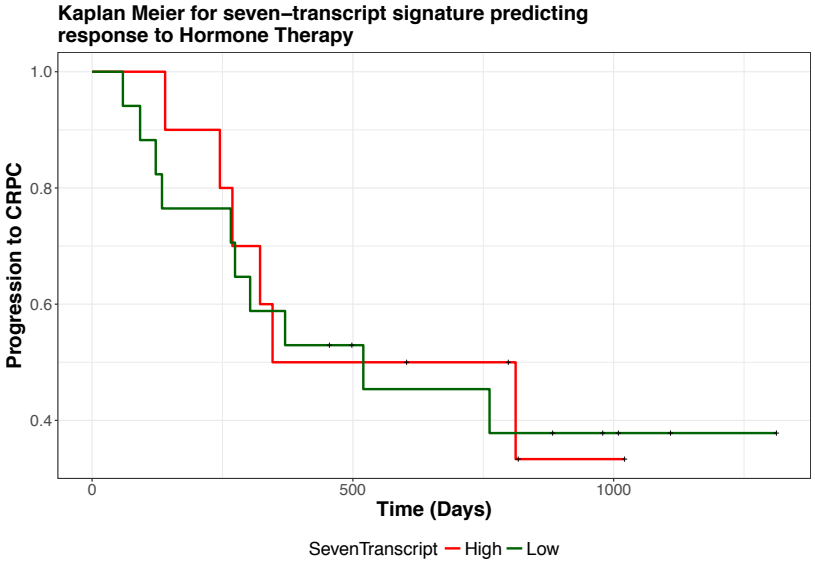


Figure 4.3 Kaplan Meier plot showing the seven transcript signature on NanoString 2 data. The signature was separated using *k*-means.

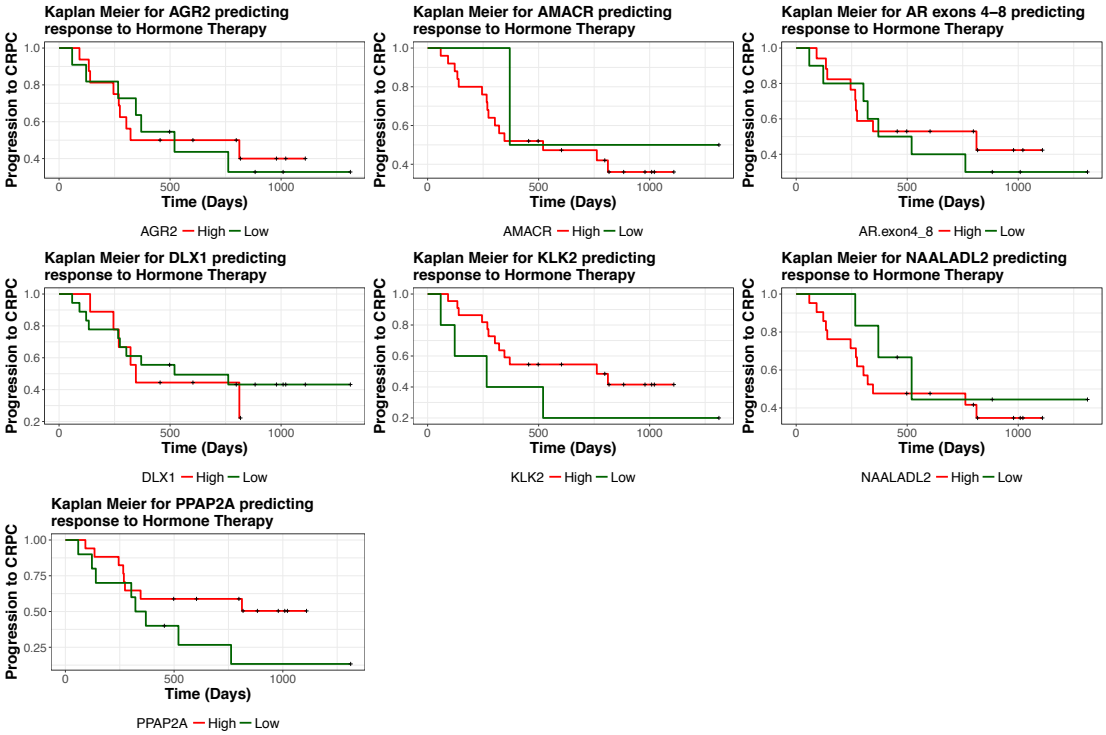


Figure 4.4 Individual Kaplan Meier plots for the seven transcripts involved in the signature

To remove any potenital batch effect, ComBat was used to normalise the second NanoString data to the pilot study data. Similar results were obtained ($p = 0.62$, HR = 0.68 (95% CI: 0.15 – 3.18)). Overall Nanostring 1 and 2 are similar.

4.4 Identifying novel progression related transcripts in the NanoString 2 data

NanoString 2 data contained expression levels from 110 more probes than NanoString 1 data. Therefore, I identified novel progression related transcripts in the NanoString 2 data. Expression of eleven transcripts were identified as significant predictors of progression using Cox regression models ($p < 0.05$), but none were significant after multiple testing correction (Table 4.21). Grouping expression into high and low using using *k*-means (section 2.5.3), found *MSMB* to be significant even after multiple testing correction ($p = 0.22\text{x}10^{-09}$, adjusted $p = 1.54\text{x}10^{-06}$). Ten other transcripts were significant using this method prior to multiple testing correction ($p < 0.05$; Table 4.22). Log rank test was also performed using the median as a separation cut off for high and low-expression, ten transcripts were observed to be significant prior to multiple testing correction ($p < 0.05$; Table 4.23).

Table 4.21 Cox regression modelling identified ten probes that were predictors of progression after HT. None were significant after multiple testing correction.

Probe	*p-value*	*Adjusted p-value*	*Coefficient*
MSMB	***0.0037***	***0.59***	***0.7***
MIR4435_1HG	***0.009***	***0.98***	***1.28***
BTG	***0.017***	***0.98***	***1.34***
PCSK6	***0.021***	***0.98***	***0.48***
MCTP1	***0.028***	***0.98***	***1.18***
IGFBP3	***0.032***	***0.98***	***1.2***
PCA3	***0.036***	***0.98***	***0.82***
SEC61A1	***0.039***	***0.98***	***1.19***
CLIC2	***0.04***	***0.98***	***1.65***
STOM	***0.048***	***0.98***	***1.15***

Table 4.22 Log-rank test identified probes that could significantly predict progression on HT. *K*-means was used to separate into high- and low-expression of each probe.

Probe	*p-value*	*Adjusted p-value*	X^2
MSMB	$9.22x10^{-09}$	$1.54x10^{-06}$	33
BTG	0.001	0.2	10.4
CLIC2	0.003	0.5	8.6
MKi67	0.006	0.9	7.7
IGFBP3	0.02	0.99	5.8
PCSK6	0.02	0.99	5.7
APOC1	0.02	0.99	5.5
COL10A1	0.02	0.99	5.4
KLK4	0.02	0.99	5.2
MIC1	0.03	0.99	4.6
SSPO	0.04	0.99	4

Table 4.23 Log-rank test identified probes that could significantly predict progression on HT. Median was used to separate into high- and low-expression of each probe.

Probe	*p-value*	*Adjusted p-value*	X^2
CLIC2	0.005	0.87	7.8
PCA3	0.008	0.99	7.1
PPAP2A	0.008	0.99	7
SEC61A1	0.012	0.99	6.3
IGFBP3	0.015	0.99	5.9
HIST1H2BG	0.016	0.99	5.8
TBP	0.02	0.99	5.4
PCSK6	0.022	0.99	5.3
BTG2	0.031	0.99	4.6
STOM	0.033	0.99	4.6

There were common transcripts identified in all three different methods: *BTG2*, *CLIC2*, *IGFBP3*, and *PCSK6*. Variable selection using LASSO and stepwise regression identified an optimal model of *BTG2*, *CLIC2* and *PCSK6* (Table 4.24). These three transcripts were also identified as having the greatest importance in a Random Forest model (Table 4.24).

Table 4.24 Optimising models using the four probes common to log-rank and cox tests. The cox model had an overall *p*-value: $p = 0.0013$.

	Lasso Beta value	*Cox p-value*	*Cox HR*	*Random Forest – Relative importance*
BTG2	***0.1***	***0.2***	***1.17***	***1***
CLIC2	***0.43***	***0.02***	***1.88***	***0.49***
PCSK6	***-0.7***	***0.003***	***0.35***	***0.21***
IGFBP3	**-**	**-**	**-**	***-0.24***

4.5 Hormone Therapy Predictor using *KLK2* ratio data on Nanostring 1

For NanoString 2 data I found that refactoring the data using *KLK2* ratio improved the ability to distinguish clinical subtypes (section 5.7.5). Therefore, here I will develop an optimal predictor of progression after HT in the Nanostring 1 data after refactoring using *KLK2*. Differential expression was assessed using the Mann-Whitney U test at three time points: initial non-responders, relapse within 6 months, within 12 months and within 24 months (Table 4.25). Cox regression models (section 2.8.2) identified nine transcripts who's expression were significantly predictors of progression (Table 4.26; $p < 0.05$; multiple testing correction not applied). Log-rank test on expression levels classified as high or low (threshold determined using *k*-means, found *AURKA* to be a significant predictor of progression prior to multiple testing correction only ($p = 0.034$, Benjamin-Hochberg adjusted $p = 0.99$). Log-rank test when using median for separation into high and low expression, found four transcripts to be differentially expressed between those that relapsed and those that continue to respond to HT (Table 4.27; $p < 0.05$; no multiple testing correction applied; Figure 4.7).

Table 4.25 Mann-Whitney U test identifies probes differentially expressed between those that have relapsed and those that are still responding to HT at different time periods (initial response relapse, within 6 month relapse, with 12 month relapse and within 24 month relapse.

Initial Response and After 6 months:

Probe	*p-value*	*Adjusted p-value*	*Log2(fold change)*
KLK3 exons 2-3	***0.016***	***0.92***	***-0.05***
PSGR	***0.029***	***1***	***-0.09***
B2M	***0.034***	***1***	***-0.06***
AURKA	***0.047***	***1***	***-0.11***

After 12 months:

Probe	*p-value*	*Adjusted p-value*	*Log2(fold change)*
PSGR	***0.008***	***0.47***	***0.08***
FOLH1	***0.022***	***0.98***	***0.04***
KLK3 exons 2-3	***0.028***	***0.98***	***0.04***
B2M	***0.038***	***0.98***	***0.06***

After 24 months:

Probe	*p-value*	*Adjusted p-value*	*Log2(fold change)*
PECI	***0.01***	***0.56***	***0.03***
PSGR	***0.016***	***0.88***	***0.06***
DLX1	***0.018***	***0.97***	***-0.05***
ALAS1	***0.045***	***1***	***0.05***

Table 4.26 Cox identified probes that are differentially expressed in NanoString 1 data normalised by KLK2 ratio.

Probe	*p-value*	*Adjusted p-value*	*Coefficient*
CAMKK2	***0.015***	***0.86***	***2.22***
UPK2	***0.027***	***0.98***	***1.49***
KLK3 exons 2-3	***0.031***	***0.98***	***3.15***
PECI	***0.031***	***0.98***	***2.27***
HPN	***0.031***	***0.98***	***2.48***
KLK4	***0.034***	***0.98***	***3.6***
GAPDH	***0.036***	***0.98***	***2.4***
ALAS1	***0.038***	***0.98***	***2.15***
KLK3 exons 1-2	***0.048***	***0.98***	***2.39***

Table 4.27 Log rank (using median for separating high and low expression) identified probes that differ between response to HT.

Probe	*p-value*	*Adjusted p-value*	X^2
STEAP4	***0.007***	***0.32***	***7.7***
PECI	***0.009***	***0.49***	***6.8***
SERPINB5	***0.013***	***0.72***	***6.1***
TBP	***0.037***	***0.99***	***4.4***

PECI was identified in all comparisons; Mann-Whitney U, Cox and Log-rank (using median for separation). Variable selection on all fourteen transcripts that were identified as candidate predictors (Table 4.25, Table 4.26, Table 4.27) were performed. Lasso identified three transcripts: *CAMKK2*, *DLX1* and *UPK2* (Table 4.28); stepwise regression identified six transcripts of which only *UPK2* was common to Lasso (Table 4.29); and using Random Forest three of the top five important genes were found in either the Lasso or Stepwise regression results (Table 4.30).

Table 4.28 Lasso selects three transcripts for HT progression prediction in *KLK2* adjusted data.

Transcript	*Beta coefficient*
CAMKK2	***0.232***
DLX1	***-0.028***
UPK2	***0.099***
Cox model: p – value = 0.2, HR = 0.999, 95% CI = 0.998 - 1	

Table 4.29 Stepwise regression selects six probes for early HT relapse prediction in *KLK2* adjusted data.

Transcript	*p-value*	*HR*
B2M	***0.096***	***2.713***
FOLH1	***0.042***	***0.246***
GAPDH	***0.122***	***0.102***
HPN	***0.083***	***5.504***
PSGR	***0.027***	***2.773***
UPK2	***0.039***	***1.955***
Cox model: p – value = 0.039, HR = 86.54, 95% CI = 2.435 - 3076		

Table 4.30 Random forest shows the importance of each transcript in distinguishing early HT relapse in *KLK2* adjusted data.

Transcript	*Importance*	*Relative Importance*
PSGR	***0.035***	***1.00***
PECI	***0.035***	***0.99***
HPN	***0.007***	***0.20***
KLK3 exons 1-2	***0.006***	***0.18***
CAMKK2	***0.006***	***0.16***
ALAS1	***0.006***	***0.16***
FOLH1	***0.005***	***0.13***
DLX1	***0.003***	***0.09***
AURKA	***0.0002***	***0.01***
KLK3 exons 2-3	***-0.0002***	***-0.01***
UPK2	***-0.0007***	***-0.02***
GAPDH	***-0.001***	***-0.03***
B2M	***-0.006***	***-0.02***
KLK4	***-0.007***	***-0.19***

Kaplan Meier plots (section 2.8.1) were produced using a *k*-means determined threshold between high and low expression (Figure 4.5). Applying variable selection on this dichotomised data, Lasso identifies *CAMKK2*, *PSGR* and *UPK2* (Table 4.31); stepwise regression selects *AURKA*, *CAMKK2*, *KLK3* exons 1-2 and *UPK2* (Table 4.32); whilst random forest suggests that *PSGR* is of most importance, followed by *UPK2* and *CAMKK2* (the same three transcripts selected via Lasso) (Table 4.33). *CAMKK2*, *PSGR* and *UPK2*, which were selected by Lasso and also the three most important transcripts according to Random forest, produce a significant cox model (p = 0.007, HR = 1.0028, 95% CI = 1.001 - 1.005).

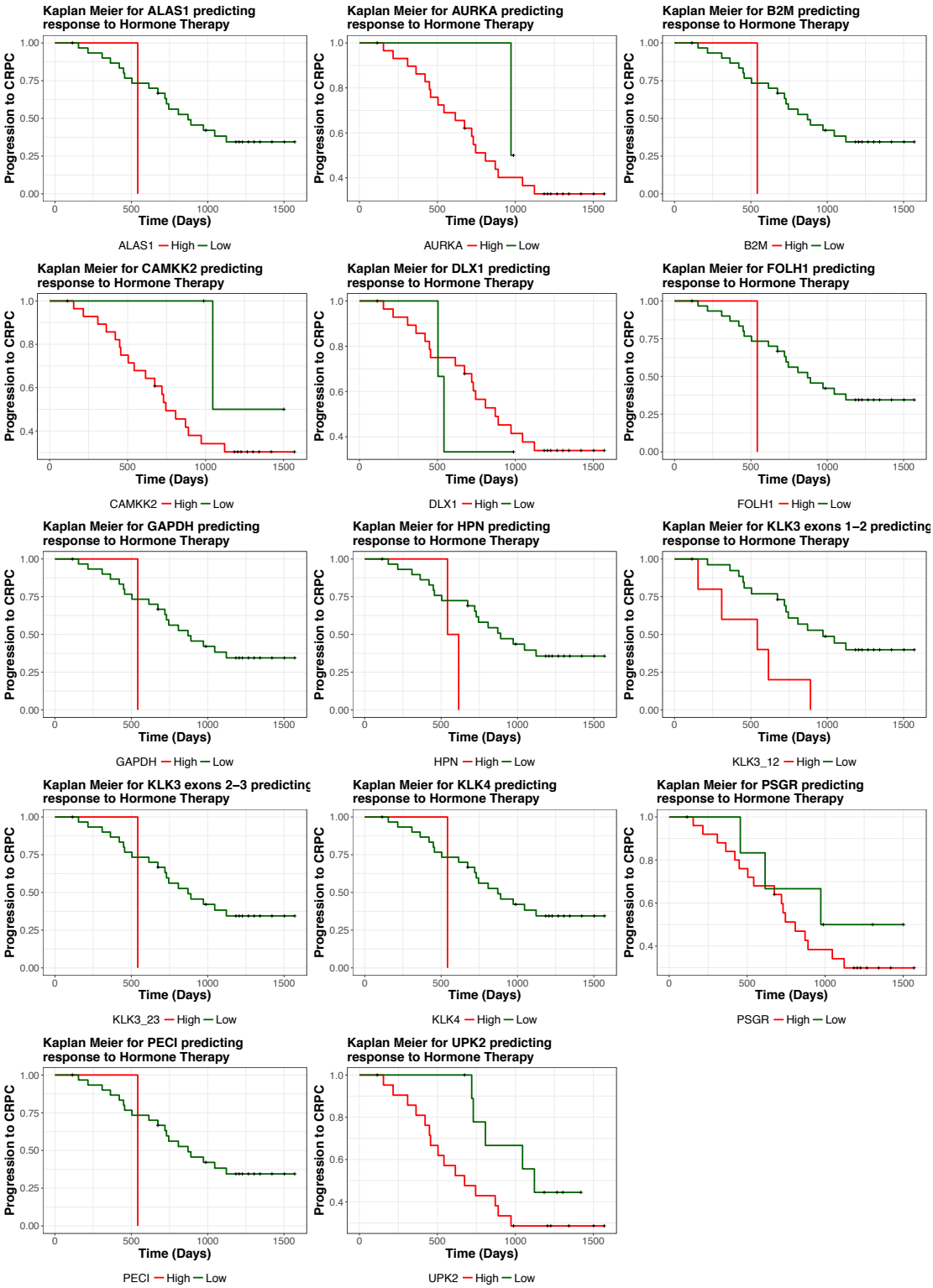


Figure 4.5 Kaplan Meier plots (expression separated via *k*-means) for the fourteen transcripts identified via Mann Whitney U, Cox and log-rank tests for early HT relapse.

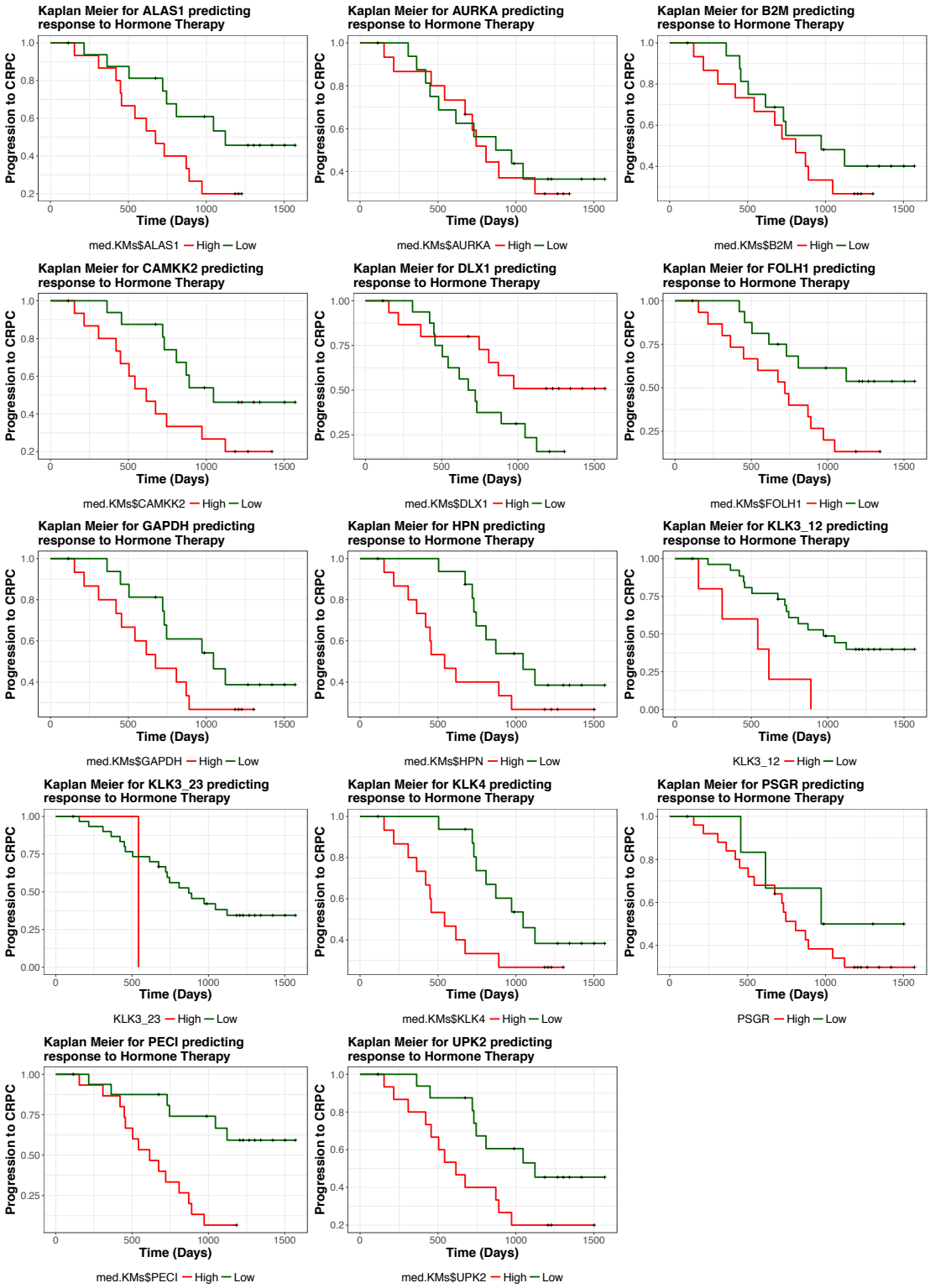


Figure 4.6 Kaplan Meier plots (expression separated via median) for the fourteen transcripts identified via Mann Whitney U, Cox and log-rank tests for early HT relapse.

Table 4.31 Lasso (with glm) selects three transcripts from the five shown to be differential from Kaplan Meier plots using *k*-means for separation. An overall Cox model using these three probes proves to be significant ($p = 0.007$).

Transcript	*Beta coefficient*
CAMKK2	***0.31***
PSGR	***0.06***
UPK2	***0.18***
Cox model: p – value = 0.007, HR = 1.0028, 95% CI = 1.001 - 1.005	

Table 4.32 Step (with Cox) selects four transcripts from the five shown to be differential from Kaplan Meier plots using k-means for separation. An overall Cox model using these four probes is not significant (p = 0.07).

Transcript	*p-value*	*HR*
AURKA	***0.13***	***1.25***
CAMKK2	***0.04***	***3.51***
KLK3 exons 1-2	***0.14***	***0.13***
UPK2	***0.06***	***2.10***
Cox model: p – value = 0.07, HR = 1, 95% CI = 1-1		

Table 4.33 Random forest shows the importance of each of the five transcripts identified via Kaplan Meier plots using *k*-means for separation. The top three important transcripts are identical to the Lasso output.

Transcript	*Importance*	*Relative Importance*
PSGR	***0.098***	***1***
UPK2	***0.018***	***0.179***
CAMKK2	***0.014***	***0.137***
AURKA	***0.003***	***0.031***
KLK3 exons 1-2	***-0.002***	***-0.017***

Lasso has consistently selected *CAMKK2* and *UPK2* along with one other transcript (*DLX1* when all that showed significance were used, *PSGR* when only those that appeared to be significant in *k*-means separated Kaplan Meier plots). This consistency of selecting *CAMKK2* and *UPK2* when the input variables are altered shows reproducibility. Though the model including *DLX1* was not significant, the model including *PSGR* was the most significant model identified ($p = 0.0023$, $HR = 1.0028$, 95% CI = 1.001 - 1.005). *CAMKK2* was always identified as important by Random forest. Step (with Cox) and Random Forest was not very consistent in creating models

with similar variables. Therefore, the most consistent and significant cox regression model identified contained *CAMKK2*, *PSGR* and *UPK2* (p = 0.007, *HR* = 1.0028, 95% CI = 1.001 - 1.005).

4.5.1 Validation of the final model on KLK2 ratio NanoString 2 data

The second set of NanoString data also refactorised using the *KLK2* ratio method was used to test the *CAMKK2*, *PSGR* and *UPK2* Cox regression mode identified in NanoString 1 data. The model did not reach statistical significance as a predictor of progression (p = 0.4, *HR* = 1.000774 (95% *CI*: 0.999 – 1.003). Looking at the Kaplan Meier plots of the transcripts individually (Figure 4.7), *CAMKK2*, *PSGR* and *UPK2* showed better survival with low expression in the pilot study, yet in the second set of data, both *CAMKK2* and *PSGR* now show better survival with higher expression.

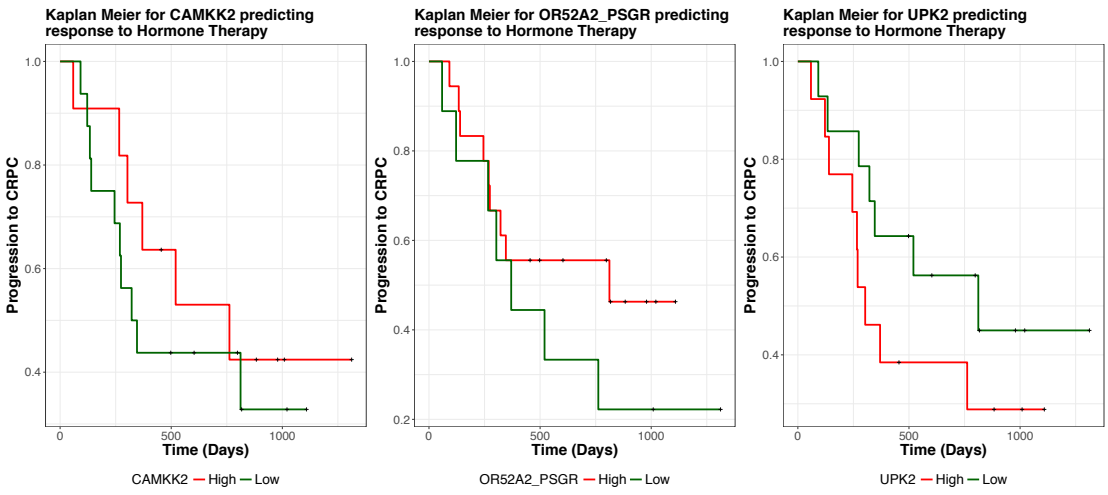


Figure 4.7 Kaplan Meier plots for the three transcripts in the model for KLK2 adjusted hormone therapy data: *CAMKK2*, *PSGR* and *UPK2*.

4.6 Conclusion

Stratified medicine enables the optimal treatment for cancer patients to be selected and improve overall survival. There have been successes in breast[208] and lung cancer[209,210], for example. However, no such robust testing and stratification exists for men with PCa and the route of treatment is not always clear. In particular, it is hard to predict the

response to treatments such as radiotherapy, hormone therapy and prostatectomy and determine whether active surveillance is a better option than treatment. These issues are key research areas for the clinical management of PCa patients.

In this chapter, I investigated if expression profiles of urinary microvesicles could be used to estimate how long a patient responded to HT. I successfully built a number of different models based on the normalisation method and dataset used. To produce a non-invasive test for the identification of those who will relapse early on HT, and thus could benefit from additional treatment, would be ground breaking for PCa patients. Depending on how the NanoString data was normalised, we saw correlation between two signatures and HT relapse.

Under the Nanostring 1 dataset with standard normalisation (via NanoString's positive probes, section 2.3.1), the optimal predictor of progression in HT patients included the expression of probes *AGR2*, *DLX1*, *KLK2*, *NAALADL2*, *AR* exons 4-8, *PPAP2A* and *AMACR* (Cox-regression model, $p = 2.3\text{x}10^{-05}$, *HR* = 0.043). This seven-probe signature was also a significant independent predictor of progression improving on other clinical factors, initial PSA, Gleason score, initial bone scan results and age (Table 4.18, Table 4.19).

After *KLK2* adjustment of the NanoString 1 data (section 2.1.1), we selected model including *CAMKK2*, *PSGR* and *UPK2* that could significantly separate those that progressed to CPRC and those that continued to respond to HT. This model was significant alone ($p = 0.0023$, *HR* = 1.0028, 95% CI = 1.001 - 1.005), which was again more significant than other clinical factors including initial PSA, Gleason score, initial bone scan results and age (Table 4.18).

Unfortunately, both of these models were not validated in the NanoString 2 dataset. There are many possible reasons why the models were not validated but one factor is that there are a relatively small number of patients in this cohort and with a relatively short feedback. This means that the models are very sensitive to outliers in the data. There are also differences between the Nanostring datasets: different centres ran the

experiment, there were different probesets, newer samples could have been collected slightly differently, and the cohorts could be somewhat different. It appears the data is very sensitive with no candidate probes being common before and after factorisation. This is common to many expression-based biomarker studies. There is a lack of robustness with proposed tests very rarely being validated in different cohorts[211]. It should also be remembered that this is a targeted based assay, the optimal probes to distinguish treatment outcome may not be included.

In probes that were unique to the second NanoString data set, the optimal model for determining time to progress for HT patients contained *BTG2*, *CLIC2* and *PCSK6*.

In this chapter I have shown the utility of urine derived microvesicle expression profiles for the prediction of outcome after treatment. This is a proof of concept that would require a much larger series with longer feedback to find the best combination of transcripts and become a usable test.

5

NanoString Data Analysis 2

5.1 Summary

The Movember GAP1 Urine Biomarker Consortium had multiple collaborators working on the identification of urinary biomarkers for the risk-stratification of PCa. Our laboratory is specifically interested in the RNA expression changes in PCa that are detectable within urinary cell sediments and extracellular vesicles (EVs) from samples collected at multiple centres. The aims of my study were to see if I could identify robust models of expression profiles using data obtained from NanoString that could answer important clinical questions in PCa management: can I detect PCa from non-PCa samples and can I risk stratify PCa, both without the need for biopsy. I therefore,

investigated different methods for normalisation of this urinary EV derived data with the aim to build optimal models from the expression of 167 markers for risk stratification and detection of cancer.

I identified robust models for the detection of PCa from non-PCa samples (AUC = 0.851) and of high-risk PCa from non-PCa samples (AUC = 0.897). Models to predict risk stratification between samples with no evidence of cancer (CB) and cancer in order of severity (CB->L->I->H) were also produced (AUC = 0.709). My models used many of the already published transcripts used in whole urine assays but also included novel transcripts that may be present in EV fractions.

5.2 Introduction

NanoString expression analysis of 167 gene-probes was applied to cell and extracellular vesicle (EV) fractions of urine from prostate cancer patients to form the NanoString 2 data set. In this chapter, quality control and technical trouble shooting (section 2.3) is applied to the whole data set, before performing exploratory analysis using just the EV samples. Investigation of the cell fraction samples can be found in chapter 6.

5.2.1 The Research Gap

Risk stratification is currently based on PSA, Gleason score and T stage but has the potential to be improved by using a novel biomarker panel. This could help tailor patients to treatment pathways and determine, at diagnosis, the aggressiveness of disease. The PCA3 test is an established biomarker that is capable of predicting PCa on a second biopsy. Therefore, showing the utility of the use of urine in PCa diagnostics and prognostics, and has shown some minor improvements to risk stratification. In chapter 3, I performed a pilot project exploring the use of NanoString applied to genetic material obtained from urinary EVs and showed that it was capable of capturing clinically relevant expression profiles.

5.2.2 Aims

In this chapter I used NanoString technology to investigate the RNA expression level changes of 167 target sequences within EVs extracted from urinary samples collected at multiple centres world-wide as part of the Movember study. The aims of this study are:

1. To identify better processing techniques for the EV NanoString data
2. To determine whether EV expression profiles are robust across variable sample cohorts collected from different centres.
3. To identify optimal models built from the expression of 167 markers for risk stratification and detection of cancer.

5.2.3 The Probe Targets

A panel of experts selected the 167 sequence targets used as probes. The probes were primarily selected from publications that highlighting genes overexpressed in prostate tumour tissue. 28 gene probes were selected from Next Generation Sequencing data of 20 urine EV RNA samples from the NNUH. Additionally, some prostate tissue specific controls and controls for kidney, bladder and blood were also included. See Supplementary Table 1 for further details.

5.2.4 Classification of prostate cancer patient samples

NanoString data from 864 samples was obtained, 95 samples were from the cell fraction. 756 samples remained after quality control (Section 5.3.2). Samples were divided in to a training set and a test set based on a 2:1 ratio while maintaining the proportions of each PCa risk category (Table 5.1) and sample collection centre (Table 5.2). The median age and PSA at diagnosis have been recorded for each clinical category within the training and tests, respectively (Table 5.2.3).

Table 5.1 Classification and Frequency of the sample types based on NICE criteria[40]. The quantity of samples for each clinical group are provided as well as the clinical description of the group in terms of Gleason score, PSA level and T stage.

Classification: NICE Groupings				
Sample Class	***Description***	***Number of Samples***	***Number of Training Samples***	***Number of Test Samples***
Advanced	***Advanced and Hhh (G8-10 PSA>100) and Hh (G8-10 PSA<100)***	***31***	***21***	***10***
High-risk	***HL= G7 PSA>20***	***107***	***72***	***35***
Intermediate-risk	***I= G3+4 PSA<20 and IL= G6 PSA>10***	***214***	***142***	***72***
Low-risk	***L= Low G6 PSA<10***	***156***	***104***	***52***
Abnormal	***High PSA no Bx, Prostatitis, Raised PSA negative Bx, HGPIN***	***137***	***92***	***45***
CB	***CB – no evidence of cancer***	***111***	***73***	***38***
Total		***756***	***504***	***252***

Table 5.2 Sample collection-site breakdown of the EV samples from NanoString2.

Location	*Training Set*	*Test Set*	*Number of Samples*
Dublin	***16***	***8***	***27***
ICR	***84***	***41***	***130***
UEA	***323***	***163***	***496***
USA	***62***	***23***	***103***
Total			***756***

Table 5.3 Median age and PSA of each clinical category within the training and test datasets.

	Training Set		**Test Set**	
	Median age	Median PSA	Median age	Median PSA
Advanced	78	273.5	82	285.75
High-risk	69	22.35	73.5	23.7
Intermediate-risk	69	9.2	67	8.45
Low-risk	64.5	6.1	64	5.5
Abnormal	67	8.19	66	7.7
CB	63	1.4	64.5	1.235

5.3 Data Preprocessing and Technical Variation

5.3.1 Normalisation and Background correction

There were six positive-control non-human ERCC probes included in the NanoString series and these were used to normalise the data for all samples as per the NanoString manual. As for the pilot data set, a large proportion (33%) of data points were less than zero after negative control correction. Therefore negative control correction was not used in this analysis. As shown in NanoString 1 (section 3.3.4) Log_2 transformation (section 2.3.3) was used to obtain a more normal distribution in the data (Figure 5.1). The Log_2 data did not follow a normal distribution using the Shapiro-Wilk test (Table 5.4), this suggests we should use non-parametric methods for analysis.

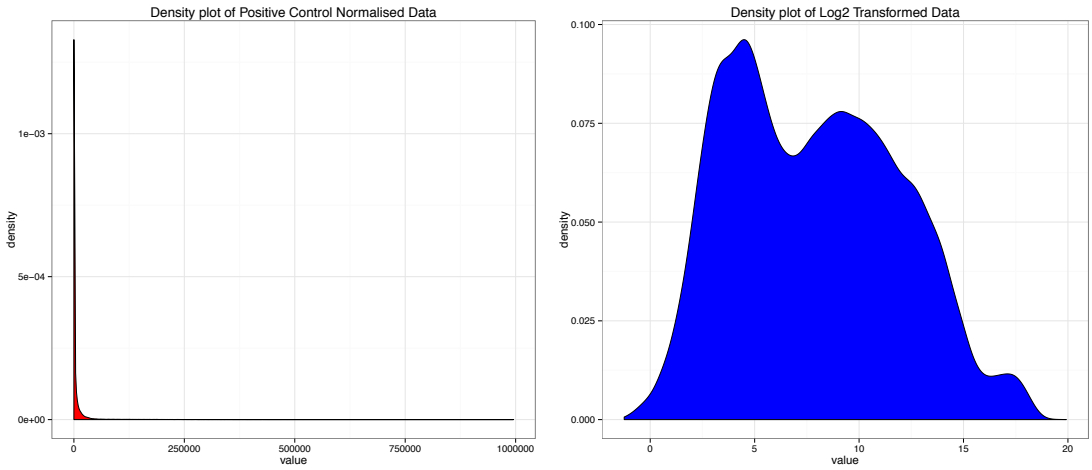


Figure 5.1 A) Positive control normalised data. B) Positive control normalised and Log_2 transformed data. The data shows a more normal distribution after Log_2 transformation.

Table 5.4 Shapiro-Wilk tests show that Log_2 data is not normally distributed.

	Log_2 transformed		
	W	***p-value***	***Normally Distributed***
The first set of randomly selected 29 samples, all probes	***0.96***	***< 2.2x10^{-16}***	***No***
The second set of randomly selected 29 samples, all probes	***0.97***	***< 2.2x10^{-16}***	***No***
The third set of randomly selected 29 samples, all probes	***0.98***	***< 2.2x10^{-16}***	***No***
The fourth set of randomly selected 29 samples, all probes	***0.97***	***< 2.2x10^{-16}***	***No***
The fifth set of randomly selected 29 samples, all probes	***0.98***	***< 2.2x10^{-16}***	***No***
The sixth set of randomly selected 29 samples, all probes	***0.97***	***< 2.2x10^{-16}***	***No***
The first set of randomly selected probes, all samples	***0.99***	***< 2.2x10^{-16}***	***No***
The second set of randomly selected probes, all samples	***0.99***	***2.205x10^{-15}***	***No***
The third set of randomly selected probes, all samples	***0.94***	***< 2.2x10^{-16}***	***No***
The fourth set of randomly selected probes, all samples	***0.96***	***< 2.2x10^{-16}***	***No***
The fifth set of randomly selected probes, all samples	***0.94***	***< 2.2x10^{-16}***	***No***
The sixth set of randomly selected probes, all samples	***0.94***	***< 2.2x10^{-16}***	***No***

5.3.2 Quality of Normalisation

The quality of the data, and its normalisation and transformation, was assessed using NanoStringNorm (section 2.3.2.1) and NanoStringQCPro (section 2.3.2.2). Overall the quality was good but a few samples and a few probes need to be treated with caution. The samples identified by the IQR/median plot were removed (A210, A216, A517, C147_1, M_97_5, M_138_7, M_149_7) along with some CBN samples, which were identified through NGS analysis (not shown as this was not performed by me).

5.3.2.1 NanoStringNorm

The negative controls had both low means and standard deviations and the positive controls showed low standard deviation, as expected. The majority of the probes

clustered around the loess curve of best fit (96 %) but a few probes were highlighted due to high means and standard deviation: *KLK4*, *RPS10*, *RPLP2*, *M5MB*, and *RPS11*. Whilst *AR* exons 4-8 and *ITPR1* were highlighted due to low mean and standard deviation.

If a sample has many missing values this could be caused by a technical failure or as a result of too little input material. There were a few samples that seemed to have missing values in the normalised data (A216, A210, A196, M_138_7, M_140_6, M_147_3, M_92_5, M_97_5 and C147_1). These were watched carefully throughout further analyses.

Each NanoString cartridge holds twelve samples. NanoStringNorm uses a *t*-test to identify cartridges that have a significantly different means, standard deviation and levels of positive controls in comparison to the other cartridges. Cartridges 22, 23, 58, 59, 60, 61, 62, 63 and 64 had higher means and standard deviation, whilst cartridges 15, 29, 36, 37, 43 and 65 through 72 had lower detection levels of positive controls.

Looking into the normalisation factors using NanoStringNorm, a number of samples had normalisation parameters that extended beyond 100% difference from the mean and could be influential outliers: (Supplementary Table 2).

5.3.2.2 NanoStringQCPro

NanoStringQCPro provided information on the binding density, field of view (FOV) and the positive controls used for initial normalisation. NanoString is only capable of reading un-overlapped barcodes when digitally scanning the image produced. Twenty-eight samples were identified as having overlapping barcodes typically caused by excess RNA input (Supplementary Table 2). No samples were identified as having less than 80% FOV, meaning there were no technical issues due to loading of cartridges (e.g. bubbles, or insufficient oiling).

The slope in the positive control data shows how well an increase in input is reflected by an increase in counts, measured using a linear model (log(counts) ~ log(input)).

Three samples were highlighted as outliers from the model: M_122_2, M_127_6 and M_131_4. Two of these samples also showed high IQR of positive controls: M_127_6 and M_131_4. NanoString recommends a positive scaling factor between 0.3 and 3. A scaling factor above this range indicates low performance of that lane during the NanoString counting protocol. Six samples' lanes were flagged as such (M_95_1, M_97_6, M_140_6, M_144_1, M_75_3 and M_147_3) and thus were considered with caution.

5.3.3 Experimental and Technical Investigations

5.3.3.1 Sample and Centre Investigations

Comparing the median with the IQR can unveil samples with low medians and/or IQRs (both of which can be problematic). Some samples were identified as such: A210, A216, A517, C147_1, M_97_5, M_138_7, M_149_7 (Figure 5.2), These samples were removed from the analysis. PCA identified a clear clustering of the cell sediment derived samples compared to EV derived samples from multiple centres (Figure 5.3), further highlighting their need to be analysed separately (Chapter 6). PCA on EV derived samples showed some clustering based on location of origin (Figure 5.4). There is evidence of significant differences in overall expression between some origin centres (Mann Whitney U tests; $p < 0.05$; Table 5.5). However, the average Log_2 expression appears to be fairly uniform across the centres (Figure 5.5).

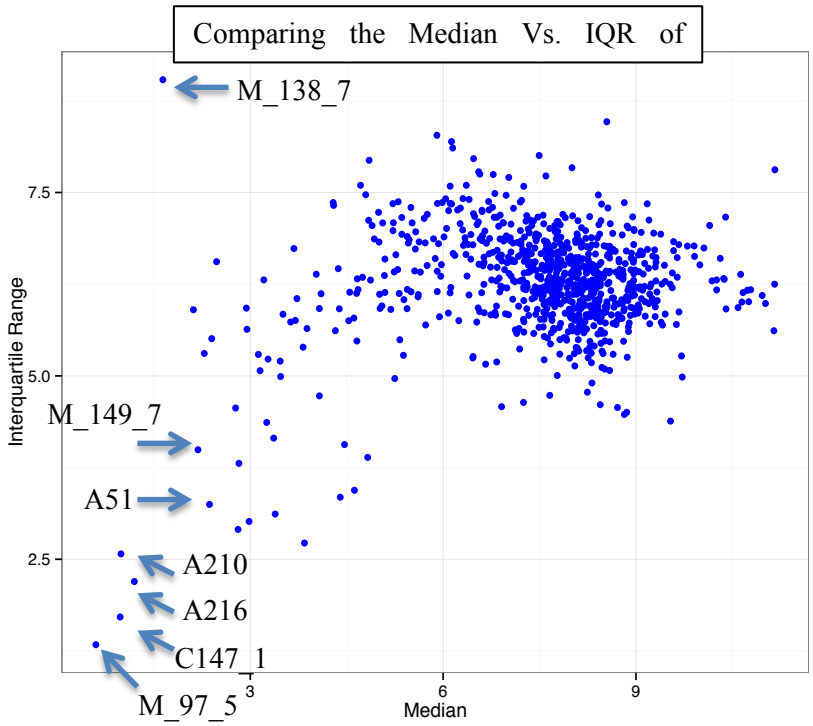


Figure 5.2 Median Vs. IQR of samples on the second NanoString study. Six samples were identified with low medians and/or IQRs, which could be problematic to further analyses.

5.3.3.2 NanoString Cartridges

NanoStringNorm showed significant differences between the mean and standard deviation of the normalised data between some cartridges; indicating there might be batch effects. Cartridge dependent variations were therefore examined using boxplots (Figure 5.6) and there was significant association between mean expression per sample and cartridge (Kruskal-Wallis rank sum test: $p < 2.2\text{x}10^{-16}$, $\chi = 329.25$). As samples from the same collection centres were loaded consecutively, there was no surprise that there was a significant association between centre and cartridge also (Chi-square test; *p*-value $< 2.2\text{x}10^{-16}$, $\chi = 2036.5$). As location was also significantly associated with median expression of samples, it was not a leap to believe this issue with cartridge discrepancies was due to location.

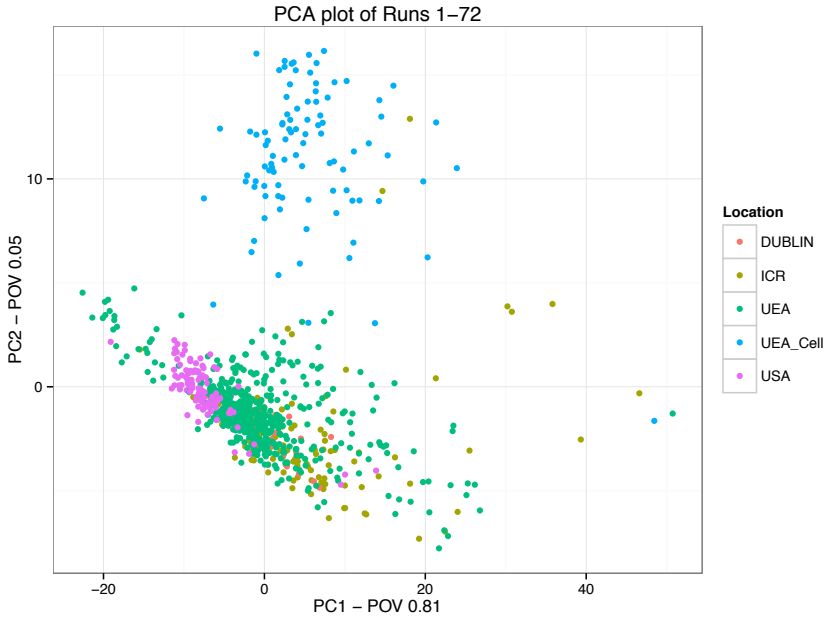


Figure 5.3 DNA extracted from EVs was collected from four different centres (Dublin, ICR, UEA, and the USA). DNA extracted from the cell pellet was only collected at UEA (UEA_Cell). PCA plot clearly identifies cell sediment derived samples as a separate cluster from EV derived samples.

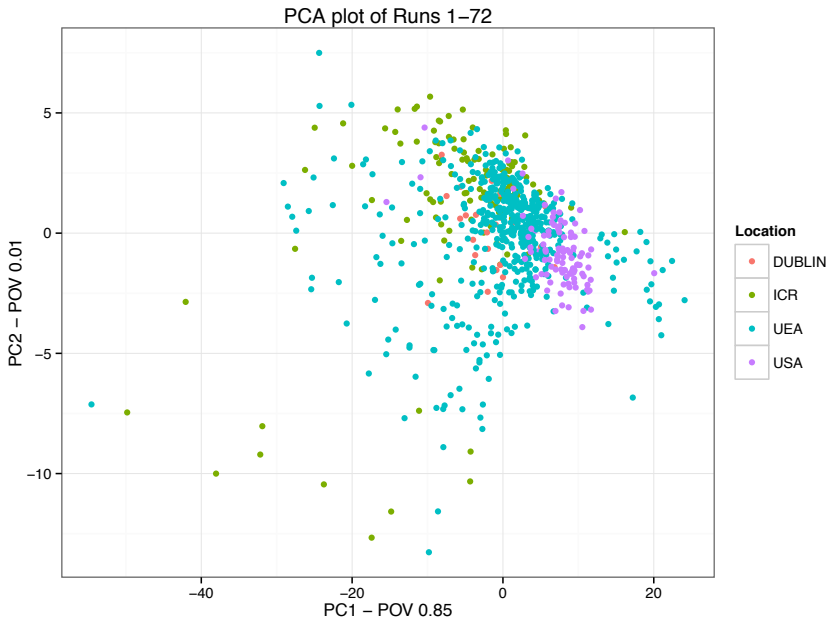


Figure 5.4 PCA plot of only EV derived DNA shows evidence of collection-centre of origin based clustering.

Table 5.5 Expression values from different collection-centres of origin compared by Mann Whitney U tests show that all centres are significantly different.

	USA	*ICR*	*DUBLIN*
UEA	$< 2.2x10^{-16}$	$< 2.2x10^{-16}$	$2.311x10^{-07}$
USA	-	$< 2.2x10^{-16}$	$< 2.2x10^{-16}$

ICR	-	-	$1.704x10^{-11}$

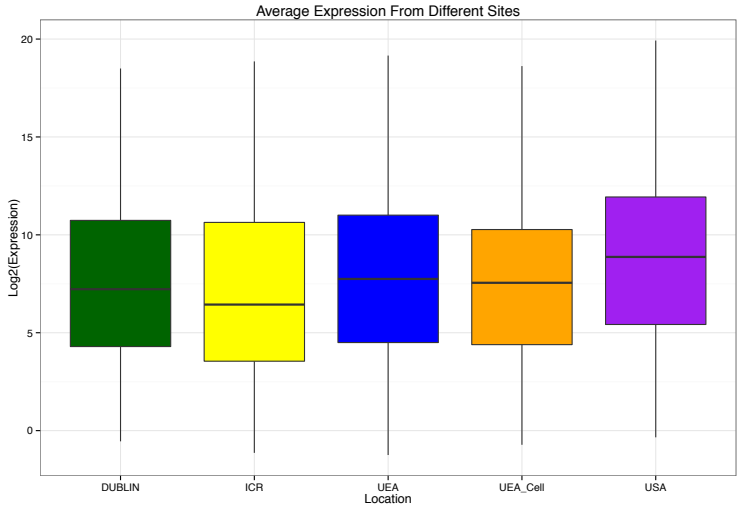


Figure 5.5 Average Log_2 expression across centres shows similar expression levels.

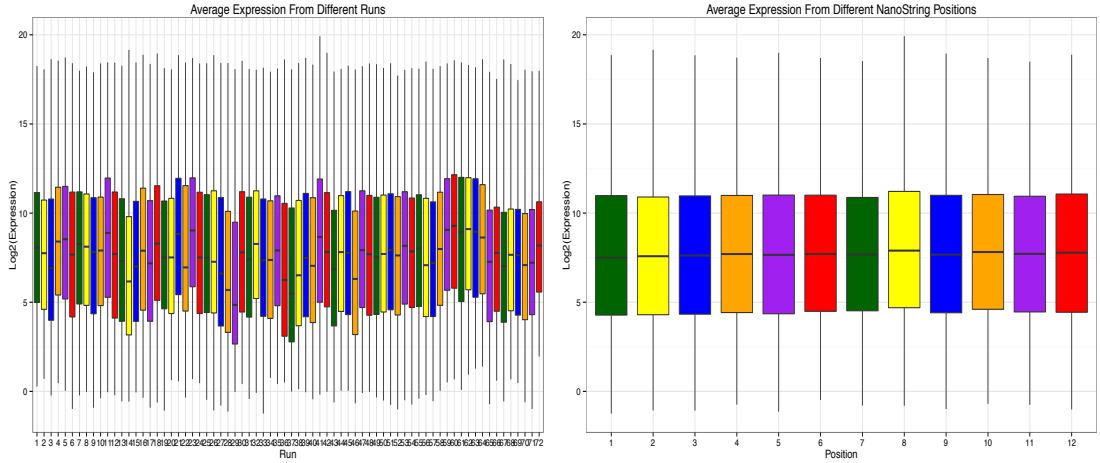


Figure 5.6 Boxplots showing average expression across cartridge and position on cartridge are similar and are showing no batch effects.

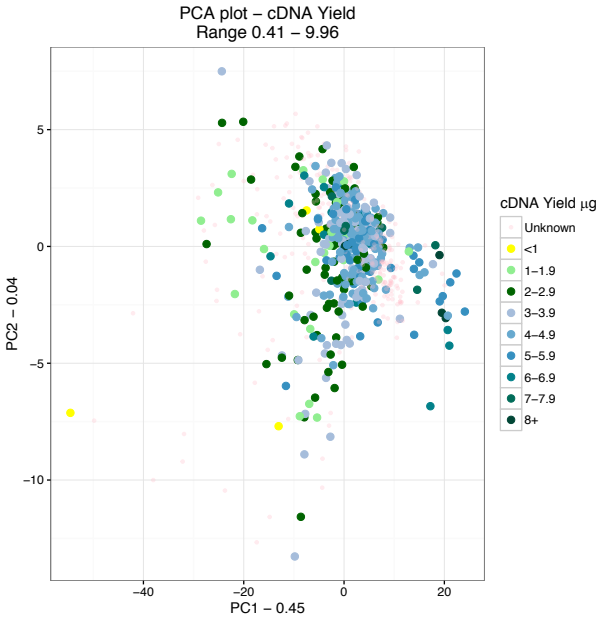


Figure 5.7 PCA plot of EV derived samples, showing a lack of clustering by cDNA yield.

5.3.3.3 RNA Amplification to cDNA

As 100 ng of RNA or cDNA is required for NanoString analysis, and the amounts of EV RNA harvestable from urine were limiting in a large proportion of samples, 15-20ng RNA from each sample was amplified using a Nugen Ovation WTA2 cDNA amplification kit. The amount of cDNA obtained after amplification (in µg) was investigated for clustering affects using PCA (Figure 5.7). cDNA yields were split into groups; <1 µg, 1-1.9µg, 2-2.9µg, 3-3.9µg, 4-4.9µg, 5-5.9µg, 6-6.9µg, 7-7.9µg and >8µg. Mild clustering affects were observed, and a significant correlation was found between cDNA yield and median $\log_2$ expression per sample ($p < 2.2\text{x}10^{-16}$, $r = 0.44$, Pearson's correlation). The distribution of clinical categories within each amplification yield group was not statistically significant; ($\chi = 125.3$, $p > 0.05$, χ^2 test (section 2.4.4)).

5.3.4 ComBat – Removing collection-centre based significance

Batch effects caused by location of sample origin (centre) were accounted for by using the ComBat function of the sva package. PCA was then used to visualise clustering in the post-ComBat data (Figure 5.9). There was no significant difference between median Log_2 expression across location (Kruskal-Wallis rank sum test: $p = 0.6488$, $\chi = 1.647$).

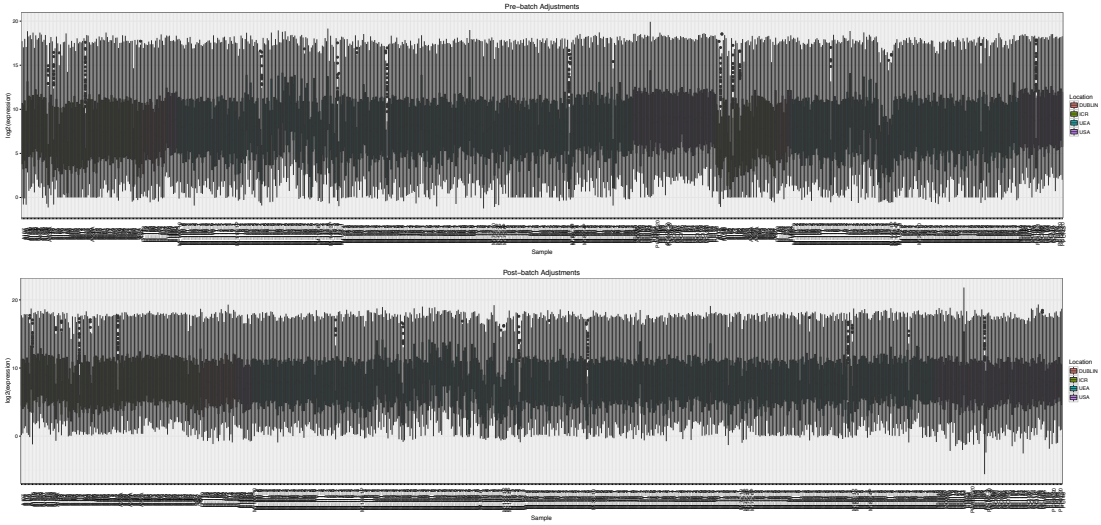


Figure 5.8 Boxplots show the log2 expression across each sample, coloured by location before and after the application of ComBat.

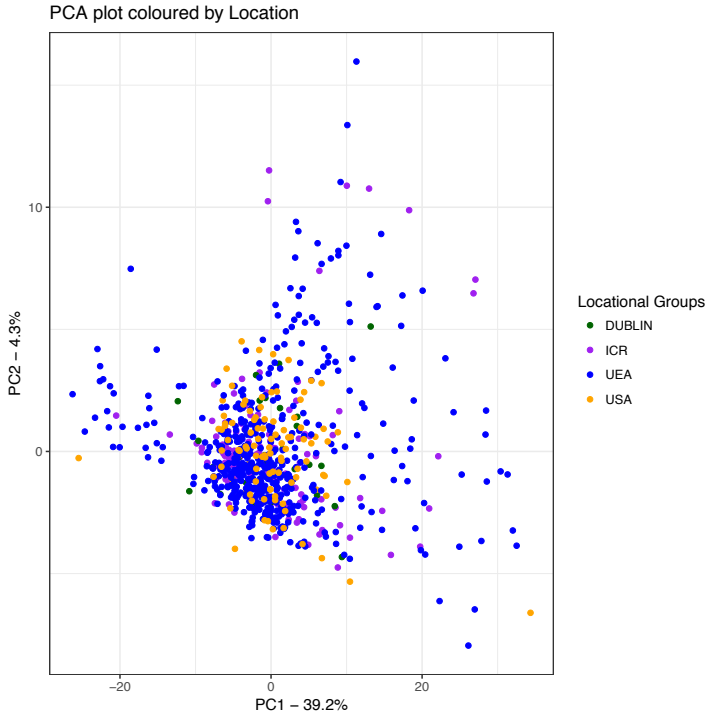


Figure 5.9 PCA plots of post-ComBat data, shows no clustering by location of origin.

5.3.5 Correlating Gene Probes

Pearson's correlation was used to identify correlating probes (Figure 5.10). There were a number of probes that correlated with $R > 0.8$. The correlations were: *CACNA1D* with *GABARAPL2* ($R = 0.965$). *ERG3*' exons 4-5 with *TMPRSS2:ERG* ($R = 0.843$). *GABARAPL2* with *CACNA1D*, *MED4* and *RPS11* ($R = 0.965$, $R = 0.805$, and R =

0.804, respectively). *RPLP2* with *RPS11* and *TWIST1* ($R = 0.859$ and $R = 0.814$, respectively). *RPS10* with *RPS11* ($R = 0.857$). *RPS11* with *GABARAPL2, RPLP2* and *RPS10* ($R = 0.804$, $R = 0.859$, and $R = 0.857$, respectively). *TWIST1* with *RPLP2* ($R = 0.814$). Whilst *KLK3* exons 1-2 and *KLK3* exons 2-3 correlated with each other ($R = 0.839$ and $R = 0.839$, respectively).

These data correlations were encouraging as many of them fitted with published expression data, for example, expression of *TMPRSS2:ERG* and *ERG3*', and the two *KLK3* probes. *RPL11* is known to be co-expressed with RPL10.

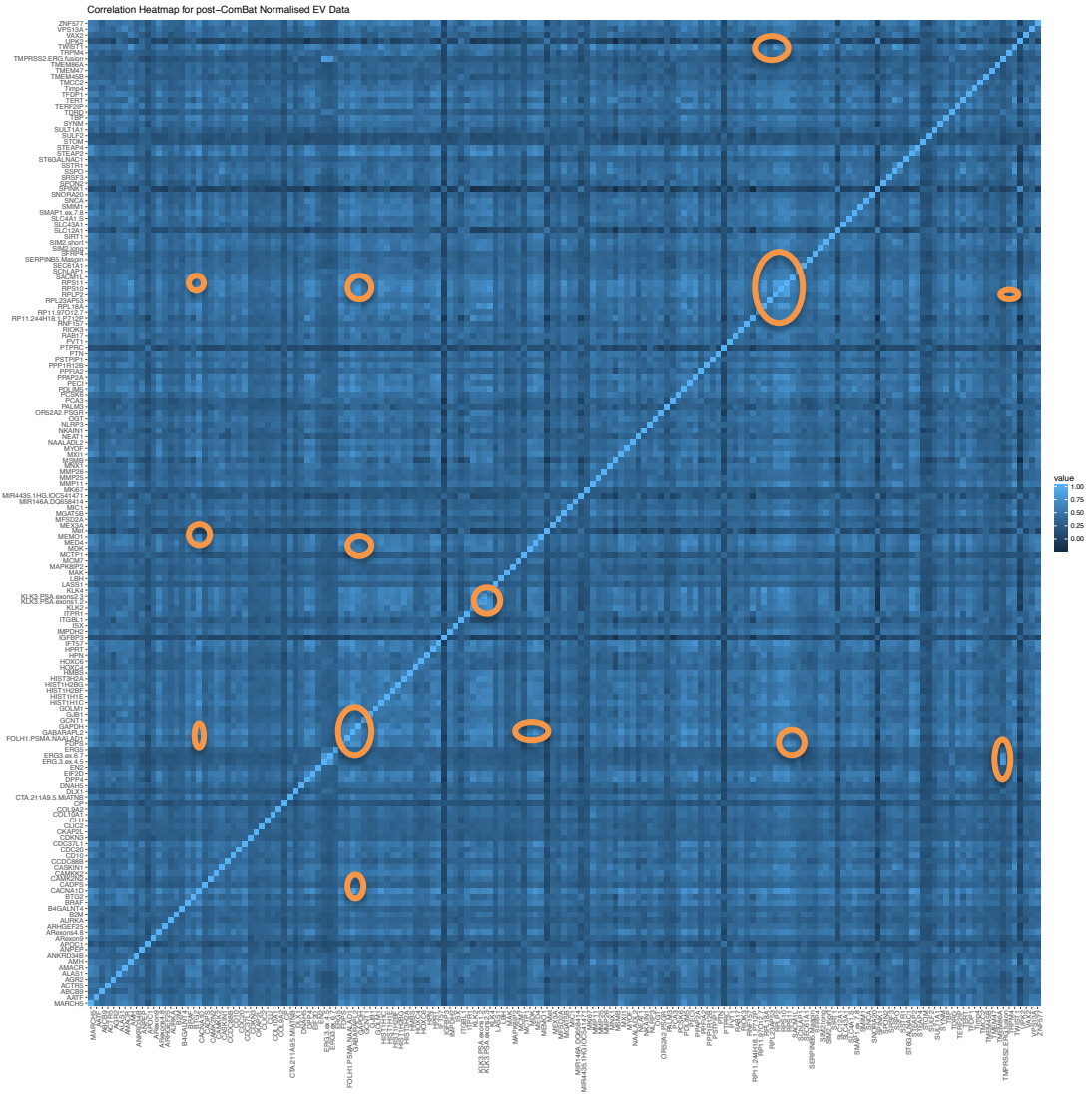


Figure 5.10 Heatmap showing correlation between NanoString Probes in post-ComBat data. *R*-values between 0 (darker) and 1 (lighter). Correlations with $R > 0.8$ have been highlighted.

5.3.6 Comparison of NanoString2 with NanoString1

Comparing the forty-nine common probes across the one hundred and thirty one common samples between NanoString1 and NanoString2 yielded three probes with a Pearson's correlation $R<0.6$: *Timp4* ($R = 0.14$), *TMPRSS2:ERG* ($R = 0.18$), and *TERT* ($R = 0.38$). Twenty-one of the probes showed high correlation, with $R>0.9$ (
Table 5.6).

Table 5.6 Pearson's Correlation between the 49 common probes and 131 common samples between NanoString1 and NanoString2.

Probe	*R*	*Probe*	*R*	*Probe*	*R*
HOXC6	***0.98***	***CLU***	***0.92***	***HPN***	***0.83***
ERG3'exons 6-7	***0.97***	***KLK3 exons 2-3***	***0.92***	***GAPDH***	***0.83***
SPINK1	***0.97***	***KLK3 exons 1-2***	***0.92***	***HOXC4***	***0.82***
SULT1A1	***0.97***	***CAMKK2***	***0.91***	***AURKA***	***0.82***
KLK2	***0.96***	***STEAP4***	***0.90***	***BRAF***	***0.81***
AR exons 4-8	***0.96***	***ANPEP***	***0.90***	***PCA3***	***0.80***
KLK4	***0.95***	***AGR2***	***0.90***	***PPAP2A***	***0.78***
AR exon 9	***0.95***	***B2M***	***0.89***	***IMPDH2***	***0.78***
UPK2	***0.95***	***PECI***	***0.89***	***OGT***	***0.77***
FOLH1	***0.95***	***PTPRC***	***0.89***	***CDC20***	***0.71***
ALAS1	***0.94***	***DLX1***	***0.89***	***MKi67***	***0.67***
AMACR	***0.94***	***MDK***	***0.89***	***ERG5'***	***0.63***
TDRD	***0.93***	***MMP26***	***0.87***	***TERT***	***0.37***
SLC12A1	***0.93***	***NAALADL2***	***0.87***	***TMPRSS2:ERG***	***0.18***
SERPINB5	***0.93***	***TBP***	***0.86***	***Timp4***	***0.14***
GOLM1	***0.93***	***CDKN3***	***0.85***		
STEAP2	***0.93***	***HPRT***	***0.83***		

5.4 Identification of Prostate and Cancer Specific Transcripts and DRE relevance

5.4.1 Kallikrein identification

NanoString median signals for the *KLK2*, *KLK3* exons 1-2, *KLK3* exons 2-3 and *KLK4* probes were again at significantly higher levels than those for the control tissue probes for blood, kidney and bladder (*PTPRC*, *SLC12A1* and *UPK2* respectively) (Mann Whitney U test: $p < 2.2\text{x}10^{-16}$ in each case, Figure 5.11). This was seen previously in NanoString1 (section 3.4.1) and shows that some of the material collected did originate

from the prostate. Once again, similar expression levels and a correlation, is observed between the two *KLK3* probes (Pearson's correlation: $R = 0.84$, $p < 2.2\text{x}10^{-16}$).

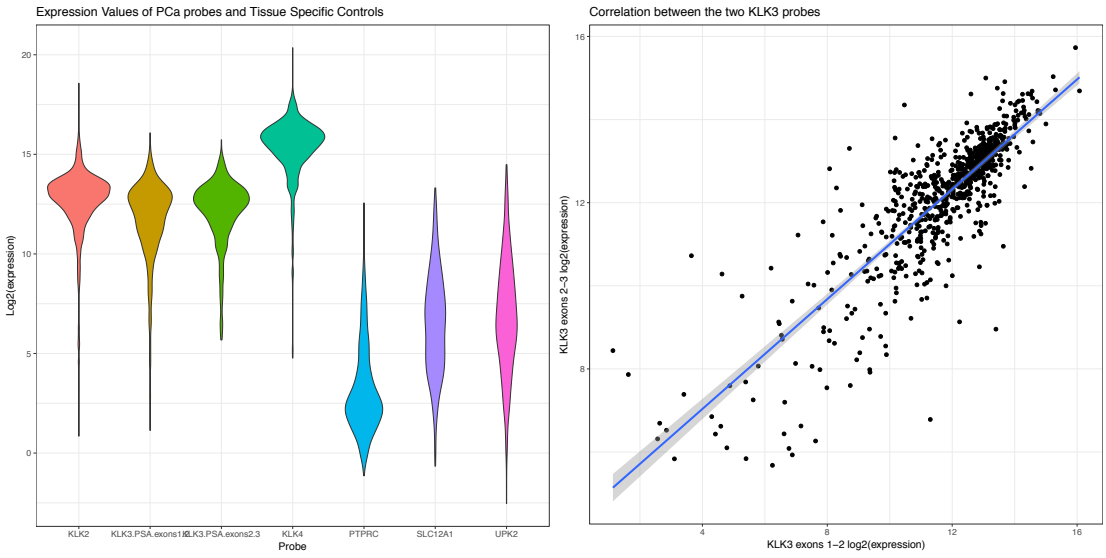


Figure 5.11 KLK2, KLK3 and KLK4 expression is higher than the tissue specific controls for blood, kidney and bladder. The two KLK3 probes are highly correlated (Pearson's correlation: $R = 0.84$, $p < 2.2\text{x}10^{16}$).

5.4.2 *TMPRSS2:ERG* Identification

Similar results can be seen in regards to the *TMPRSS2:ERG* fusion gene, the *ERG3'* probes and *ERG5'*, as in NanoString1 (section 3.4.2). *TMPRSS2:ERG* fusions, *ERG* 3' and *ERG* 5' expression are linked to PCa, and are therefore expected to be seen more prevalently in samples obtained from men with known PCa compared to those with no clinical evidence of PCa (CBN samples) (Mann Whitney U test between respective probe's expression values and local cancer (low-, intermediate- and high-risk cancer)/CBN groupings. (*TMPRSS2:ERG*: $p < 2.2\text{x}10^{-16}$, *ERG* 3' exons 4-5: $p < 2.2\text{x}10^{-16}$; *ERG* 3' exons 6-7: $p < 2.2\text{x}10^{-16}$; and *ERG* 5': $p = 1.572\text{x}10^{-08}$). The density plots for *TMPRSS2:ERG* and the *ERG3'* probes (Figure 5.12) have two peaks which would be compatible with an on/off pattern for that probe suggesting that approximately 50% of the samples from men with cancer have detectable *TMPRSS2:ERG* fusions (which is in agreement with the literature available (section 1.4.6) and the results from NanoString1, (section 3.4.2)).

A larger proportion of the CBN and raised PSA negative Bx (S) samples do not have high expression of *TMPRSS2:ERG,* compared to the cancer samples. The cancer samples across all clinical categories and abnormal (including HG:PIN, prostatitis and atypia samples) have fewer samples with lower *TMPRSS2:ERG* expression. The CBN samples also show lower numbers with high *TMPRSS2:ERG* expression, however there are a few (as expected).

The *ERG5*' probe, which is not part of the *TMPRSS2:ERG* fusion transcript, is not significantly different between clinical risk categories. This is also seen in NanoString1 (section 3.4.2). These results suggest that the second set of NanoString data is detecting transcripts accurately and that a proportion of the genetic material identified is coming from PCa or HG-PIN, again.

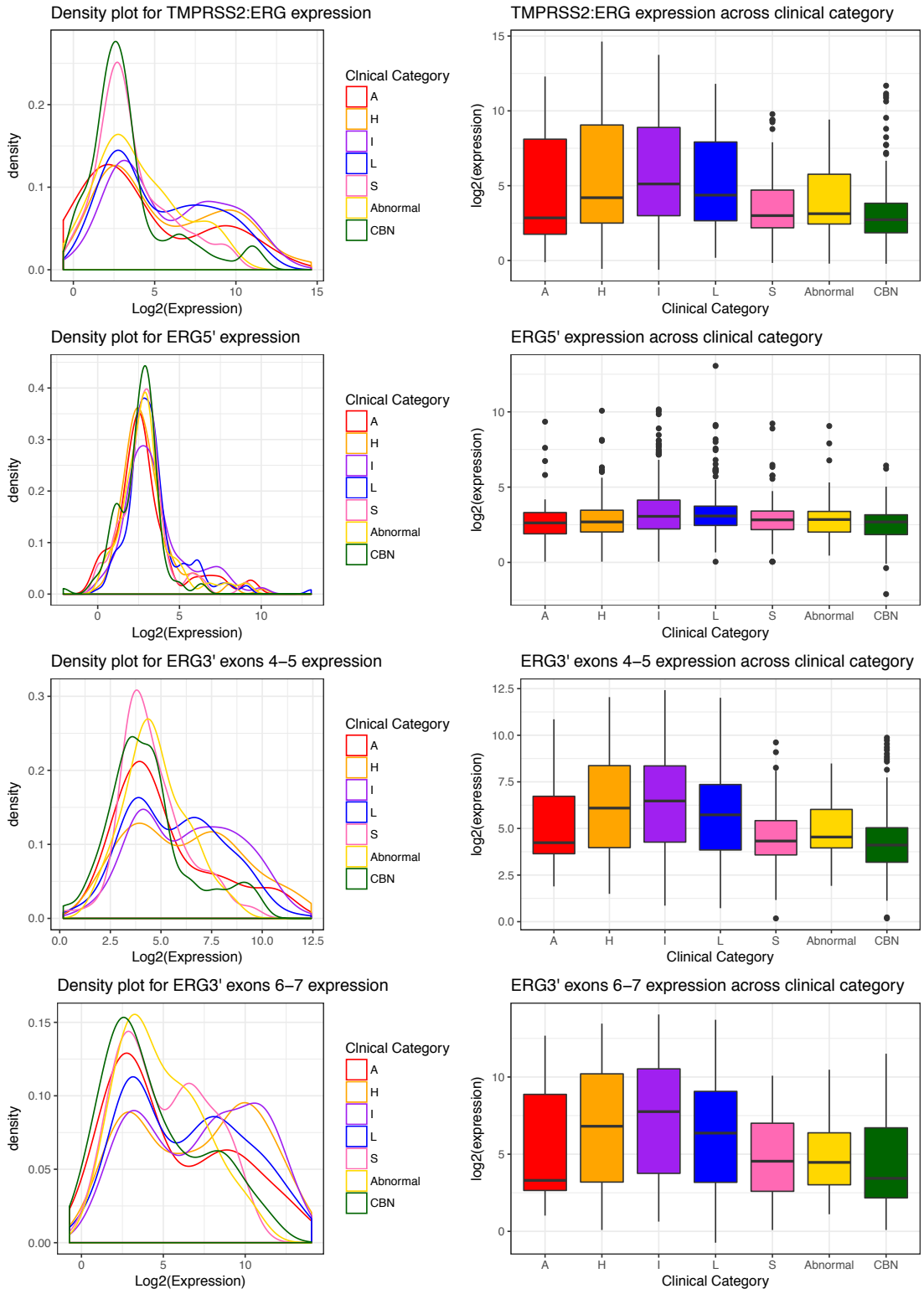


Figure 5.12 Density plots and Boxplots showing the expression changes of *TMPRSS2:ERG*, two *ERG* 3' probes, and *ERG* 5' across clinical categories.

5.4.3 PCA3 Test

As in the NanoString1 (section 3.4.3) data, the PCA3 test was significantly different between PCa (Advanced, high-risk, intermediate-risk and low-risk) samples and CBN

samples (Kruskal-Wallis rank sum test: $p = 6.2\text{x}10^{-09}$, $\chi^2 = 33.76$ and Mann Whitney U test: $p < 2.2\text{x}10^{-16}$, Figure 5.13). There are some significant differences across clinical categories also ($p < 0.05$; Mann-Whitney U test; Table 5.7).

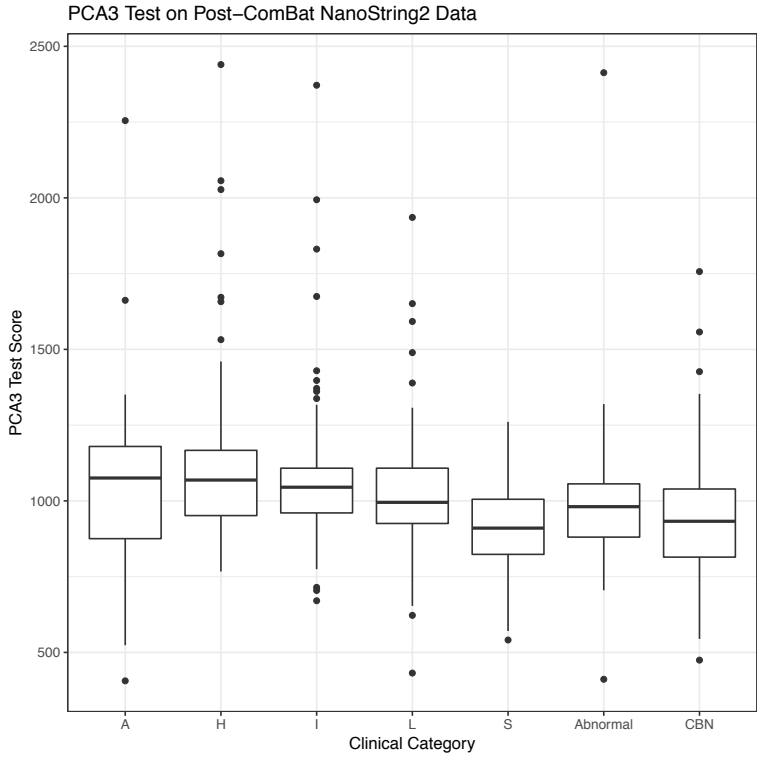


Figure 5.13 PCA3 Test on post-ComBat NanoString2 data (PCA3 transcript expression/average KLK3 transcript expression * 1000)

Table 5.7 Mann Whitney U test of PCA3 Test scores between the different clinical categories.

p-value	*Advanced*	*High-risk*	*Intermediate-Risk*	*Low-Risk*	*High PSA negative Bx*	*Abnormal*	*CBN*
Advanced		***0.657***	***0.756***	***0.255***	***0.003 (Up in A)***	***0.095***	***0.021 (Up in A)***
High-Risk	***0.657***		***0.126***	***0.004 (Up in H)***	***3.14x10^{-10} (Up in H)***	***5.5x10^{-04} (Up in H)***	***1.3x10^{-07} (Up in H)***
Intermediate-Risk	***0.756***	***0.126***		***0.024 (Up in I)***	***2.7x10^{-12} (Up in I)***	***0.001 (Up in I)***	***1.2x10^{-08} (Up in I)***
Low-Risk	***0.255***	***0.004 (Up in H)***	***0.024 (Up in I)***		***3.4x10^{-07} (Up in L)***	***0.101***	***1.0x10^{-04} (Up in L)***
High PSA negative Bx	***0.003 (Up in A)***	***3.14x10^{-10} (Up in H)***	***2.7x10^{-12} (Up in I)***	***3.4x10^{-07} (Up in L)***		***0.029 (Up in Abnormal)***	***0.408***
Abnormal	***0.095 (Up in A)***	***5.5x10^{-04} (Up in H)***	***0.001 (Up in I)***	***0.101***	***0.029 (Up in Abnormal)***		***0.189***
CBN	***0.021 (Up in A)***	***1.3x10^{-07} (Up in H)***	***1.2x10^{-08} (Up in I)***	***1.0x10^{-04}***	***0.408***	***0.189***	

5.5 Clustering

5.5.1 Principal Component Analysis

PCA (section 2.5.1) shows no significant clustering by clinical category (Kruskal-Wallis rank sum test: $p = 0.2064$, $\chi = 8.5$).

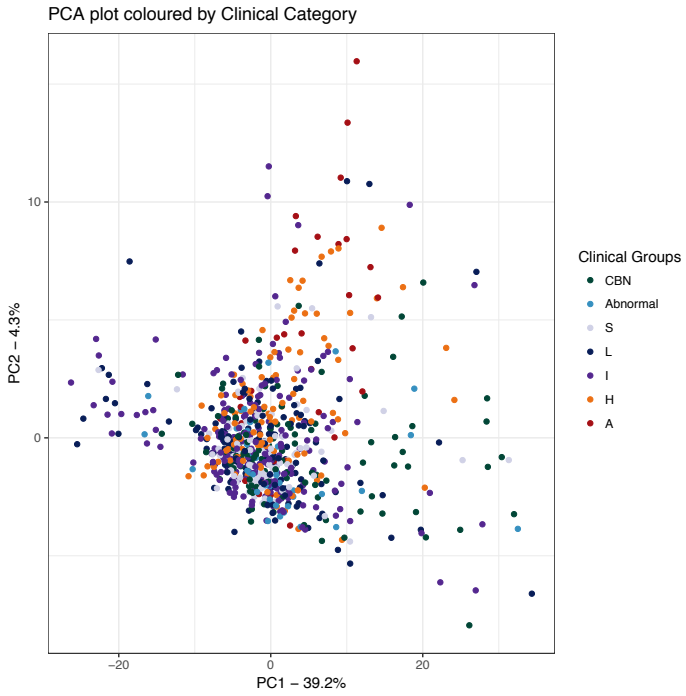


Figure 5.14 PCA plot of post-ComBat data, shows no clustering by clinical category.

5.5.2 Latent Process Decomposition (LPD)

LPD (section 2.5.5) was applied to the dataset for three hundred and forty-six of the training samples. There were predicted to be five clusters in the data, with a sigma parameter of -1. LPD analysis was then performed 100 times using these parameters. A significant association was found between LPD group and clinical risk group (Chi-square: $p = 7.46\text{x}10^{-14}$, $\chi = 115$, Figure 5.15) but not the sample origin (Chi-square: $p = 0.095$, $\chi = 18.7$, Figure 5.17, Table 5.8, Figure 5.18). This suggests that this data set is picking up on underlying processes in the NanoString2 data that effects clinical risk. Figure 5.16 shows the clinical breakdown of each LPD group. There appeared to

be an over-representation of CBN samples in LPD 1 but this was not significant (Chi-square test CBN vs. low-, intermediate- and high-risk cancer: p-value = 0.09, X^2 = 2.8). LPD2 had an over representation of localised cancer (low-risk and intermediate-risk) Chi-square test: p-value = 0.037, X^2 = 4.3. Whilst LPD3 showed a significant over-representation of more progressed cancer (high-risk/advanced cancer) Chi-square test: p-value = 1.671×10^{-07}, X^2 = 31.2. There was no significant over-representation of cancer (advanced, high-, intermediate- and low-risk) or CBN samples in either LPD4 or LPD5. All cancer vs. CBN, more progressed cancer vs. localised cancer vs. CBN were both tested.

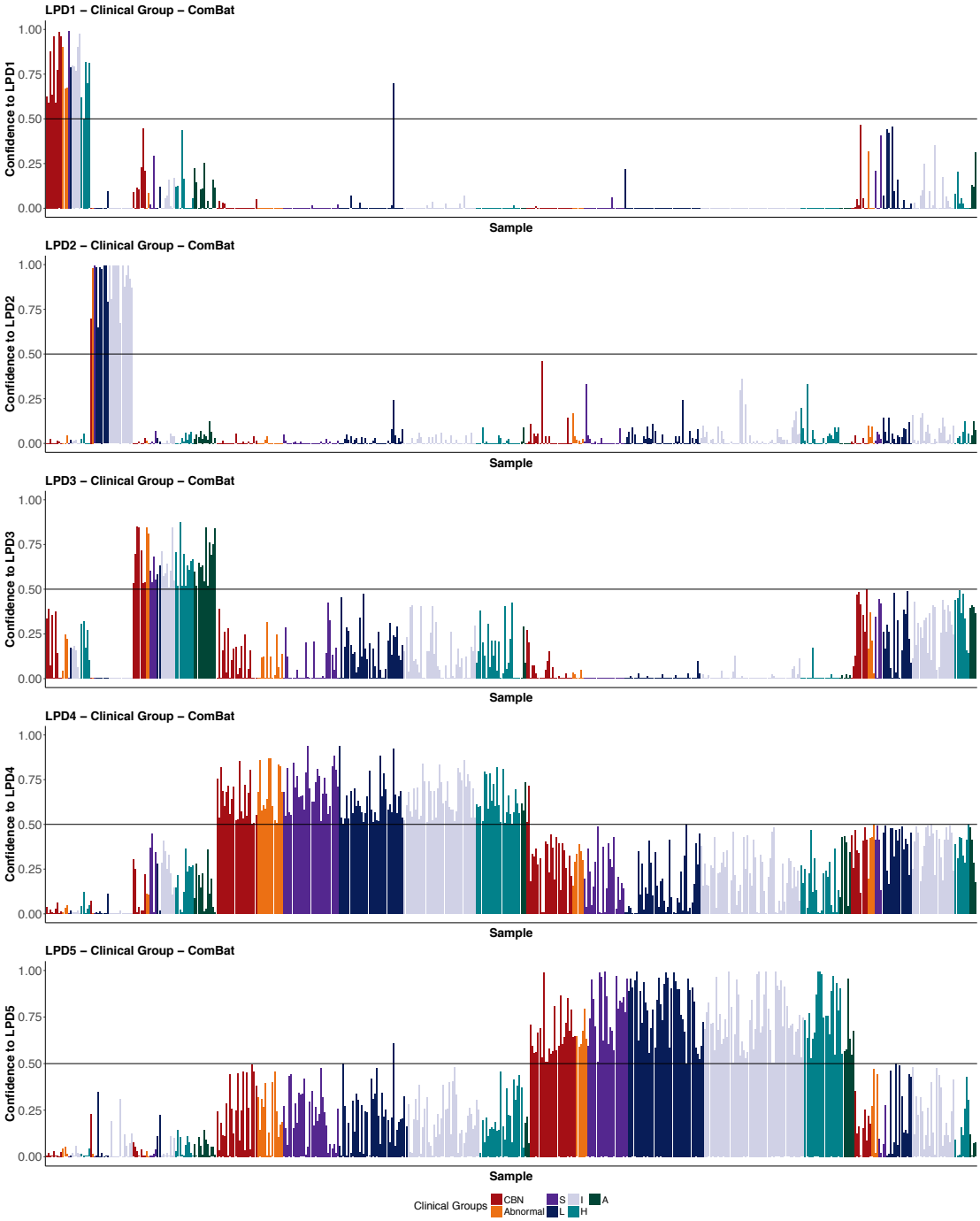


Figure 5.15 LPD of post-ComBat data separated into five processes and coloured by clinical category.

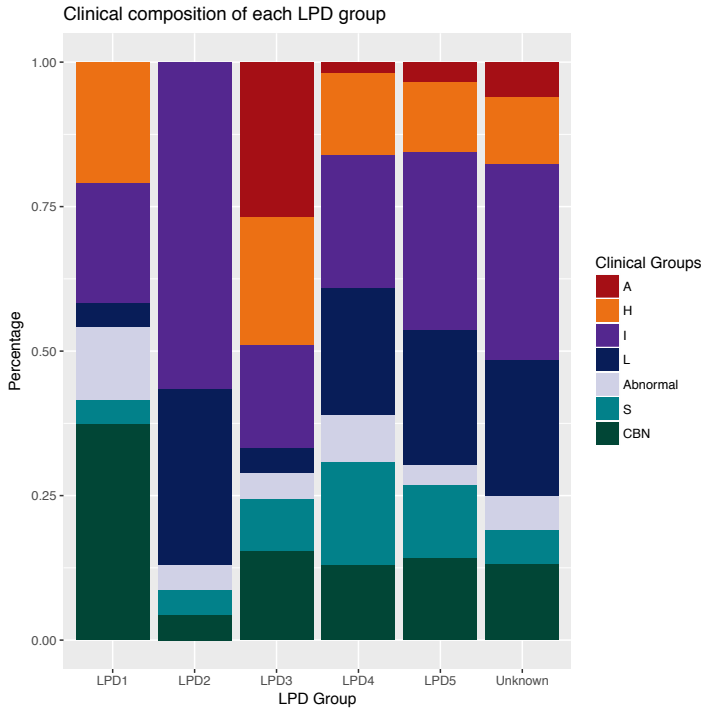


Figure 5.16 Clinical breakdown of each LPD group. Chi-square test: *p*-value = 7.46×10^{-14}, $X^2 = 115$ (ignoring samples from unknown LPD group).

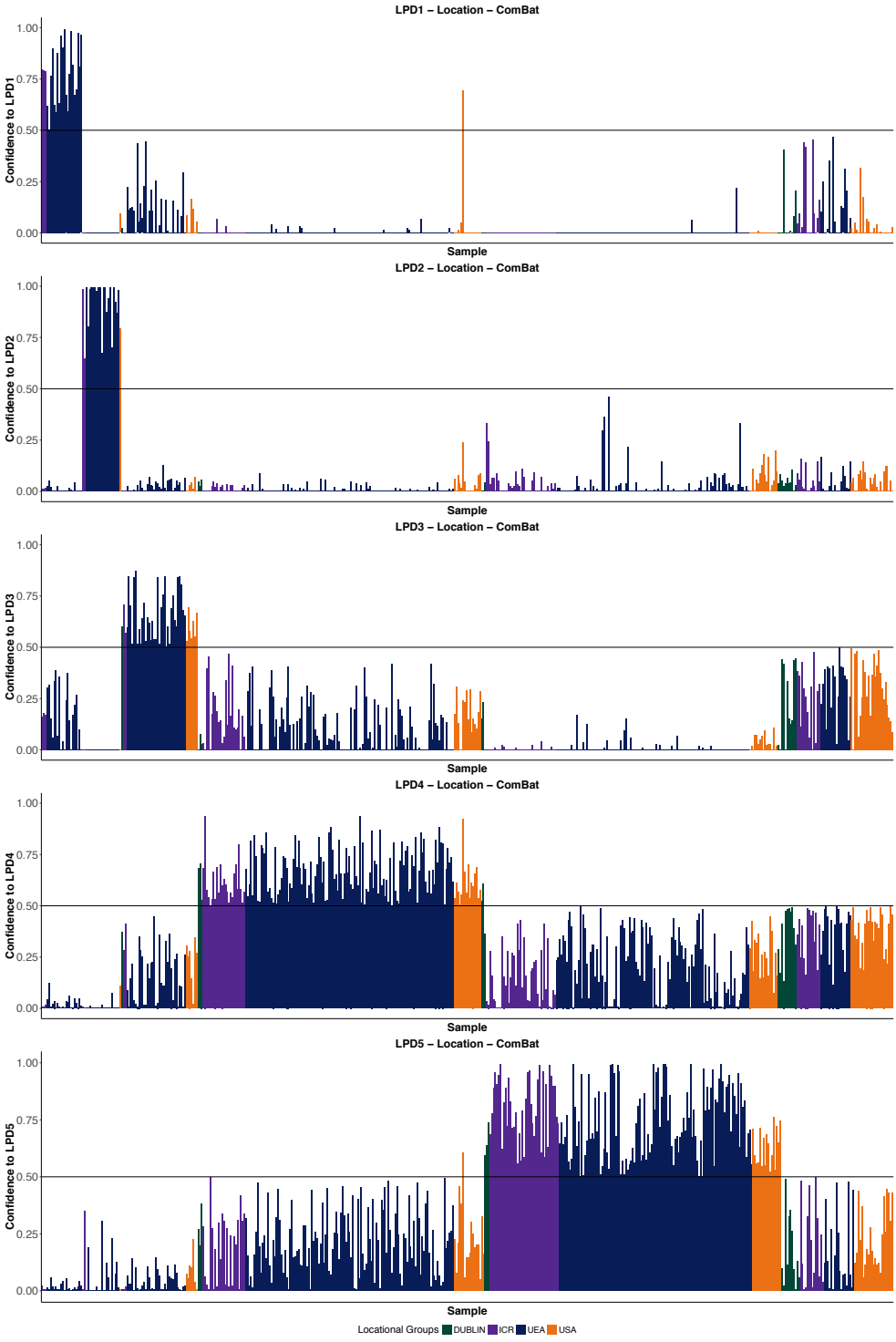


Figure 5.17 LPD of post-ComBat data separated into five processes and coloured by location of origin.

Table 5.8 Location of origin breakdown of LPD groups.

	LPD1	*LPD2*	*LPD3*	*LPD4*	*LPD5*
DUBLIN	***0***	***0***	***1***	***3***	***3***
ICR	***3***	***2***	***2***	***25***	***41***
UEA	***21***	***20***	***35***	***123***	***114***
USA	***0***	***1***	***7***	***18***	***17***
Total	***24***	***23***	***45***	***169***	***175***

There were 167, 166, 131, 61, & 153 transcripts that were significantly differentially expressed in LPD processes 1-5 respectively vs. the rest ($p < 0.05$ after multiple testing correction, Mann-Whitney U test: section 2.4.1). Looking at the top 10 most significant associated transcripts shows a decrease in expression in LPD groups 1, 3 and 4 and an increase in expression in LPD groups 2 and 5 (Table 5.9).

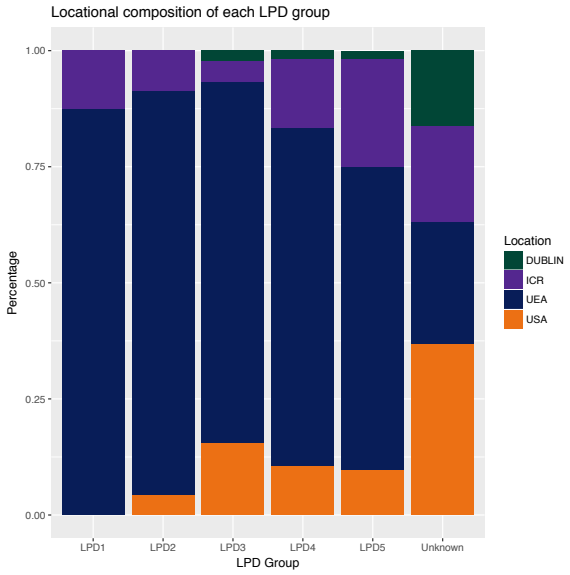


Figure 5.18 Location of origin breakdown of each LPD group. Chi-square test: *p*-value = 0.095, X^2 = 18.7 (ignoring unknown LPD group samples).

Table 5.9 Top ten significantly associated transcripts involved in the separation of samples into LPD groups. The p-value shown is adjusted using Benjamin Hochberg multiple testing correction.

LPD Group	*LPD1*	*p-value*	*Log_2(FC)*	*LPD2*	*p-value*	*Log_2(FC)*	*LPD3*	*p-value*	*Log_2(FC)*
# Sig Genes	***167***			***166***			***131***		
Top 10:	***CAMKK2***	***$4.80X10^{-14}$***	***-1.17***	***IFT57***	***$9.50X10^{-14}$***	***0.18***	***KLK2***	***$1.97X10^{-12}$***	***-0.20***
	CACNA1D	***$1.10X10^{-13}$***	***-0.49***	***OGT***	***$1.26X10^{-13}$***	***0.27***	***DPP4***	***$2.20X10^{-12}$***	***-0.23***
	GABARAPL2	***$1.10X10^{-13}$***	***-0.32***	***GABARAPL2***	***$1.31X10^{-13}$***	***0.17***	***CASKIN1***	***$1.29X10^{-10}$***	***-0.21***
	RPS11	***$3.13X10^{-13}$***	***-0.13***	***DPP4***	***$1.56X10^{-13}$***	***0.19***	***MSMB***	***$1.34X10^{-10}$***	***-0.08***
	RPL23AP53	***$3.74X10^{-13}$***	***-1.35***	***IMPDH2***	***$1.65X10^{-13}$***	***0.26***	***CACNA1D***	***$1.55X10^{-10}$***	***-0.20***
	PPAP2A	***$3.94X10^{-13}$***	***-0.33***	***HPRT***	***$1.68X10^{-13}$***	***0.30***	***GABARAPL2***	***$1.71X10^{-10}$***	***-0.14***
	CTA.211A9.5/MIATNB	***$4.44X10^{-13}$***	***-2.43***	***EIF2D***	***$1.69X10^{-13}$***	***0.25***	***TERT***	***$2.02X10^{-10}$***	***-0.24***
	STEAP2	***$5.07X10^{-13}$***	***-0.60***	***MXI1***	***$2.05X10^{-13}$***	***0.22***	***ZNF577***	***$2.69X10^{-10}$***	***-0.26***
	IFT57	***$8.73X10^{-13}$***	***-0.33***	***PECI***	***$2.09X10^{-13}$***	***0.25***	***SSPO***	***$3.12X10^{-10}$***	***-0.20***
	MIC1	***$8.77X10^{-13}$***	***-1.20***	***RP11.97O12.7***	***$2.10X10^{-13}$***	***0.28***	***CAMK2N2***	***$3.32X10^{-10}$***	***-0.52***

LPD Group	***LPD4***	***p-value***	***Log_2(FC)***	***LPD5***	***p-value***	***Log_2(FC)***
# Sig Genes	***61***			***153***		
Top 10:	***VPS13A***	***$3.38X10^{-06}$***	***-0.11***	***GABARAPL2***	***$2.26X10^{-22}$***	***0.07***
	TERF2IP	***$3.79X10^{-06}$***	***-0.05***	***CACNA1D***	***$2.71X10^{-21}$***	***0.09***
	ABCB9	***$1.47X10^{-05}$***	***-0.21***	***STEAP2***	***$3.26X10^{-17}$***	***0.09***
	MARCH5	***$1.64X10^{-05}$***	***-0.08***	***KLK2***	***$4.09X10^{-17}$***	***0.07***

	MMP25	***1.89X10^{-05}***	***-0.25***	***MED4***	***2.31X10^{-16}***	***0.09***
	TMEM45B	***1.92X10^{-05}***	***-0.14***	***CASKIN1***	***1.66X10^{-15}***	***0.13***
	RPLP2	***1.93X10^{-05}***	***-0.03***	***DPP4***	***7.40X10^{-15}***	***0.07***
	PECI	***2.41X10^{-05}***	***-0.06***	***IFT57***	***8.66X10^{-15}***	***0.07***
	CASKIN1	***2.64X10^{-05}***	***-0.10***	***RPS11***	***8.75X10^{-14}***	***0.03***
	MEMO1	***3.23X10^{-05}***	***-0.08***	***MARCH5***	***9.67X10^{-14}***	***0.09***

5.6 Further processing techniques

After positive control normalisation and $\log_2$ transformation, four further processing techniques were used. These included adjusting the data to focus on the prostate derived proportion by: using *KLK3* as per the PCA3 test (section 2.1.1), using *KLK2* in a similar way, and using a *KLK2* ratio (section 2.1.1). In addition, *RPLP2* and *GAPDH* were identified as novel housekeeper genes and used to normalise to the amount of material. *RPLP2* and *GAPDH* did not have any significant association with clinical category ($p < 0.05$; Tukey test (section 2.4.7)) and had a strong correlation ($r = 2.2\text{x}10^{-16}$, Pearson's correlation, section 2.4.3). Each of these methods were used to create a data set and subsequently to build clinical prediction models (Table 5.10). The *KLK2* and *KLK3* adjusted data also included the removal of CBN with high *TMPRSS2:ERG*. As CBN samples were from patients with no clinical evidence of cancer rather than strictly benign, it was expected that there would be some cancer present in some of the samples. Removal of high *TMPRSS2:ERG* CBN samples, was a step towards correcting for this.

Samples with low *KLK2* and *KLK3* values were also removed. These are prostate-expression specific control transcripts. Eliminating these data, removed samples where the majority of the RNA was not originating from the prostate.

Table 5.10 The different normalisations of the data that the predictive models were built using (separately).

Data	*Description*
KLK2 ratio	***The ratio of KLK2 was used to normalise the data***
KLK2 adjusted	***Low KLK2 removed and high TMPRSS2:ERG removed.*** ***Median and IQR used to adjust data***
KLK3 adjusted	***Low KLK3 removed and high TMPRSS2:ERG removed.*** ***Median and IQR used to adjust data***
Housekeeper normalised – GAPDH and RPLP2	***KLK2 ratio data, further normalised via GAPDH and RPLP2***

5.7 Clinical Prediction models

The data were stratified into test and training sets in the ratio 1:2 (Table 5.1) weighted according to sample origin and clinical risk category. Models were built to predict four different response variables i.e. clinical questions (Table 5.11) using each of the four different processed datasets (Section 5.6) using the training samples.

For models predicting a binary variable, logistic regression (section 2.6.1) and Mann Whitney U (section 2.4.1) tests were used to identify transcripts that individually could predict the two groups ($p < 0.05$). For models predicting an ordinal variable, univariate proportional odds models (polr) were used to identify significant transcripts ($p < 0.05$). Multiple testing correction using Benjamin Hochberg was applied.

For each clinical question, final models were built using LASSO using three input criteria:

1. All 167 probes
2. Probes that were identified as significant in univariate analyses ($p < 0.05$; no multiple testing correction)
3. Probes that were identified as significant in univariate analyses when multiple testing correction was applied (Benjamin Hochberg corrected $p < 0.05$)

Models were then applied to the test datasets, where the specificity, sensitivity and PPV of each model was determined (Table 5.13, Table 5.15, Table 5.17, Table 5.21).

Table 5.11 Clinical predictive models built using the training set and tested using the test set.

Model	*Samples*	*Model type*
CB vs. Cancer	***Clinically benign samples Vs low-, intermediate-, and high-risk cancer samples grouped together***	***Binary***
CB vs. High risk cancer	***Clinically benign Vs. high-risk cancer (extreme ends of no evidence of cancer and and those with higher grade)***	***Binary***
CB, low-, intermediate-, and high-risk trend	***Each sample category is a separate group and ordered***	***Ordinal***
CB, cancer, metastatic cancer trend	***Clinically benign samples, with low-, intermediate-, and high-risk cancer samples grouped together, and metastatic cancer samples in groups ordered by severity***	***Ordinal***

5.7.1 Models predicting presence of cancer CB and cancer (L, I, H) samples

Expression of 80, 63, 49, 55 probes had a significant association with whether a sample had no evidence for cancer (CB) or not (L, I, H) in the four processed datasets (*KLK2* ratio, *KLK2* adjusted, *KLK3* adjusted, HK normalised, respectively) (Supplementary Table 4). The top probe was *ERG3'* exons 4-5 ($p = 1.54\text{x}10^{-09}$, $\log_2\text{FC} = 1.58$), *PCA3* ($p = 4.5\text{x}10^{-07}$, $\log_2\text{FC} = 0.19$), *PCA3* ($p = 1.61\text{x}10^{-06}$, $\log_2\text{FC} = 0.14$), and *ERG3'* exons 4-5 ($p = 4.5\text{x}10^{-09}$, $\log_2\text{FC} = 0.699$), respectively.

Multivariate models were built to predict whether a patient had cancer (L, I, H samples) or had no evidence for cancer (CB) (Table 5.12, Table 5.13). The ROC curves and probes involved in each model can be found in the supplementary figures (Supplementary Figure 2, Supplementary Figure 3, Supplementary Figure 4,

Supplementary Table 5, Supplementary Table 6, Supplementary Table 7 and Supplementary Table 8, respectively).

In this comparison there were large differences in the number of samples in each of the two categories, with CB having approximately only a quarter of the sample size of cancer. Therefore, random sampling was used to select a similar number of cancer samples to CB samples, and the model predictive process was run iteratively 1,000 times. The model with the mean AUC was selected to be applied to the test dataset. Again, the AUC, Sensitivity, Specificity and PPV and the selected probes were recorded for each model on the training set (Table 5.14) and the test set (Table 5.15) and the curves and probes involved in each model can be observed in the supplementary figures (Supplementary Figure 5, Supplementary Figure 6, Supplementary Figure 7, Supplementary Figure 8, Supplementary Table 9, Supplementary Table 10, Supplementary Table 11, and Supplementary Table 12, respectively).

The models were generally good predictors of whether cancer was present or not (median AUC = 0.8045, IQR = 0.06). In general, AUC in the test data was better in the *KLK2* ratio and the *GAPDH* and *RPLP2* normalised data (all had AUC > 0.8) compared to the *KLK2* and *KLK3* adjusted data (mostly AUC > 0.7). There was not much difference observed between those with the randomly selected cancer samples (median AUC =0.847, IQR =0.11), and those with all of the cancer samples (median AUC = 0.846, IQR = 0.098).

The accuracy measures remained very high in the test sets (median AUC = 0.915, IQR = 0.05, but were slightly lower than the training data set (median AUC = 0.8045, IQR = 0.06), showing the models in general were robust and useful.

The model with the best AUC in the training data, was when using all of the probes from the *RPLP2* and *GAPDH* normalised data (Training AUC = 0.925, Test AUC = 0.851) in detail as an example. 18 transcripts were selected by Lasso and went into these models; *TMPRSS2:ERG, ERG3'* exons 4-5, *APOC1, ISX, SLC12A1, HOXC6, MCTP1, TDRD, PDLIM5, CD10, GABARAPL2, PTN, AR* exon 9, *PPP1R12B, CP,*

MXI1, and *KLK4*. The training model had 85% sensitivity, 73% specificity and 94% PPV (Figure 5.19).

Table 5.12 Training model outcomes comparing CB with Cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
AUC	***0.949***	***0.886***	***0.891***	***0.91***	***0.929***	***0.849***	***0.966***	***0.935***	***0.824***	***0.925***	***0.902***	***0.859***
Sensitivity	***89%***	***77%***	***71%***	***95%***	***93%***	***72%***	***95%***	***89%***	***68%***	***88%***	***81%***	***75%***
Specificity	***89%***	***87%***	***92%***	***71%***	***81%***	***87%***	***89%***	***89%***	***86%***	***85%***	***87%***	***90%***
PPV	***97%***	***96%***	***97%***	***92%***	***94%***	***95%***	***97%***	***96%***	***94%***	***97%***	***97%***	***98%***
Threshold	***0.689*** ***9804***	***0.771*** ***395 7***	***0.813*** ***742 6***	***0.648*** ***9155***	***0.654*** ***5943***	***0.778*** ***8927***	***0.675*** ***1793***	***0.723022***	***0.768*** ***3476***	***0.773*** ***5114***	***0.823*** ***5587***	***0.831*** ***7314***
Number of Probes	***21***	***4***	***8***	***26***	***31***	***6***	***50***	***29***	***4***	***18***	***10***	***6***

Table 5.13 Test model outcomes comparing CB with Cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
AUC	***0.846***	***0.819***	***0.816***	***0.772***	***0.776***	***0.775***	***0.745***	***0.762***	***0.718***	***0.851***	***0.838***	***0.816***
Sensitivity	***89%***	***89%***	***68%***	***59%***	***69%***	***62%***	***72%***	***74%***	***68%***	***85%***	***83%***	***60%***
Specificity	***67%***	***63%***	***83%***	***91%***	***82%***	***82%***	***77%***	***77%***	***68%***	***73%***	***73%***	***93%***
PPV	***91%***	***90%***	***92%***	***96%***	***93%***	***92%***	***91%***	***90%***	***87%***	***94%***	***93%***	***97%***
Threshold	***0.6338898***	***0.6324655***	***0.8325578***	***0.7983429***	***0.7875803***	***0.8177205***	***0.7625193***	***0.7126755***	***0.7689744***	***0.7482762***	***0.7599851***	***0.8785139***
Number of Probes	***21***	***4***	***8***	***26***	***31***	***6***	***50***	***29***	***4***	***18***	***10***	***6***

Table 5.14 Training model outcomes comparing CB with randomly selected Cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
AUC	***0.957***	***0.916***	***0.876***	***0.991***	***0.924***	***0.893***	***0.915***	***0.943***	***0.851***	***0.936***	***0.915***	***0.87***
Sensitivity	***87%***	***94%***	***73%***	***98%***	***94%***	***83%***	***86%***	***94%***	***81%***	***87%***	***85%***	***75%***
Specificity	***94%***	***71%***	***87%***	***94%***	***79%***	***90%***	***87%***	***83%***	***79%***	***90%***	***85%***	***86%***
PPV	***92%***	***92%***	***84%***	***92%***	***82%***	***88%***	***86%***	***84%***	***94%***	***88%***	***89%***	***85%***
Threshold	***0.447312***	***0.3629038***	***0.4870657***	***0.4036753***	***0.4046788***	***0.4623698***	***0.4487793***	***0.3792794***	***0.403381***	***0.4397159***	***0.4691263***	***0.5171299***
Number of Probes	***17***	***9***	***5***	***35***	***19***	***6***	***16***	***20***	***4***	***8***	***7***	***5***

Table 5.15 Test model outcomes comparing CB with randomly selected Cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
AUC	***0.843***	***0.803***	***0.813***	***0.806***	***0.751***	***0.768***	***0.72***	***0.713***	***0.695***	***0.821***	***0.828***	***0.808***
Sensitivity	***78.00%***	***71%***	***69%***	***67%***	***57%***	***64%***	***85%***	***67%***	***74%***	***65%***	***82%***	***63%***
Specificity	***80.00%***	***80%***	***83%***	***91%***	***88%***	***88%***	***56%***	***71%***	***59%***	***90%***	***70%***	***87%***
PPV	***92.00%***	***92%***	***93%***	***96%***	***92%***	***93%***	***85%***	***87%***	***85%***	***96%***	***93%***	***96***
Threshold	***0.4179643***	***0.4595509***	***0.5235882***	***0.4789435***	***0.5916455***	***0.5855912***	***0.3973965***	***0.4905224***	***0.3726214***	***0.508384***	***0.4018513***	***0.5944663***
Number of Probes	***17***	***9***	***5***	***35***	***19***	***6***	***16***	***20***	***4***	***8***	***7***	***5***

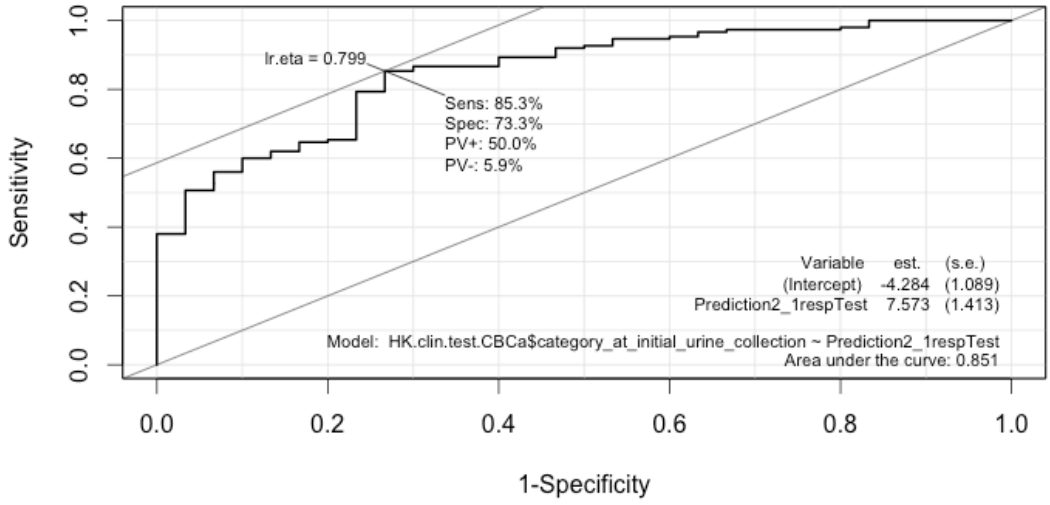


Figure 5.19 ROC curve of top performing model for the prediction of CB vs. Cancer (Low-, Intermediate- and High-risk).

5.7.2 Models to distinguish the extreme categories i.e. CB and high-risk cancer samples

Expression of 98, 43, 39, 39 probes had a significant association with whether a sample was high-risk (H) or there was no evidence for cancer (CBN) in the four processed datasets (*KLK2* ratio, *KLK2* adjusted, *KLK3* adjusted, HK normalised, respectively) (Supplementary Table 13). The top probe was *ERG3*' exons 4-5 ($p = 6.995\text{x}10^{-07}$, logFC = 1.87), *HPN* ($p = 3.767\text{x}10^{-06}$, logFC = 0.24), *HPN* ($p = 1.317\text{x}10^{-05}$, logFC = 0.19), and *ERG3'* exons 4-5 ($p = 1.42\text{x}10^{-06}$, logFC = 0.79), respectively.

Binomial models were built to predict whether a patient was at high risk of cancer (H) or had no evidence for cancer (CB) (Table 5.16, Table 5.17, see Supplementary Table 14, Supplementary Table 15, Supplementary Table 16 and Supplementary Table 17).

The models were decent predictors (test model median AUC = 0.957, IQR = 0.036, training model median AUC = 0.831, IQR = 0.07). In general, the metrics of the models didn't seem to differ much between the different normalisations (slightly lower AUCs in the *KLK3* adjusted data), or the input probe subset. Models with AUC of up to 0.9 were seen in the training sets, and models with AUC of up to 0.8 were seen when applying the models to the test data. Sensitivities in the 90%s and PPVs in the 80%s

were observed on the test data, suggesting these models were capable to distinguishing well between the CB and high-risk cancer samples.

The model built using the adjusted significant probe lists from the *GAPDH* and *RPLP2* normalised data gave a high AUC of 0.897 in the training data (AUC = 0.924 in the test data). This model had high sensitivity (91%), 80% specificity and 83% PPV (ROC - Figure 5.20). The transcripts used to build this model were *PCA3, APOC1, HPN*, *ERG3*' exons 4-5 and *TMPRSS2:ERG*.

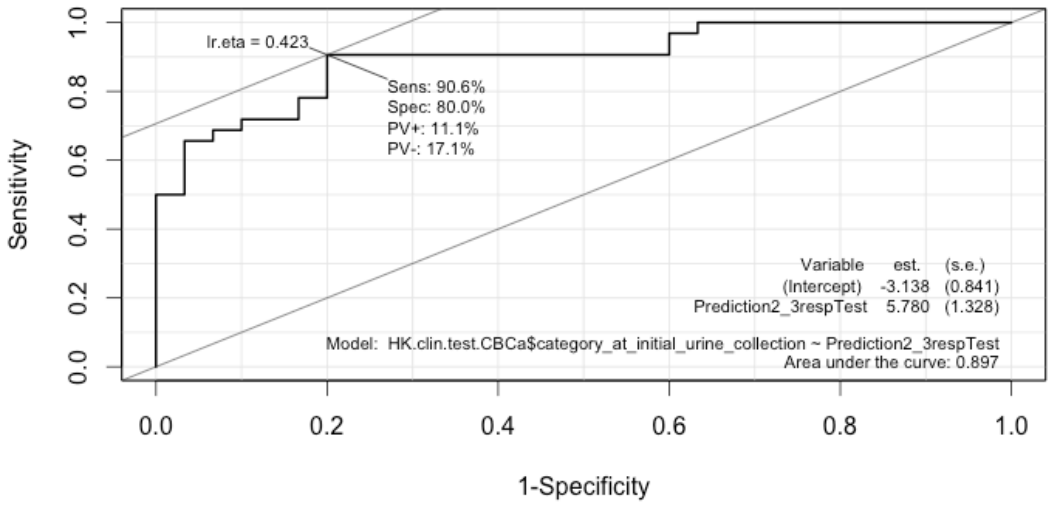


Figure 5.20 ROC curve of the training set for the GAPDH and RPLP2 normalised model built using the 5 significant probes post multiple testing correction.

Table 5.16 Training model outcomes comparing CB with high-risk Cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
AUC	***0.991***	***0.97***	***0.94***	***0.952***	***0.955***	***0.866***	***0.962***	***0.959***	***0.85***	***0.976***	***0.992***	***0.924***
Sensitivity	***100%***	***84%***	***86%***	***86%***	***86%***	***71%***	***91%***	***84%***	***74%***	***94%***	***97%***	***96%***
Specificity	***92%***	***98%***	***90%***	***94%***	***94%***	***91%***	***90%***	***97%***	***90%***	***92%***	***96%***	***77%***
PPV	***93%***	***97%***	***92%***	***92%***	***94%***	***89%***	***91%***	***95%***	***90%***	***94%***	***97%***	***84%***
Threshold	***0.4065925***	***0.5787023***	***0.5417758***	***0.5080049***	***0.503871***	***0.554859***	***0.4488001***	***0.538367***	***0.5328716***	***0.4702714***	***0.5099732***	***0.4034014***
Number of Probes	***26***	***16***	***9***	***19***	***17***	***5***	***21***	***19***	***3***	***13***	***21***	***5***

Table 5.17 Test model outcomes comparing CB with high –risk Cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
AUC	***0.851***	***0.859***	***0.832***	***0.822***	***0.829***	***0.738***	***0.789***	***0.796***	***0.738***	***0.897***	***0.883***	***0.897***
Sensitivity	***88%***	***97%***	***94%***	***91%***	***91%***	***91%***	***97%***	***97%***	***91%***	***88%***	***84%***	***91%***
Specificity	***77%***	***63%***	***60%***	***65%***	***77%***	***59%***	***65%***	***65%***	***59%***	***83%***	***83%***	***80%***
PPV	***80%***	***73%***	***70%***	***71%***	***76%***	***71%***	***70%***	***72%***	***67%***	***82%***	***84%***	***83%***
Threshold	***0.3834388***	***0.2640011***	***0.2297271***	***0.3562361***	***0.4580838***	***0.4023346***	***0.3300849***	***0.2864755***	***0.4023346***	***0.5286275***	***0.4354893***	***0.4889143***
Number of Probes	***26***	***16***	***9***	***19***	***17***	***5***	***21***	***19***	***3***	***13***	***21***	***5***

5.7.3 Models to predict risk categories using trends in expression

Expression of 114, 45, 50, 53 probes had a significant association with increasing risk category (CB->L->I->H) in the four processed datasets (*KLK2* ratio, *KLK2* adjusted, *KLK3* adjusted, HK normalised, respectively) (Supplementary Table 18). The top probe was *ERG3'* exons 4-5 (p = 1.86x10^{-13}), *PCA3* (p = 1.45x10^{-08}), *PCA3* (p = 1.52x10^{-07}), and *ERG3'* exons 4-5 (p = 1.44x10^{-08}) respectively (Figure 5.21).

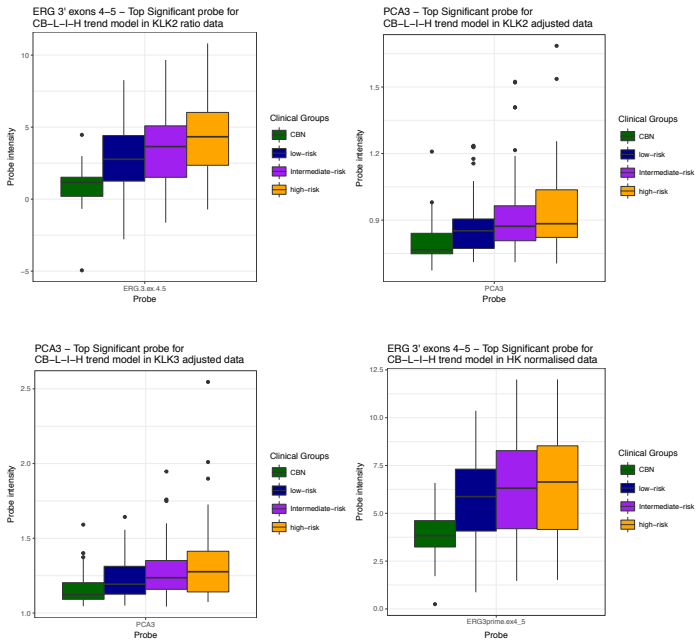


Figure 5.21 Top Significant Probe for CB, low-risk, intermediate-risk and high-risk cancer trend in all four data normalisations.

Multivariate proportional odds models were built to predict CB samples, the low-, intermediate- and high-risk cancer samples (section 2.6.1) (Table 5.18, Table 5.19). The probes involved in each model can be observed (Supplementary Table 19, Supplementary Table 20, Supplementary Table 21, and Supplementary Table 22).

The metrics of the models for the *KLK2* ratio and *KLK2* adjusted data were very similar (median = 0.67015, IQR = 0.06 and median AUC 0.6689, IQR = 0.08). Slightly lower AUCs were observed in the *KLK3* adjusted data (median AUC = 0.669, IQR = 0.1), and slightly higher AUCs were observed in the *GAPDH* and *RPLP2* normalised data (median AUC = 0.73385, IQR = 0.05). The average model metrics for the test data were poorer than with previous clinical questions (median AUC = 0.65005, IQR = 0.05). The

sensitivity of all of the models were fairly low (median= 29%, IQR=0.45), whilst specificity fairly high (median= 91% IQR=0.23). This suggested that separating between the different risk categories of cancer can be difficult.

The model built using the *GAPDH* and *RPLP2* normalised data and only the probes still significant post multiple testing correction has the highest AUC = 0.7088. The probes used to build this model were *APOC1, DPP4, ERG* 3' exons 4-5, *ERG* 3' exons 6-7, *GABARAPL2, HOXC6, HPN, ITGBL1, KLK4, MYOF, PCA3, TDRD*, and *TMPRSS2:ERG* (Figure 5.22). The Sensitivities of this model ranged from 9%-79% and the specificities ranged from 46%-95%.

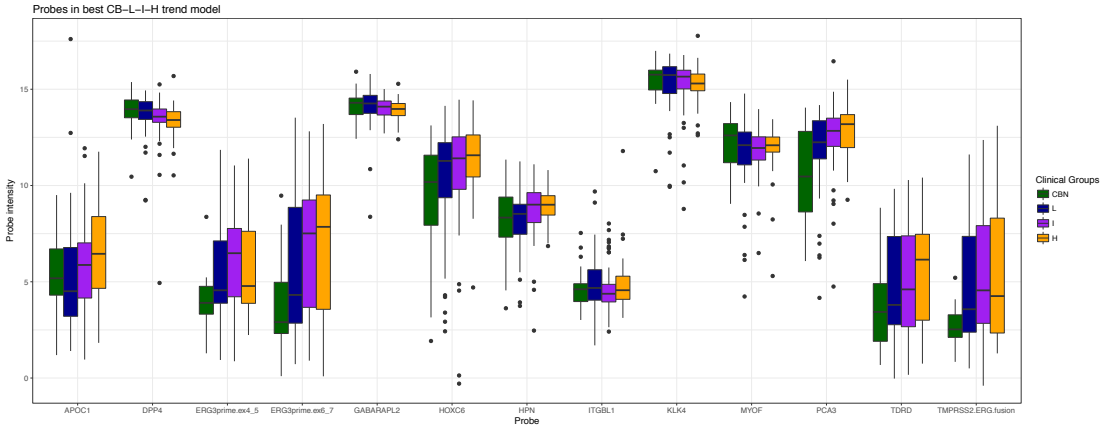


Figure 5.22 Boxplot showing the expression level of each transcript featured in the CB-L-I-H model built using the multiple tested correction significant probes from the GAPDH and RPLP2 normalised data. This model showed the best test data AUC (0.7008).

Table 5.18 Training model outcomes comparing CB, low-, intermediate- and high- risk cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
Accuracy	***0.5112***	***0.4581***	***0.4413***	***0.5412***	***0.522***	***0.4643***	***0.6749***	***0.576***	***0.5124***	***0.5***	***0.4944***	***0.5112***
AUC	***0.7663***	***0.6757***	***0.6196***	***0.7802***	***0.7469***	***0.6856***	***0.8146***	***0.7606***	***0.6929***	***0.7587***	***0.7728***	***0.7608***
Sensitivity:												
CB	***52%***	***38%***	***27%***	***45%***	***48%***	***36%***	***84%***	***74%***	***60%***	***50%***	***56%***	***54%***
L	***19%***	***8%***	***4%***	***26%***	***24%***	***12%***	***0%***	***0%***	***0%***	***15%***	***19%***	***25%***
I	***84%***	***88%***	***94%***	***84%***	***81%***	***83%***	***89%***	***80%***	***75%***	***84%***	***74%***	***76%***
H	***25%***	***10%***	***7%***	***37%***	***32%***	***25%***	***59%***	***41%***	***36%***	***25%***	***34%***	***31%***
Specificity:												
CB	***97%***	***96%***	***96%***	***97%***	***96%***	***94%***	***95%***	***91%***	***85%***	***97%***	***95%***	***96%***
L	***87%***	***91%***	***96%***	***87%***	***89%***	***92%***	***100%***	***100%***	***100%***	***86%***	***84%***	***83%***
I	***42%***	***26%***	***18%***	***47%***	***45%***	***36%***	***59%***	***53%***	***48%***	***40%***	***47%***	***50%***
H	***98%***	***98%***	***100%***	***97%***	***97%***	***96%***	***96%***	***92%***	***93%***	***98%***	***97%***	***97%***
PPV:												
CB	***73%***	***65%***	***54%***	***74%***	***68%***	***54%***	***81%***	***69%***	***53%***	***74%***	***67%***	***68%***
L	***32%***	***23%***	***25%***	***44%***	***42%***	***32%***	***NA***	***NA***	***NA***	***27%***	***28%***	***34%***
I	***49%***	***44%***	***44%***	***51%***	***49%***	***46%***	***58%***	***52%***	***48%***	***49%***	***48%***	***51%***
H	***78%***	***58%***	***83%***	***72%***	***72%***	***58%***	***82%***	***63%***	***63%***	***78%***	***73%***	***69%***
Number of Probes	***36***	***13***	***5***	***12***	***37***	***14***	***78***	***39***	***12***	***37***	***34***	***13***

Table 5.19 Test model outcomes comparing CB, low-, intermediate- and high- risk cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction.

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
Accuracy	***0.4611***	***0.45***	***0.4278***	***0.4444***	***0.4222***	***0.3944***	***0.3944***	***0.222***	***0.4056***	***0.4716***	***0.4659***	***0.4773***
AUC	***0.6894***	***0.6646***	***0.6115***	***0.6479***	***0.6522***	***0.6273***	***0.6372***	***0.4993***	***0.6468***	***0.6791***	***0.709***	***0.7088***
Sensitivity:												
CB	***37%***	***37%***	***33%***	***30%***	***43%***	***27%***	***47%***	***90%***	***53%***	***35%***	***35%***	***42%***
L	***28%***	***15%***	***9%***	***28%***	***15%***	***7%***	***0%***	***0%***	***0%***	***28%***	***35%***	***28%***
I	***79%***	***85%***	***88%***	***75%***	***76%***	***78%***	***68%***	***18%***	***72%***	***82%***	***76%***	***79%***
H	***6%***	***6%***	***0%***	***13%***	***31%***	***13%***	***25%***	***0%***	***16%***	***6%***	***6%***	***9%***
Specificity:												
CB	***97%***	***94%***	***95%***	***97%***	***91%***	***89%***	***85%***	***15%***	***81%***	***96%***	***94%***	***93%***
L	***84%***	***90%***	***93%***	***86%***	***90%***	***92%***	***100%***	***100%***	***100%***	***85%***	***83%***	***86%***
I	***42%***	***33%***	***23%***	***39%***	***39%***	***31%***	***42%***	***90%***	***40%***	***39%***	***46%***	***46%***
H	***95%***	***97%***	***97%***	***93%***	***93%***	***95%***	***84%***	***99%***	***91%***	***97%***	***95%***	***95%***
PPV:												
CB	***69%***	***55%***	***59%***	***64%***	***50%***	***33%***	***38%***	***18%***	***36%***	***60%***	***50%***	***50%***
L	***37%***	***35%***	***31%***	***41%***	***33%***	***21%***	***NA***	***NA***	***NA***	***41%***	***42%***	***42%***
I	***48%***	***46%***	***43%***	***45%***	***45%***	***43%***	***44%***	***54%***	***44%***	***48%***	***50%***	***50%***
H	***22%***	***29%***	***0%***	***29%***	***8%***	***33%***	***26%***	***0%***	***28%***	***29%***	***22%***	***30%***
Number of Probes	***36***	***13***	***5***	***12***	***37***	***14***	***78***	***39***	***12***	***37***	***34***	***13***

5.7.4 Models to predict patient type using trends in expression

Expression of 152, 57, 56, 45 probes had a significant association with increasing severity of disease type i.e. no evidence for cancer (CB), organ confined cancer (L, I, & H) and metastatic disease (A) in the four processed datasets (*KLK2* ratio, *KLK2* adjusted, *KLK3* adjusted, HK normalised respectively) (Supplementary Table 23). The top probe was *HOXC6* (p = 5.19x10^{-10}), *UPK2* (p = 2.91x10^{-08}), *UPK2* (p = 2.4x10^{-08}), and *HOXC6* (p = 3.39x10^{-06}) respectively.

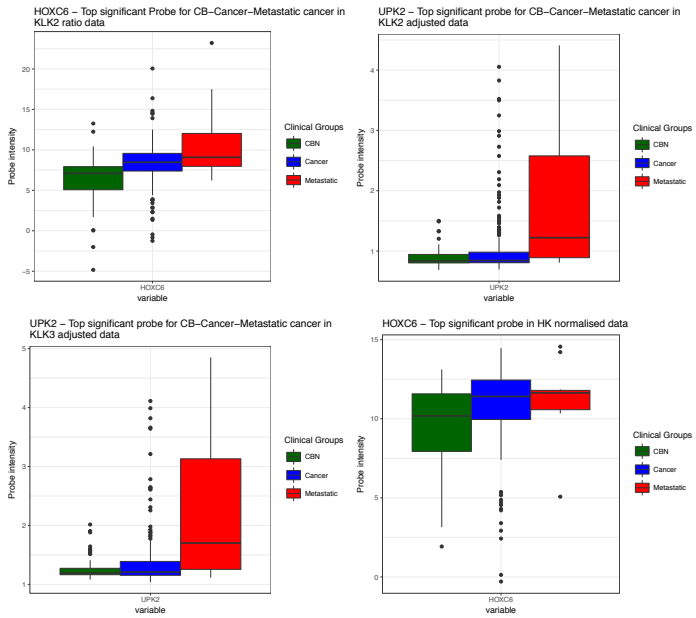


Figure 5.23 Top Significant Probe for CB, Cancer, Metastatic trend in all four data normalisations.

Multivariate proportional odds models were built to predict clinical categories (section 2.6.1), no evidence for cancer (CB), organ confined cancer (L, I, & H) and metastatic disease (Table 5.20, Table 5.21). The probes involved in each model can be observed (Supplementary Table 24, Supplementary Table 25, Supplementary Table 26 and Supplementary Table 27).

Low AUCs were observed across all inputs and data sets (median AUC = 0.57365, IQR = 0.08). The *GAPDH* and *RPLP2* normalised data showed slightly higher AUCs (median AUC = 0.6388, IQR = 0.997). The sensitivity of the sample categories in all of the models were fairly low (median = 18%, IQR = 87%). Whilst the specificity is fairly

high but not uniformly across the models (median = 98%, IQR = 75%). Inclusion of the advanced samples could be a reason for this poor model quality. Advanced tumours tend to be firm to the touch and it is thought that upon compression tend to release fewer cells into the urine (section 1.3.4.2). This is further supported by the lower levels of prostate specific transcripts observed in advanced samples (section 3.4) and *UPK2* (the bladder specific marker) is one of the most significant differential probes comparing these samples.

Again, the model with the best AUC (0.6469) is from the *GAPDH* and *RPLP2* (HK) normalised data. The model was built using the significant probes (*MARCH5, AMACR, APOC1, CACNA1D, CP, DLX1, ERG* 3' exons 4-5, *ERG* 3' exons 6-7, *GABARAPL2, GCNT1, GJB1, HOXC6, IFT57, ITGBL1, KLK2, KLK4, MCTP1, Met, MIR4435_1HG, MSMB, PALM3, PCA3, PTN, SLC12A1, SSTR1, STOM, SULF2, TDRD, TMCC1, TMEM45B, ZNF577*). The model's sensitivity ranged from 17% - 93% and it's specificity ranged from 26%- 98%.

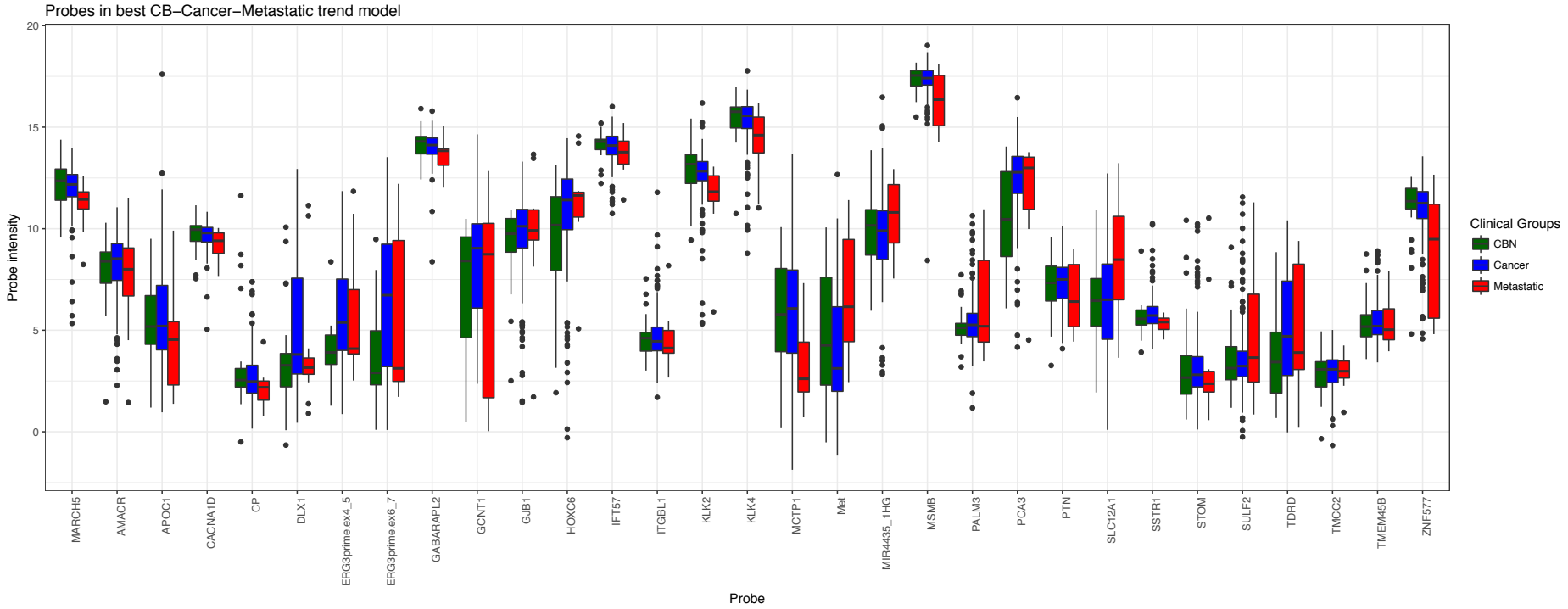


Figure 5.24 Boxplot showing the expression level of each transcript featured in the CB-Cancer-Metastatic cancer model built using the significant probes from the *GAPDH* and *RPLP2* normalised data. This model showed the best test data AUC (0.6469).

Table 5.20 Training model outcomes comparing CB, Cancer (low-, intermediate- and high- risk) and metastatic (A) cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
Accuracy	***0.8136***	***0.811***	***0.811***	***0.8114***	***0.8372***	***0.7855***	***0.781***	***0.8072***	***0.7353***	***0.8819***	***0.8504***	***0.8005***
AUC	***0.5554***	***0.5495***	***0.5495***	***0.5878***	***0.6566***	***0.5201***	***0.6267***	***0.7126***	***0.5913***	***0.7375***	***0.6685***	***0.541***
Sensitivity:												
CB	***4%***	***2%***	***2%***	***5%***	***24%***	***2%***	***23%***	***35%***	***6%***	***44%***	***37%***	***4%***
Cancer	***100%***	***100%***	***100%***	***100%***	***99%***	***99%***	***100%***	***98%***	***98%***	***100%***	***98%***	***98%***
Metastatic	***13%***	***13%***	***13%***	***22%***	***26%***	***4%***	***17%***	***35%***	***22%***	***35%***	***17%***	***9%***
Specificity:												
CB	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***99%***	***100%***	***100%***	***99%***	***99%***
Cancer	***7%***	***5%***	***5%***	***10%***	***25%***	***2%***	***21%***	***35%***	***11%***	***41%***	***31%***	***5%***
Metastatic	***100%***	***100%***	***100%***	***100%***	***100%***	***99%***	***100%***	***99%***	***99%***	***100%***	***99%***	***100%***
PPV:												
CB	***100%***	***100%***	***100%***	***100%***	***93%***	***100%***	***100%***	***92%***	***80%***	***96%***	***86%***	***33%***
Cancer	***81%***	***81%***	***81%***	***81%***	***83%***	***79%***	***77%***	***80%***	***74%***	***87%***	***85%***	***81%***
Metastatic	***75%***	***75%***	***75%***	***100%***	***86%***	***20%***	***100%***	***80%***	***56%***	***100%***	***67%***	***67%***
Number of Probes	***11***	***7***	***8***	***39***	***39***	***11***	***35***	***39***	***9***	***69***	***31***	***9***

Table 5.21 Test model outcomes comparing CB, Cancer (low-, intermediate- and high- risk) and metastatic (A) cancer samples for the four different normalisations of data. Three input probe sets were used: all probes, those significant via GLM testing and those significant post - multiple testing correction

	KLK2 ratio			*KLK2 Adjusted*			*KLK3 Adjusted*			*GAPDH and RPLP2 normalised*		
	All probes	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***	***All probes***	***Significant probes***	***Adjusted Significant Probes***
Accuracy	***0.7865***	***0.7812***	***0.7812***	***0.7917***	***0.776***	***0.776***	***0.8021***	***0.3542***	***0.7656***	***0.7819***	***0.7926***	***0.7819***
AUC	***0.5111***	***0.5***	***0.5333***	***0.5595***	***0.5799***	***0.5657***	***0.5911***	***0.5778***	***0.5695***	***0.6307***	***0.6469***	***0.5***
Sensitivity:												
CB	***3%***	***0%***	***0%***	***10%***	***17%***	***0%***	***20%***	***90%***	***13%***	***27%***	***31%***	***0%***
Cancer	***100%***	***100%***	***100%***	***99%***	***95%***	***99%***	***98%***	***28%***	***95%***	***91%***	***93%***	***98%***
Metastatic	***0%***	***0%***	***0%***	***8%***	***8%***	***17%***	***8%***	***19%***	***8%***	***25%***	***17%***	***0%***
Specificity:												
CB	***100%***	***100%***	***100%***	***100%***	***99%***	***100%***	***99%***	***28%***	***96%***	***93%***	***96%***	***98%***
Cancer	***24%***	***0%***	***0%***	***10%***	***14%***	***2%***	***17%***	***83%***	***12%***	***29%***	***26%***	***0%***
Metastatic	***100%***	***100%***	***100%***	***99%***	***97%***	***99%***	***99%***	***100%***	***99%***	***98%***	***98%***	***100%***
PPV:												
CB	***100%***	***100%***	***100%***	***100%***	***71%***	***NA***	***86%***	***19%***	***40%***	***39%***	***53%***	***0%***
Cancer	***79%***	***78%***	***78%***	***80%***	***80%***	***78%***	***81%***	***85%***	***79%***	***84%***	***83%***	***79%***
Metastatic	***NA***	***NA***	***NA***	***33%***	***17%***	***3%***	***33%***	***NA***	***33%***	***50%***	***33%***	***NA***
Number of Probes	***11***	***7***	***8***	***39***	***39***	***11***	***35***	***39***	***9***	***69***	***31***	***9***

5.7.5 Conclusions

Use of the housekeeping probes *GAPDH* and *RPLP2* provided normalised data that produced good prediction models (this data provided the best AUC for prediction models for all four clinical questions (Table 5.11)). Data was otherwise treated similarly to NanoString1 (chapter 3). Identification of *GAPDH* and *RPLP2* as housekeepers to normalise urinary EV RNA derived NanoString data increased the robustness of my prediction models.

All models were built using a training set that included samples from all four centres, and particularly the binomial tests were robust (high AUCs). The models therefore, can predict cancer from samples with no evidence of cancer (CB) regardless of sample origin.

Optimal models built from the expression of 167 markers for risk stratification and detection of cancer were found using the *GAPDH* and *RPLP2* normalised data, however, input lists varied from all probes, significant probes (identified by polr) and adjusted significant probes (Benjamin Hochberg multiple testing correction).

The Prostate Cancer Prevention Trial risk calculator (PCPTrc) and the Prostate Cancer Prevention Trial high-grade risk calculator (PCPThg) are logistic regression models, which incorporate PSA level, PSA velocity, DRE result, previous biopsy results, age at biopsy, race and family history of PCa[212]. These models have been combined with urinary (whole cell) *TMPRSS2:ERG* and urine *PCA3* levels to improve model AUC: PCPTrc alone had an AUC of 0.639, whilst inclusion of urinary *TMPRSS2:ERG* and *PCA3* improved the AUC to 0.762. Urinary *TMPRSS2:ERG* and *PCA3* also improved the predictive power of serum PSA (AUC = 0.651 increased to AUC = 0.772)[213]. Similarly, PCA3, which is used in the PCA3 test, which was the first commercialy available urinary test for PCa, is capable of predicting cancer from non-cancer samples (AUC = 0.98)[214]. The models achieved similar AUCs, when predicting cancer (L, I and H) from samples with no evidence of cancer (CB): AUC = 0.851 for the best model and

median AUC = 0.8045. I found that in EV harvested material the *ERG*3' exons 4-5 and *PCA3* data were the most highly differentiating between cancer and samples with no evidence of cancer. The probes used in the top model were *TMPRSS2:ERG, ERG3'* exons 4-5, *APOC1, ISX, SLC12A1, HOXC6, MCTP1, TDRD, PDLIM5, CD10, GABARAPL2, PTN, AR* exon 9, *PPP1R12B, CP, MXI1,* and *KLK4.*

The high-grade predictor also benefitted from the addition of urinary *TMPRSS2:ERG* and *PCA3* data (AUC = 0.707 increased to AUC = 0.779)[213]. A second high-grade predictor was produced by Van Neste *et al.*, which used whole urine mRNA levels of *HOXC4, HOXC6, TDRD1, DLX1* and *PCA3* (with *KLK3* as a reference) alongside clinical factors (including PSA density, previous biopsies, PSA, age and family history[215]. This model reached an overall AUC of 0.9 in their validation set. The high-risk (H) Vs. no evidence of cancer (CB) models also achieved high AUCs (top AUC = 0.897, median AUC = 0.831). The model also used *PCA3* and *TMPRSS2:ERG* levels, along with *APOC1*, *HPN* and *ERG3'* exons 4-5 (from EV harvested RNA). The top most significant probes when comparing high-risk cancer with samples with no evidence of cancer was *HPN* and *ERG* 3' exons 4-5.

The ExoDx Prostate IntelliScore urine exosome assay uses *ERG* and *PCA3* data normalised using *SPDEF* combined with clinical factors (including PSA level, age, race and family history) to predict between Gleason 6 and Gleason 7 PCa with AUC = 0.73[216]. The models to predict between different risk categories (CB->L->I->H) had similar AUCs (median = 0.67015, highest AUC = 0.709). The model was built using *APOC1, DPP4, ERG* 3' exons 4-5, *ERG* 3' exons 6-7, *GABARAPL2, HOXC6, HPN, ITGBL1, KLK4, MYOF, PCA3, TDRD*, and *TMPRSS2:ERG.* It is not surprising that *PCA3* and *ERG3*' exons 4-5 were also the most highly significant in all four data normalisations.

I have shown that EV derived material from multiple centres can be quantified by NanoString to produce models that can predict cancer presence and aggression without biopsy. However, much greater numbers and model refinements (such as RF etc.)

would be needed to strengthen the models into a test that could be used in the clinic. A multivariate regression with the combination of RNA signatures and clinical factors should also be investigated.

6

Expression Profile of the Cell Sediment Urine Fraction

6.1 Summary

In this chapter, I compared the transcriptome profiles of two urine fractions from prostate cancer patients and controls, and examined whether the transcriptomes from cell sediment were better than EV transcriptomes for PCa diagnosis. I found that the cell sediments have a very different transcriptome profile to the EV fractions, which is similar to what was found in renal cancer[217]. Transcripts found by microarray analysis to be significantly more abundant in the EV fraction compared to the cell sediment were more commonly expressed in prostatic tissue and also had more known associations with prostate cancer. This suggested that the majority of RNA within the extracellular vesicle fraction comes from prostatic tissue, both normal and cancerous. These analyses support the hypothesis that EVs are a better fraction to study for biomarkers in prostate cancer.

Analysis of cell sediment NanoString data to identify transcripts that could be used diagnostically to identify D'Amico clinically categories show common transcripts being selected by both Lasso and Random forest analyses when i) different input subsets of transcripts were used, and ii) when data was normalised with different control genes. Different probes for *ERG* 3' sequences were identified in cell (*ERG* probe targeting exons 6-7) and EV models (*ERG* probe targeting exons 4-5). *HOXC6* and *TDRD* were found in both cell and EV models. Interestingly, *PCA3* and *TMPRSS2:ERG* were found in EV models and not cell models. This supports other work that the majority of PCa RNA content in whole urine is originating from EVs and not whole cells.

6.2 Introduction

NanoString technology was applied to cell and extracellular vesicle (EV) fractions of urine from prostate cancer patients to form the NanoString 2 data set. Urine samples were divided into two fractions by centrifugation: i) cell sediment and ii) supernatant containing extracellular vesicles (section 2.1.2). In this chapter, analysis of the cell fraction will be completed. The investigation of the EV fraction can be found in Chapter 5.

6.2.1 The Research Gap

Since the production of the PCA3 test[214], urine has been investigated for PCa biomarkers. Whole urine and cell sediment are commonly used and many models have been developed or built upon to include urinary expression of transcripts as biomarkers[213,215]. However, little work has been done on the EV fraction. The EV fraction has been identified to be a useful source of biomarkers in renal cancer[217] and PCa associated transcripts have been quantified from PCa urine EVs[218]. No comparisons between transcript expression levels in EV fractions and cell fractions

have been published. It is therefore, unknown which may be the better source of PCa biomarkers.

6.2.2 The Aims

In the first part of this chapter, I will examine the NanoString data for differences in the expression profiles of the cell sediment between different clinical categories and try to construct models to predict clinical categories: These comparisons include comparing i) CB (no evidence of cancer) samples with D'Amico cancer risk groups: Low, Intermediate and High, and ii) CB vs high-risk cancer samples. Two trends will be investigated, CB, Low-, Intermediate-, High-risk cancers, ordered as such and CB, cancer and metastatic samples; ordered as such. In each comparison I have used two methods of analysis (logistic regression analysis and Mann Whitney U test), and will compare and contrast the selected gene transcripts from each. These investigations have already been presented for the extracellular vesicle fraction (chapter 5).

In the second part of this chapter, I will compare and contrast the matched EV and cell sediment fraction data from microarray and NanoString analyses. Other studies have observed that the transcriptomes of urinary extracellular vesicles and whole urine are different in renal cancer[217]. I will identify transcripts that are significantly differentially expressed between the cell sediment and EVs in both NanoString data and microarray data.

6.2.3 The Data

The cell and EV fractions were analysed in 95 samples from a range of clinical categories (Table 6.1) based on the D'Amico classification using 167 NanoString probes. Three of these samples were taken pre-DRE, and as shown previously, these samples are not fully comparable with those obtained post-DRE and were not used in this chapter. These data were normalised with the spiked in positive controls as per the

NanoString manual (section 2.3.1) and $\log_2$ transformed (section 2.3.3) to produce the baseline normalised data. Investigations into use of housekeeper transcripts were completed and data was normalised (section 2.3.2) using *RPLP2* and *TWIST1*. These probes showed no association to clinical categories ($p > 0.05$ ANOVA-Tukey test) and were heavily correlated ($p < 2.2\text{x}10^{-16}$, $r = 0.83$). Then similarly to the EV fraction *KLK2* ratio normalisation (section 2.1.1) was also performed. PCA plots (not shown) were used to visualise the *RPLP2* and *TWIST1* normalised ("HK normalised data") and the *KLK2* ratio data. There were two outlier samples; M_86_1 (an Intermediate – risk sample) and M_147_1 (a CB sample), which were removed from the HK normalised data. M_147_1 was also removed from the *KLK2* ratio normalised data, as forty-six of the one hundred and sixty-seven values for M_147_1 were zero. The values for M_88_5 looked normal in the KLK2 ratio data. Four clinical questions (Table 6.2) were investigated in the data and prediction models were produced accordingly. Due to limited numbers of samples, the data was not divided into test and training data.

Table 6.1 Clinical breakdown of cell sediment fraction samples subjected to NanoString (within the second NanoString set). Twelve samples were CB (no evidence of cancer). Thirty raised PSA samples were negative for PCa on biopsy, but other abnormalities were found such as, HGPIN, prostatitis and atypia. Forty-six had localised cancer on TRUS biopsy of which four were D'Amico graded as Low risk, twenty-eight Intermediate risk and fourteen High-risk. Four samples had shown signs of metastasis.

	CB	*Abnormal*	*L*	*I*	*H*	*A*	*Total*
Number of Samples	***12***	***30***	***4***	***28***	***14***	***4***	***92***
Percentage	***13%***	***33%***	***4%***	***30%***	***15%***	***4%***	***100%***
Median Age	***65***	***66***	***63.5***	***71***	***67.5***	***82***	***68***
Median PSA	***0.9***	***7.9***	***6.4***	***7.8***	***16.8***	***377***	***8.1***

Table 6.2 Clinical predictive models built using the cell dataset.

Model	*Samples*	*Model type*
CB vs. Cancer	***Clinically benign samples Vs low-, intermediate-, and high-risk cancer samples grouped together***	***Binary***
CB vs. High risk cancer	***Clinically benign Vs. high-risk cancer (extreme ends of no evidence of cancer and***	***Binary***

	those with higher grade)	
CB, low-, intermediate-, and high-risk trend	***Each sample category is a separate group and ordered***	***Ordinal***
CB, cancer, metastatic cancer trend	***Clinically benign samples, with low-, intermediate-, and high-risk cancer samples grouped together, and metastatic cancer samples in groups ordered by severity***	***Ordinal***

6.3 Models predicting presence of cancer CB and cancer (L, I, H) samples using cell sediment data

6.3.1.1 Differentially expressed transcripts

Expression of 85, 28, and 24 transcripts had a significant association (via logistic regression section 2.6.1) with whether a sample had no evidence for cancer (CB) or not (L, I, H) in the three processed datasets (the baseline data, *KLK2* ratio, and HK normalised, respectively) (Supplementary Table 31). Only *MCTP1* remained significant post multiple testing correction (adjusted $p = 0.04$) in the baseline data and none remained significant in the *KLK2* ratio and HK normalised data. The top significant probe in these datasets was *ERG* 3' exons 6-7 ($p = 0.001$) and *NAALADL2* ($p = 3.33\text{x}10^{-05}$), respectively.

Expression of 94, 33, and 56 transcripts had a significant association (via Mann Whitney U (MWU) testing, section 2.4.1) with whether a sample had no evidence for cancer (CB) or not (L, I, H) in the three processed datasets (the baseline data, *KLK2* ratio, and HK normalised, respectively) (Supplementary Table 32). The top significant probes identified by MWU were *SULF2* ($p = 9.18\text{x}10^{-06}$), *PCA3* ($p = 3.72\text{x}10^{-05}$) and *SPINK1* ($p = 3.72\text{x}10^{-05}$), respectively.

Between the two tests 79, 20 and 26 transcripts were common between the two methods suggesting a good level of robustness. The top ten transcripts with the biggest log2 fold

change for the baseline data, *KLK2* ratio, and HK normalised data are shown (Table 6.3, Table 6.4, Table 6.5, respectively).

Table 6.3 Top ten transcripts with biggest log2 fold change in the baseline normalised data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***0.0002***	***0.024***	***0.0014***	***0.2049***	***1.64***
ERG3' exons 6-7	***$2.84x10^{-07}$***	***$4.74x10^{-05}$***	***0.0008***	***0.128***	***1.38***
TMPRSS2:ERG	***$4.52x10^{-05}$***	***0.0069***	***0.0013***	***0.1979***	***1.31***
SLC43A1	***0.0003***	***0.0406***	***0.0019***	***0.2745***	***1.17***
CLIC2	***$2.66x10^{-05}$***	***0.0042***	***0.001***	***0.1645***	***1.05***
B4GALNT4	***$3.38x10^{-05}$***	***0.0053***	***0.0012***	***0.1807***	***1.04***
CADPS	***$1.37x10^{-05}$***	***0.0022***	***0.0004***	***0.0682***	***1.04***
CKAP2L	***0.0116***	***1***	***0.0033***	***0.4318***	***1.01***
HPN	***$7.04x10^{-05}$***	***0.0103***	***0.0006***	***0.1041***	***0.97***
LASS1	***0.0002***	***0.022***	***0.0011***	***0.1703***	***0.97***

Table 6.4 Top ten transcripts with biggest log2 fold change in the *KLK2* ratio data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***$6.80x10^{-05}$***	***0.01***	***0.004***	***0.63***	***0.21***
ERG3' exons 6-7	***$7.80x10^{-05}$***	***0.01***	***0.001***	***0.24***	***0.18***
TDRD	***0.0004***	***0.06***	***0.004***	***0.72***	***0.18***
SLC43A1	***0.002***	***0.32***			***0.17***
CADPS	***0.004***	***0.67***	***0.01***	***1***	***0.16***
ERG5'	***0.01***	***0.99***			***0.15***
B4GALNT4	***0.01***	***0.87***			***0.14***
SLC12A1	***0.003***	***0.54***	***0.03***	***1***	***0.13***
TMCC2	***0.05***	***0.99***	***0.05***	***1***	***0.13***
TMPRSS2:ERG	***0.001***	***0.17***	***0.01***	***1***	***0.13***

Table 6.5 Top ten transcripts with biggest log2 fold change in the HK normalised data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***0.0002***	***0.0374***	***0.0019***	***0.3087***	***1.5***
ERG3' exons 6-7	***0.0006***	***0.1045***	***0.0228***	***0.9861***	***1.1***
TMPRSS2:ERG	***0.0036***	***0.5527***	***0.0069***	***0.9861***	***1.1***
CP	***0.0146***	***0.9924***	***0.0109***	***0.9861***	***-1***
TDRD	***0.001***	***0.153***	***0.0105***	***0.9861***	***0.9***
NAALADL2	***$3.33x10^{-05}$***	***0.0056***	***0.0012***	***0.2012***	***-0.8***
SLC43A1	***0.0005***	***0.0895***	***0.0168***	***0.9861***	***0.8***
ST6GALNAC1	***0.0008***	***0.1311***	***0.0238***	***0.9861***	***-0.8***
SPINK1	***$7.80x10^{-05}$***	***0.0129***			***-0.7***
UPK2	***0.0007***	***0.1128***	***0.0026***	***0.4313***	***-0.7***

6.3.1.2 Models and gene selection

A number of different transcript subsets were input to Lasso for probe shrinkage and selection, these included i) all of the transcripts (n = 167). Transcripts identified as having significantly different expression between cancer and CB using ii) Mann Whitney U (n = 94, n = 33 and n = 56) and iii) logistic regression (n = 85, n = 28 and n = 24), separately, and iv) transcripts common to both those identified by Mann Whitney U and logistic regression (n = 79, n = 20 and n = 26) for each of the three normalisations (the baseline data, *KLK2* ratio, and HK normalised), respectively. The AUC, sensitivity and specificity of each model on the same data was collected (Table 6.6) and transcript lists (Table 6.7, Table 6.8, Table 6.9) and boxplots of the Lasso selected probes were produced (Supplementary Figure 13, Supplementary Figure 14 and Supplementary Figure 15).

Table 6.6 AUC, Sensitivity and Specificity of models to predict CB vs. Cancer (L, I, H) in different data normalisations of cell NanoString data.

	HK normalised				*KLK2 ratio*				*Baseline*			
	All Transcripts	***MWU***	***glm***	***Both***	***All Transcripts***	***MWU***	***glm***	***Both***	***All Transcripts***	***MWU***	***glm***	***Both***
AUC	***0.989***	***0.998***	***0.998***	***0.998***	***0.996***	***0.998***	***0.993***	***0.995***	***0.998***	***1***	***1***	***1***
Sensitivity	***100%***	***98%***	***98%***	***98%***	***98%***	***98%***	***94%***	***96%***	***98%***	***100%***	***100%***	***100%***
Specificity	***92%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***
Number of Probes	***8****	***8****	***8****	***8****	***9***	***7***	***7****	***7****	***13***	***17***	***14*****	***14*****

*, ** and *** selected probes are identical in model.

Table 6.7 Beta values of individual transcripts within models suggested by Lasso using different input transcripts for the baseline normalised data.

Transcript	*Beta – All Transcripts*	*Beta - MWU*	*Beta - glm*	*Beta - both*
ACTR5		***0.572***	***0.206***	***0.226***
APOC1	***0.045***	***0.022***	***0.054***	***0.052***
ARHGEF25		***-0.176***	***0.396***	
CADPS	***0.273***	***0.481***		***0.399***
CAMKK2	***0.055***			
ERG 3' exons 6-7	***0.082***	***0.203***	***0.135***	***0.137***
EN2		***0.32***	***0.146***	***0.164***
HIST1H2BG	***0.006***		***0.015***	***0.013***
HOXC6	***0.096***	***0.138***	***0.114***	***0.116***
IGFBP3		***-0.148***		
LASS1	***0.115***	***0.314***	***0.26***	***0.263***
MCTP1	***0.159***			
MMP25	***0.042***	***0.3470***	***0.219***	***0.224***
MMP26	***-0.124***	***-0.137***		
NAALADL2		***-0.515***	***-0.356***	***-0.371***
PCA3	***0.019***	***0.084***	***0.076***	***0.078***
RIOK3	***0.095***	***0.0290***	***0.012***	***0.003***
SPINK1	***-0.05***	***-0.0220***	***-0.056***	***-0.058***
SLC12A1		***0.1020***		
TDRD		***0.1260***	***0.041***	***0.044***

Table 6.8 Beta values of individual transcripts within models suggested by Lasso using different input transcripts for the *KLK2* ratio data.

Transcript	*Beta – All transcripts*	*Beta - MWU*	*Beta - glm*	*Beta - Both*
CADPS	***0.075***	***0.1795***	***0.1052***	***0.1691***
CKAP2L	***0.1992***	***0.2766***	***0.1894***	***0.2467***
EN2	***0.0828***			
ERG 3' exons 6-7	***0.7197***	***0.7699***	***0.7411***	***0.8396***
HOXC6	***0.3855***	***0.6533***	***0.3772***	***0.5718***
MFSD2A	***0.0104***			
NAALADL2	***-1.456***	***-2.084***	***-1.3254***	***-1.994***
SFRP4	***0.1328***	***0.228***		
SIM2 long			***0.0251***	***0.2771***
TDRD	***0.157***	***0.3875***	***0.1587***	***0.258***

Table 6.9 Beta values of individual transcripts within models suggested by Lasso using different input transcripts for the HK normalised data.

Transcript	*Beta – All transcripts*	*Beta - MWU*	*Beta - glm*	*Beta - Both*
CADPS	***0.1096***	***0.1433***	***0.1971***	***0.2285***
CLIC2	***0.1062***	***0.1143***	***0.1241***	***0.128***
ERG 3' exons 6-7	***0.0716***	***0.0819***	***0.0967***	***0.1051***
HOXC6	***0.1962***	***0.2044***	***0.2182***	***0.2272***
NAALADL2	***-0.287***	***-0.315***	***-0.353***	***-0.3719***
SIM2 long	***0.0411***	***0.0536***	***0.0602***	***0.0575***
TDRD	***0.0502***	***0.064***	***0.088***	***0.1031***
UPK2	***-0.081***	***-0.073***	***-0.061***	***-0.055***

Random forest was also applied to i) all transcripts, ii) significant transcripts identified by MWU and iii) significant transcripts identified by glm for the three different normalisations (the baseline data, *KLK2* ratio, and HK normalised), respectively (Supplementary Table 34, Supplementary Table 35 and Supplementary Table 36). The random forest model with the least error (Table 6.11) was built using the glm identified significant probes from the KLK2 ratio data (the mean square of residuals = 0.088). *ERG* 3' exon 6-7 was in the top 5 transcripts in 8/9 random forests, whilst *APOC1* was in the top 5 transcripts in 6/9 random forests. *HOXC6, CADPS, RIOK3, TMPRSS2:ERG, SLC12A1* and *SPINK1* occur in 3/9 random forests (Table 6.10).

Table 6.10 Frequency of transcripts in top 5 for random forests.

Transcript	*Frequency in top 5 random forest important transcripts*	*Data*
APOC1	***6***	***Baseline, KLK2 and HK***
CADPS	***3***	***HK***
CCDC88B	***2***	***HK***
ERG 3' exons 6-7	***8***	***Baseline, KLK2 and HK***
NEAT1	***2***	***Baseline***
RIOK3	***3***	***Baseline***
TMPRSS2:ERG	***3***	***Baseline***
SLC12A1	***3***	***KLK2***
SPINK1	***3***	***Baseline + HK***
HOXC6	***4***	***KLK2 + HK***
PCA3	***2***	***KLK2***

Table 6.11 Mean square of residuals for random forest models for predicting CB vs. cancer (L, I and H) samples using different input probes across three different normalisations.

HK				*KLK2*			*Baseline*		
Input:	***All***	***glm***	***MWU***	***All***	***glm***	***MWU***	***All***	***glm***	***MWU***
Mean square of residuals	***0.114***	***0.118***	***0.138***	***0.115***	***0.088***	***0.104***	***0.106***	***0.099***	***0.103***

6.3.2 CB vs High risk cancer patients

6.3.2.1 Differentially expressed transcripts

The 12 samples with no evidence of cancer (CB) were compared to the 14 high-risk cancer samples (H) using glm and MWU tests. 51, 12, and 20 transcripts had a significant association (via logistic regression, section 2.6.1) with whether a sample had no evidence for cancer (CB) or was high-risk cancer (H) in the three processed datasets (the baseline data, *KLK2* ratio, and HK normalised, respectively, Table 6.12, Table 6.13, Table 6.14, Supplementary Table 37). None remained significant post multiple testing correction in the baseline, *KLK2* ratio or the HK normalised data. The top significant probe in these datasets was *NEAT1* ($p = 0.004$), *ERG* 3' exons 6-7 ($p = 0.008$), and *HOXC6* ($p = 0.005$), respectively.

Expression of 65, 25, and 35 transcripts had a significant association (via Mann Whitney U (MWU) testing, section 2.4.1) with whether a sample had no evidence for cancer (CB) or if the samples were high-risk cancer (H) in the three processed datasets (the baseline data, *KLK2* ratio, and HK normalised, respectively, Table 6.12, Table 6.13, Table 6.14, Supplementary Table 38). Post multiple testing correction, the expression of 10 (*ERG* 3' exons 6-7, *B4GALNT4, RIOK3, CADPS, MCTP1, HOXC6, NEAT1, CLIC2, APOC1* and *SIM2* long), 1 (*HOXC6*) and 0, remained significant. The top significant probes identified by MWU were *ERG* 3' exons 6-7 ($p = 6.21\text{x}10^{-06}$), *HOXC6* ($p = 4.28\text{x}10^{-05}$ and adjusted $p = 0.007$) and *HOXC6* ($p = 0.0005$), respectively.

Table 6.12 Top 10 transcripts with biggest log_2 fold change between CB and HR-cancer in the baseline data.

	glm		*MWU*		
Transcript	***p - value***	***adjusted p - value***	***p - value***	***adjusted p - value***	***Log_2(FC)***
HOXC6	***0.0002***	***0.0299***	***0.004***	***0.6711***	***2***
ERG3' exons 6-7	***$6.21x10^{-06}$***	***0.001***	***0.0371***	***0.9942***	***1.6***
TDRD	***0.0011***	***0.1558***	***0.0333***	***0.9942***	***1.5***
TMPRSS2:ERG	***0.0004***	***0.0668***	***0.0386***	***0.9942***	***1.3***
B4GALNT4	***$2.88x10^{-05}$***	***0.0048***	***0.0409***	***0.9942***	***1.2***
SLC43A1	***0.002***	***0.2897***	***0.0117***	***0.9942***	***1.2***
CADPS	***$6.70x10^{-05}$***	***0.011***	***0.02***	***0.9942***	***1.1***
CLIC2	***0.0002***	***0.0386***	***0.0087***	***0.9942***	***1***
HPN	***0.0008***	***0.1258***	***0.0092***	***0.9942***	***0.9***
LASS1	***0.0011***	***0.1558***	***0.0103***	***0.9942***	***0.9***

Table 6.13 Top 10 transcripts with biggest log_2 fold change between CB and HR-cancer in the *KLK2* ratio data.

	glm		*MWU*		
Transcript	***p - value***	***adjusted p - value***	***p - value***	***adjusted p - value***	***Log_2(FC)***
TMPRSS2:ERG	***0.004***	***0.68***	***0.028***	***1.000***	***0.25***
ERG 3' exons 6-7	***0.000***	***0.07***	***0.008***	***1.000***	***0.25***
HOXC6	***4.28E-05***	***0.01***			***0.25***
TDRD	***0.001***	***0.09***	***0.017***	***1.000***	***0.24***
SLC43A1	***0.002***	***0.27***	***0.022***	***1.000***	***0.21***
CADPS	***0.007***	***1***			***0.18***
B4GALNT4	***0.002***	***0.33***	***0.035***	***1.000***	***0.17***
ERG 5'	***0.027***	***1***			***0.16***
SLC12A1	***0.013***	***1***			***0.15***
ERG 3' exons 4-5	***0.046***	***1***	***0.050***	***1.000***	***0.14***

Table 6.14 Top 10 transcripts with biggest log_2 fold change between CB and HR-cancer in the HK normalised data.

	glm		*MWU*		
Transcript	***p - value***	***adjusted p - value***	***p - value***	***adjusted p - value***	***Log_2(FC)***
HOXC6	***0.0005***	***0.0882***	***0.0059***	***0.9765***	***1.6***
ERG3' exons 6-7	***0.0013***	***0.2186***	***0.0266***	***0.9765***	***1.4***
TDRD	***0.0031***	***0.4948***	***0.0272***	***0.9765***	***1.1***
TMPRSS2:ERG	***0.0094***	***1***	***0.033***	***0.9765***	***1.1***
ST6GALNAC1	***0.0037***	***0.5969***	***0.0168***	***0.9765***	***-1***
SLC43A1	***0.0013***	***0.2186***	***0.0197***	***0.9765***	***0.9***
B4GALNT4	***0.0202***	***1***			***0.8***
HPN	***0.0077***	***1***	***0.0314***	***0.9765***	***0.8***
CADPS	***0.0145***	***1***	***0.0326***	***0.9765***	***0.7***
CCDC88B	***0.031***	***1***	***0.0482***	***0.9765***	***0.7***

6.3.2.2 Models and gene selection

A number of different transcript subsets were input to Lasso for probe shrinkage and selection, these included i) all of the transcripts (n = 167). Transcripts identified as having significantly different expression between cancer and CB using ii) Mann Whitney U (n = 65, n = 25 and n = 35) and iii) logistic regression (n = 51, n = 12, and n = 20), separately, and iv) transcripts common to both those identified by Mann Whitney U and logistic regression (n = 49, n = 12 and n = 20) for each of the three normalisations (the baseline data, *KLK2* ratio, and HK normalised), respectively. The AUC, sensitivity and specificity of each model on the same training data was collected (Table 6.15) and transcript lists (Table 6.16, Table 6.17, Table 6.18) and boxplots of the Lasso selected probes were produced (section 2.6.1, Supplementary Figure 16, Supplementary Figure 17 and Supplementary Figure 18).

Table 6.15 AUC, Sensitivity and Specificity of models to predict CB vs. high-risk cancer (H) in different data normalisations of cell NanoString data.

	HK normalised				*KLK2 ratio*				*Baseline*			
	All Transcripts	***MWU***	***glm***	***Both***	***All Transcripts***	***MWU***	***glm***	***Both***	***All Transcripts***	***MWU***	***glm***	***Both***
AUC	***1***	***1***	***1***	***1***	***0.952***	***0.905***	***0.958***	***0.905***	***1***	***1***	***1***	***1***
Sensitivity	***100%***	***100%***	***100%***	***100%***	***93%***	***86%***	***93%***	***86%***	***100%***	***100%***	***100%***	***100%***
Specificity	***100%***	***100%***	***100%***	***100%***	***92%***	***83%***	***92%***	***83%***	***100%***	***100%***	***100%***	***100%***
Number of Probes	***6****	***6****	***7***	***6****	***2****	***2*****	***2****	***2*****	***9***	***9***	***10***	***4***

*, ** and *** have identical probes selected for the model.

Table 6.16 Beta values of individual transcripts within HR cancer and CB models suggested by Lasso using different input transcripts for the baseline normalised data.

Transcript	*Beta – All Transcripts*	*Beta - MWU*	*Beta - glm*	*Beta - both*
AATF	***0.094***	***0.696***	***0.982***	
CADPS	***0.773***	***1.255***	***1.798***	***0.337***
CAMKK2		***0.055***		
CCDC88B		***0.037***		
CDKN3			***-0.101***	
CKAP2L	***0.042***		***0.358***	
ERG 3' exons 6-7	***0.135***	***0.219***	***0.193***	***0.096***
HOXC6	***0.197***	***0.218***	***0.168***	***0.115***
IGFBP3	***-0.051***	***-0.288***		
LASS1	***0.187***	***0.186***	***0.623***	
MCTP1	***0.003***			***0.029***
MMP25		***0.092***	***0.197***	
NAALADL2	***-0.121***			
SIM2 long			***0.337***	
TDRD		***0.084***		

Table 6.17 Beta values of individual transcripts within HR cancer and CB models suggested by Lasso using different input transcripts for the *KLK2* ratio data.

Transcript	*Beta – All Transcripts*	*Beta - MWU*	*Beta - glm*	*Beta - both*
ERG 3' exons 6-7	***0.3394***	***0.5927***	***0.391***	***0.391***
HOXC6	***0.0287***	***0.1029***		
SIM2 long			***0.0349***	***0.0349***

Table 6.18 Beta values of individual transcripts within HR cancer and CB models suggested by Lasso using different input transcripts for the HK normalised data.

Transcript	*Beta – All transcripts*	*Beta - MWU*	*Beta – glm*	*Beta - both*
CADPS	***0.3134***	***0.8661***	***0.4861***	***1.2907***
ERG 3' exons 6-7	***0.0636***	***0.0684***	***0.0784***	***0.0329***
GJB1	***-0.0503***	***-0.009***	***-0.0518***	
HOXC6	***0.1835***	***0.2967***	***0.2332***	***0.3504***
NAALADL2	***-0.1273***	***-0.3281***	***-0.1872***	***-0.4819***
SIM2 long		***0.1225***		***0.3251***
SPINK1	***-0.0754***	***-0.0949***	***-0.0812***	***-0.1063***

Random forest was also applied to i) all transcripts, ii) significant transcripts identified by MWU and iii) significant transcripts identified by glm for the three different normalisations (the baseline data, *KLK2* ratio, and HK normalised), respectively (Supplementary Table 40, Supplementary Table 41 and Supplementary Table 42).

Using the glm identified significant probes in the HK normalised data gives the models with the smallest error (mean square of residuals: 0.117), although all models are very similar (Table 6.20).

HOXC6 was in the top 5 transcripts in 8/9 random forests, whilst *CADPS* was in the top 5 transcripts in 7/9 random forests. *ERG3'* exons 6-7 and *SPINK1* occur in 4/9, and *ST6GALNAC1* and *TDRD* occur in 3/9 random forests (Table 6.19).

Table 6.19 Frequency of transcripts in top 5 for random forests (CB vs high-risk cancer models).

Transcript	*Frequency in top 5 random forest important transcripts*	*Data*
CADPS	***7***	***Baseline + KLK2 + HK***
CCDC88B	***2***	***Baseline***
ERG3' exons 6-7	***4***	***Baseline + KLK2 + HK***
HOXC6	***8***	***Baseline + KLK2 + HK***
SIM2 long	***2***	***KLK2***
SLC43A1	***2***	***KLK2***
SPINK1	***4***	***Baseline + HK***
ST6GALNAC1	***3***	***HK***
TDRD	***3***	***KLK2***
VAX2	***2***	***HK***

Table 6.20 Mean Square of residuals error for each random forest model produced using different input probes in three different normalisations.

	HK			*KLK2*			*Baseline*		
Input:	***All***	***glm***	***MWU***	***All***	***glm***	***MWU***	***All***	***glm***	***MWU***
Mean square of residuals	***0.18***	***0.117***	***0.138***	***0.149***	***0.18***	***0.14***	***0.146***	***0.138***	***0.145***

6.3.3 Trend CBN-L-I-H

6.3.3.1 Significant transcripts

Trend (increase or decrease) in expression across the 12 CB samples, 4 low-risk, 28 intermediate risk and 14 high-risk samples was investigated. Two methods of ordered multinomial regression were used: i) proportional odds logistic regression (polr) and ii)

logistic regression setting clinical group to an ordered integer. Using polr there were 70, 20 and 15 transcripts that significantly modelled the trend in the three processed datasets (the baseline data, *KLK2* ratio, and HK normalised, respectively), ($p < 0.05$, Supplementary Table 43). Of these only 7 (*B4GALNT4, HOXC6, ERG3'* exons 6-7, *APOC1*, *TMPRSS2:ERG, NEAT1,* and *MCTP1*), 2 (*HOXC6* and *ERG* 3' exons 6-7) and 1 (*HOXC6*) remained significant post multiple testing correction. The top significant probes identified by polr were *APOC1*, *HOXC6* and *ERG* 3' exons 6-7 jointly ($p = 0.0001$), *HOXC6* ($p = 1.36\text{x}10^{-05}$) and *HOXC6* ($p = 4.54\ \text{x}10^{-6}$), respectively.

Using logistic regression there were 87, 36 and 19 transcripts that modelled trend with statistical significance in the three processed datasets (the baseline data, *KLK2* ratio, and HK normalised, respectively), ($p < 0.05$, Supplementary Table 44). Of these 19, 4 (*HOXC6*, *ERG3*' exons 6-7, *TMPRSS2:ERG* and *TDRD*), and 1 (*HOXC6*) remained significant post multiple testing correction, respectively. The top significant probes identified by polr were *APOC1* ($p = 2.90\text{x}10^{-06}$, adjusted p-value = 0.0005), *ERG* 3' exons 6-7 and *HOXC6* jointly ($p = 0.0002$) and *HOXC6* ($p = 6.37\text{x}10^{-05}$), respectively (Table 6.21, Table 6.22, Table 6.23).

Polr identifies fewer transcripts than glm but all but one transcript identified by polr were also identified by logistic regression in each case, showing robustness in their identification. Similar probes were identified as most significant by the two methods also.

Table 6.21 Top 15 significant transcripts identified by polr to have trend across CB - L - I - H clinical categories in the baseline normalised cell data.

Transcript	glm *p* -value	glm adjusted *p*-	polr *p*-value	polr adjusted *p*-
APOC1	**2.90x10^{-06}**	**0.0005**	**0.0001**	**0.0246**
ERG3' exons 6-7	5.87x10^{-06}	0.001	**0.0001**	**0.0191**
HOXC6	1.85x10^{-05}	0.003	**0.0001**	**0.0191**
TMPRSS2:ERG	6.00x10^{-05}	0.0095	0.0002	0.0385
MCTP1	4.10x10^{-06}	0.0007	0.0003	0.0481
NEAT1	0.0002	0.0265	0.0003	0.047
RIOK3	6.64x10^{-06}	0.0011	0.0003	0.0515
ISX	3.91x10^{-05}	0.0063	0.0004	0.066
HPN	6.51x10^{-05}	0.0102	0.0007	0.1079
GCNT1	0.0002	0.025	0.0008	0.1318
SULF2	2.23x10^{-05}	0.0036	0.0008	0.1288
CAMKK2	5.44x10^{-05}	0.0087	0.0012	0.1813
MMP25	0.0002	0.0277	0.0014	0.2178
CADPS	0.0001	0.019	0.0017	0.261
LASS1	0.0002	0.0331	0.0019	0.2854

Table 6.22 Top 15 significant transcripts identified by polr to have trend across CB - L - I - H clinical categories in the *KLK2* ratio cell data.

Transcript	glm *p* -value	glm adjusted *p*-	polr *p*-value	polr adjusted *p*-
***ERG3'* exons 6-7**	2.18 x10^{-05}	0.0036	**0.0002**	**0.0283**
HOXC6	**1.36x10^{-05}**	**0.0023**	0.0002	0.0406
TMPRSS2:ERG	6.86 x10^{-05}	0.0112	0.0007	0.1136
TDRD	0.0002	0.0263	0.0011	0.1757
SIM2 long	0.0031	0.5021	0.0056	0.9028
HPN	0.0027	0.4419	0.0081	0.994
GCNT1	0.0066	0.998	0.0104	0.994
CADPS	0.0052	0.8154	0.0112	0.994
TMEM86A	0.0079	0.998	0.0184	0.994
CKAP2L	0.0046	0.731	0.0187	0.994
LASS1	0.005	0.7872	0.0209	0.994
ERG3' exons 4-5	0.0275	0.998	0.0243	0.994
FOLH1	0.0124	0.998	0.0348	0.994
ISX	0.0022	0.3607	0.0367	0.994
ANKRD34B	0.0112	0.998	0.0374	0.994

Table 6.23 Top 15 significant transcripts identified by polr to have trend across CB - L - I - H clinical categories in the HK normalised cell data.

Transcript	glm *p* -value	glm adjusted *p*-	polr *p*-value	polr adjusted *p*-
HOXC6	**4.54 x10^{-6}**	**0.0008**	**6.37x10^{-05}**	**0.0106**
TDRD	0.0012	0.2024	0.0034	0.564
SIM2 long	0.0032	0.5147	0.0043	0.7056
SLC43A1	0.0011	0.1895	0.006	0.978
UPK2	0.0028	0.4609	0.0077	0.9994
ERG 3' exons 6-7	0.0043	0.6877	0.0098	0.9994
NAALADL2	0.0018	0.2913	0.0098	0.9994
TMPRSS2:ERG fusion	0.004	0.6414	0.0127	0.9994
ST6GALNAC1	0.0049	0.7755	0.0179	0.9994
FOLH1	0.0174	0.9941	0.0191	0.9994
MEX3A	0.0243	0.9941	0.0337	0.9994
TMEM86A	0.0107	0.9941	0.0337	0.9994
SERPINB5	0.0162	0.9941	0.0425	0.9994
PALM3	0.027	0.9941	0.0461	0.9994
EN2			0.0463	0.9994

6.3.3.2 Models and gene selection

A number of different transcript subsets were input to Lasso for probe shrinkage and selection, these included i) all of the transcripts (*n* = 167). Transcripts identified as having significant decrease or increase in expression across CB->L->I->H clinical categories using ii) polr (*n* = 70, *n* = 20 and *n* = 15), and iii) logistic regression (*n* = 87, *n* = 36, and *n* = 19), separately, for each of the three normalisations (the baseline data, *KLK2* ratio, and HK normalised), respectively (Supplementary Table 45). In addition, the transcripts common to both those identified by polr and glm for the HK normalised data only was also submitted to Lasso (*n* = 14), these were the only significant transcript lists where polr did not contain all of the glm identified probes. *APOC1* was the only probe selected by Lasso in all three transcript inputs for the baseline normalised data. *HOXC6* was the only probe selected by Lasso in the *KLK2* ratio data. *HOXC6*, *NAALADL2* and *UPK2* were common probes selected by Lasso in the HK normalised data.

The AUC, sensitivity and specificity of each model on the same training data was collected (Table 6.27) and transcript lists (Table 6.24, Table 6.25, Table 6.26) and boxplots of the Lasso selected probes were produced (Supplementary Figure 19, Supplementary Figure 20 and Supplementary Figure 21).

Table 6.24 Optimal multinomial models for predicting clinical category (CB, low-risk, intermediate-risk, and high-risk cancer) with different subsets of input transcripts (from preliminary ordered glm and polr tests) in the baseline normalised cell data.

Transcript	*All transcripts (n = 167) - Beta*	*glm (n = 87) - Beta*	*polr (n = 70) - Beta*
AATF	***0.1***	***0.115***	***0.109***
APOC1		***0.056***	
B4GALNT4	***0.004***	***0.121***	
CADPS	***0.068***		***0.105***
CAMKK2		***0.062***	
CCDC88B		***0.036***	
EN2			***0.024***
ERG 3' exons 6-7	***0.154***	***0.139***	***0.144***
HOXC6	***0.16***		***0.121***
KLK3 exons 2-3	***-0.022***		
LASS1	***0.034***		***0.017***
MCTP1	***0.037***		***0.095***
MMP25	***0.007***		***0.088***
NAALADL2	***-0.104***	***-0.034***	
RIOK3	***0.016***		***0.095***
SPINK1			***-0.1***
SULF2	***0.088***		***0.046***
VAX1	***-0.131***		
Cp1	***1.198***	***0.702***	***1.283***
Cp2	***2.018***	***1.786***	***2.132***
Cp3	***-0.414***	***-0.368***	***-0.436***

Table 6.25 Optimal multinomial models for predicting clinical category (CB, low-risk, intermediate-risk, and high-risk cancer) with different subsets of input transcripts (from preliminary ordered glm and polr tests) in *KLK2* ratio cell data

Transcript	*All transcripts (n = 94) - Beta*	*glm (n = 36) - Beta*	*polr (n = 20) - Beta*
CADPS	***0.027***	***0.075***	***0.026***
CKAP2L		***0.074***	
ERG3' exons 4-5		***-0.121***	
ERG3' exons 6-7	***0.809***	***0.784***	***0.591***
HOXC4	***-0.134***		
HOXC6	***0.264***	***0.475***	***0.34***
ITGBL1	***-0.118***		
NAALADL2	***-0.479***		
PALM3	***-0.017***		
RIOK3		***-0.487***	
TDRD		***0.031***	
TMPRSS2:ERG	***0.008***	***0.12***	***0.131***
Cp1	***0.998***	***1.094***	***0.837***
Cp2	***1.938***	***2.065***	***1.937***
Cp3	***-0.401***	***-0.427***	***-0.400***

Table 6.26 Optimal multinomial models for predicting clinical category (CB, low-risk, intermediate-risk, and high-risk cancer) with different subsets of input transcripts (from preliminary ordered glm and polr tests) in HK normalised cell data.

Transcript	*All transcripts (n = 167) - Beta*	*glm (n = 19) - Beta*	*polr (n = 15) - Beta*	*glm and polr (n = 14) - Beta*
CADPS		***0.035***		
CLIC2		***0.043***		
ERG 3' exons 6-7	***0.088***	***0.19***	***0.179***	***0.211***
GJB1	***-0.057***	***-0.233***		
HOXC6	***0.151***	***0.266***	***0.205***	***0.234***
NAALADL2	***-0.094***	***-0.183***	***-0.235***	***-0.285***
PALM3		***-0.028***	***-0.045***	***-0.074***
SLC43A1		***0.015***	***0.033***	***0.058***
TDRD		***0.045***		
TMEM86A				***0.023***
UPK2	***-0.014***		***-0.003***	
Cp1	***0.65***	***1.622***	***1.276***	***1.563***
Cp2	***1.76***	***2.321***	***2.089***	***2.26***
Cp3	***-0.361***	***-0.466***	***-0.425***	***-0.456***

Table 6.27 AUC, Sensitivity and Specificity of models to predict trend across clinical categories: CB > L- > I- > H-risk cancer in different data normalisations of cell NanoString data.

Data type:	*Baseline*			*KLK2*			*HK*			
Model Input:	***All transcripts***	***Glm***	***Polr***	***All transcripts***	***Glm***	***Polr***	***All transcripts***	***Glm***	***Polr***	***glm + polr***
Accuracy	***0.7069***	***0.6552***	***0.6897***	***0.7069***	***0.6379***	***0.6207***	***0.6552***	***0.6897***	***0.6897***	***0.6897***
AUC	***0.7604***	***0.7242***	***0.7669***	***0.7504***	***0.7252***	***0.702***	***0.6944***	***0.7609***	***0.7609***	***0.7609***
Sensitivity: CB	***83%***	***67%***	***92%***	***75%***	***67%***	***58%***	***67%***	***83%***	***83%***	***83%***
L	***0%***	***0%***	***0%***	***0%***	***0%***	***0%***	***0%***	***0%***	***0%***	***0%***
I	***96%***	***93%***	***93%***	***96%***	***89%***	***89%***	***100%***	***93%***	***93%***	***93%***
H	***29%***	***26%***	***21%***	***36%***	***29%***	***29%***	***14%***	***29%***	***29%***	***29%***
Specificity: CB	***100%***	***100%***	***100%***	***98%***	***100%***	***98%***	***100%***	***100%***	***100%***	***100%***
L	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***	***100%***
I	***47%***	***40%***	***47%***	***47%***	***40%***	***37%***	***33%***	***47%***	***47%***	***47%***
H	***98%***	***95%***	***95%***	***100%***	***93%***	***95%***	***100%***	***95%***	***95%***	***95%***
PPV: CB	***100%***	***100%***	***100%***	***90%***	***100%***	***88%***	***100%***	***100%***	***100%***	***100%***
L	***NA***	***NA***	***NA***	***NA***	***NA***	***NA***	***NA***	***NA***	***NA***	***NA***
I	***63%***	***59%***	***62%***	***63%***	***58%***	***57%***	***58%***	***62%***	***62%***	***62%***
H	***80%***	***67%***	***60%***	***100%***	***57%***	***67%***	***100%***	***67%***	***67%***	***67%***
Number of Probes	***13***	***7***	***11***	***8***	***8***	***4***	***5***	***9***	***6***	***6***

Random forest was also applied to i) all transcripts, ii) significant transcripts identified by polr and iii) significant transcripts identified by glm for the three different normalisations (the baseline data, *KLK2* ratio, and HK normalised), respectively (Supplementary Table 46, Supplementary Table 47 and Supplementary Table 48). Using all probes in the baseline normalised data gives the models with the smallest error (OOB error: 27.6%, Table 6.28). *ERG3'* exons 6-7 was present in 8/9 random forest models, whilst *TMPRSS2:ERG* and *HOXC6* were present in 6/9 RF models (Table 6.29).

Table 6.28 OOB error rates for random forest models built to predict trend over clinical categories: CB > L > I >H

	Baseline			*KLK2*			*HK*			
Input:	***All***	***glm***	***polr***	***All***	***glm***	***polr***	***All***	***glm***	***polr***	***Glm + polr***
OOB error	***27.6 %***	***50%***	***50%***	***46.6 %***	***48.3 %***	***51.7 %***	***44.8 %***	***43.1 %***	***41.4 %***	***44.8 %***

Table 6.29 Frequency of transcripts in top 5 for random forests (CB > L > I >H trend models).

Transcript	*Frequency in top 5 random forest important transcripts*	*Data*
ERG3' exons 6-7	***8***	***Baseline, KLK2 and HK***
TMPRSS2:ERG	***6***	***Baseline, KLK2 and HK***
HOXC6	***6***	***KLK2 and HK***
PCA3	***3***	***KLK2 and HK***
PALM3	***3***	***HK***
RIOK3	***2***	***Baseline***
NEAT1	***2***	***Baseline***
CADPS	***2***	***Baseline***
APOC1	***2***	***Baseline***
FOLH1	***2***	***KLK2***
NAALADL2	***2***	***HK***
UPK2	***2***	***HK***

6.4 Summary of Predictive Models

In the *KLK2* ratio data half of the models had better AUCs on the training set from the cell data and half from the EV data (Table 6.30). However, in the HK data, the AUCs were higher in the cell data. I am limited in the number of samples for the cell fraction and so the models have not been applied to a test data set, this means that the models could be over fitting the data.

Table 6.30 Comparison of AUCs from models using cell and EV data.

	Cell	EV
KLK2 ratio data		
CB vs Cancer (L, I, H) All transcripts	0.996	0.949
CB vs Cancer (L, I, H) Significant transcripts	0.998	0.886
CB vs HR Cancer (H) All transcripts	0.952	0.991
CB vs HR Cancer (H) Significant transcripts	0.958	0.97
CB > L > I > H All transcripts	0.7504	0.7663
CB > L > I > H Significant transcripts	0.702	0.6757
HK normalised data		
CB vs Cancer (L, I, H) All transcripts	0.989	0.925
CB vs Cancer (L, I, H) Significant transcripts	0.998	0.902
CB vs HR Cancer (H) All transcripts	1	0.976
CB vs HR Cancer (H) Significant transcripts	1	0.992
CB > L > I > H All transcripts	0.7609	0.7587
CB > L > I > H Significant transcripts	0.7609	0.7728

Table 6.31 Transcripts identified by all selection models for the different clinical category tests across the different normalisations on the cell NanoString data.

Normalisation	*Clinically Benign vs. Cancer*	*Clinically Benign vs. High risk cancer*	*Trend Clinically Benign, low-risk, intermediate-risk and high-risk*
Baseline	***ACTR5*** ***APOC1*** ***ARHGEF25*** ***CADPS*** ***CAMKK2*** ***ERG3' exon 6-7*** ***EN2*** ***HIST1H2BG*** ***HOXC6*** ***IGFBP3*** ***LASS1*** ***MCTP1*** ***MMP25*** ***MMP26*** ***NAALADL2*** ***PCA3*** ***RIOK3*** ***SPINK1*** ***SLC12A1*** ***TDRD***	***AATF*** ***CADPS*** ***CAMKK2*** ***CCDC88B*** ***CDKN3*** ***CKAP2L*** ***ERG3' exon 6-7*** ***HOXC6*** ***ITGFBP3*** ***LASS1*** ***MCTP1*** ***MMP25*** ***NAALADL2*** ***SIM2 long*** ***TDRD***	***AATF*** ***APOC1*** ***B4GALNT4*** ***CADPS*** ***CAMKK2*** ***CCDC88B*** ***EN2*** ***ERG3' exon 6-7*** ***HOXC6*** ***KLK3 exons 2-3*** ***LASS1*** ***MCTP1*** ***MMP25*** ***NAALADL2*** ***RIOK3*** ***SPINK1*** ***SULF2*** ***VAX1***
KLK2 ratio	***CADPS*** ***CKAP2L*** ***EN2*** ***ERG3' exons 6-7*** ***HOXC6*** ***MFSD2A*** ***NAALADL2*** ***SFRP4*** ***SIM2 long*** ***TDRD***	***ERG3' exons 6-7*** ***HOXC6*** ***SIM2 long***	***CADPS*** ***CKAP2L*** ***ERG3' exons 4-5*** ***ERG3' exons 6-7*** ***HOXC4*** ***HOXC6*** ***ITGBL1*** ***NAALADL2*** ***PALM3*** ***RIOK3*** ***TDRD*** ***TMPRSS2:ERG***
RPLP2 and TWIST1 normalised	***CADPS*** ***CLIC2*** ***ERG3' exons 6-7*** ***HOXC6*** ***NAALADL2*** ***SIM2 long*** ***TDRD*** ***UPK2***	***CADPS*** ***ERG3' exon 6-7*** ***GJB1*** ***HOXC6*** ***NAALADL2*** ***SIM2 long*** ***SPINK1***	***CADPS*** ***CLIC2*** ***ERG3' exons 6-7*** ***GJB1*** ***HOXC6*** ***NAALADL2*** ***PALM3*** ***SLC43A1*** ***TDRD*** ***TMEM86A*** ***UPK2***

Table 6.32 Transcripts selected for models in EV data.

Normalisation	*Clinically Benign vs. Cancer*	*Clinically Benign vs. High risk cancer*	*Trend Clinically Benign, low-risk, intermediate-risk and high-risk*
KLK2 ratio	***AMACR***	***ACTR5***	***AMACR***
	APOC1	***ALAS1***	***ANKRD34B***
	AR exon 9	***AMACR***	***APOC1***
	CP	***ANKRD34B***	***AR exon 9***
	DLX1	***APOC1***	***AR exons 4-8***
	ERG3' exon 4-5	***AR exon 9***	***BTG2***
	GJB1	***AR exons 4-8***	***CD10***
	HOXC6	***AURKA***	***CP***
	IGFBP3	***BTG2***	***DLX1***
	ISX	***CD10***	***DPP4***
	KLK4	***CKAP2L***	***ERG 3' exons 4-5***
	MXI1	***CP***	***ERG 3' exons 6-7***
	NEAT1	***DLX1***	***GABARAPL2***
	PCA3	***DPP4***	***HIST1H1E***
	PPP1R12B	***ERG 3' exons***	***HOXC6***
	RNF157	***4-5***	***HPN***
	ST6GALNAC	***HOXC6***	***IGFBP3***
	SULT1A1	***HPN***	***ISX***
	TDRD	***IGFBP3***	***ITGBL1***
	TMEM47	***ISX***	***KLK4***
	TMPRSS2:ERG	***KLK4***	***MED4***
		MAK	***MEMO1***
		MED4	***MXI1***
		MMP25	***MYOF***
		NEAT1	***NEAT1***
		PCA3	***PCA3***
		PDLIM5	***PPP1R12B***
		PPFIA2	***PSGR***
		PSTPIP1	***PSTPIP1***
		PTPRC	***SLC12A1***
		RPL18A	***SRSF3***
		SRSF3	***SULT1A1***
		STEAP4	***TDRD***
		TMEM47	***Timp4***
		TMPRSS2:ERG	***TMEM47***
			TMPRSS2:ERG
			ZNF577
RPLP2 and GAPDH normalised	***APOC1***	***AMACR***	***ACT5R***
	AR exon 9	***ANKRD34B***	***AMH***
	CD10	***APOC1***	***ANKRD34B***
	CP	***AR exon 9***	***APOC1***
	ERG3' exons 4-5	***AR exon 4-8***	***AR exon 9***
	GABARAPL2	***CD10***	***AR exon 4-8***
	HOXC6	***DLX1***	***CD10***
	HPN	***DPP4***	***CP***
	ISX	***ERG3' exons 4-5***	***DPP4***
	KLK4	***GABARAPL2***	***ERG3' exons 4-5***
	MCTP1	***HOXC6***	***ERG3' exons 6-7***
			FDPS

PCA3	***HPN***	***GABARAPL2***
PDLIM5	***KLK4***	***GCNT1***
PPP1R12B	***MYOF***	***GJB1***
PTN	***NEAT1***	***HIST1H1E***
SLC12A1	***PCA3***	***HIST1H2BF***
SULT1A1	***PDLIM5***	***HOXC6***
TDRD	***SLC12A1***	***HPN***
TMPRSS2:ERG	***SRSF3***	***IGFBP3***
	STOM	***ISX***
	SULT1A1	***ITGBL1***
	TMPRSS2:ERG	***KLK4***
		MED4
		MEMO1
		MIATNB
		MSMB
		MXI1
		MYOF
		NEAT1
		PCA3
		PPP1R12B
		RPS10
		SLC12A1
		SPINK1
		SRSF3
		SULT1A1
		TDRD
		Timp4
		TMPRSS2:ERG
		TRPM4
		UPK2
		ZNF577

Comparing the transcripts selected for models in the cell *KLK2* ratio data (Table 6.31) and the EV *KLK2* ratio data (Table 6.32), only 5 transcripts were selected for both sets of models (*CKAP2L, HOXC6, TDRD, ITGBL1* and *TMPRSS2:ERG*). The same comparison for the HK normalised data yielded a different 5 transcripts in common (*ERG* 3' exons 6-7, *HOXC6, TDRD, GJB1* and *UPK2*). This shows that different probes are selected as important for predictive models between the different fractions of urine (cell vs EV).

6.5 Comparison of the urine expression profiles of Extracellular vesicle and Cell fractions in Prostate Cancer

6.5.1 Microarray comparison of the global expression profile of Extracellular vesicle and Cell fractions.

I examined Affymetrix microarray expression data from the cell sediment and EV fraction of urine collected from prostate cancer patients from either the NNUH or the Royal Marsden Hospital NHS foundation trust (n = 3). Genes that were significantly differentially expressed between the two fractions were determined by Dr. Daniel Brewer using the Limma package and the method proposed by Mootha *et al*., 2003, to give a value for variance of expression and if it significantly differs between fractions[219]. 98 genes were found to be up-regulated in the extracellular vesicles and 116 up-regulated in the cell sediment fraction.

Table 6.33 A list of the top 20 microarray detected transcripts out of 98 that were found to be significantly more abundant in extracellular vesicles compared with sediment from the same urine.

Rank	*Gene*	*Log_2(FC)*	*p-value*	*Tissue Expression*	*Known Cancer Associations*
1	***TMSB15A***	***5.19***	***0.048***	***Prostate***	***Prostate[220], Other[221,222,223]***
2	***PRKG2***	***5.02***	***0.035***	***Prostate, Other***	***N***
3	***TCEA3***	***4.87***	***0.048***	***Other***	***N***
4	***PRAC***	***4.85***	***0.035***	***Prostate, Other***	***Prostate[224], Other[225]***
5	***KLK4***	***4.82***	***0.041***	***Prostate***	***Prostate[226,227]***
6	***FOLH1, FOLH1B***	***4.59***	***0.048***	***Prostate, Other***	***Prostate[228]***
7	***EPHX2***	***4.58***	***0.041***	***Expressed in all***	***Prostate[229], Other[230,231,232]***
8	***GMPR***	***4.57***	***0.042***	***Expressed in all (higher expression in Prostate, Other)***	***Prostate[233]***
9	***RANBP3L***	***4.5***	***0.046***	***Prostate, Kidney, Other***	***Multiple[234,235]***
10	***MPPED2***	***4.39***	***0.047***	***Prostate, Other***	***Other[236]***
11	***CKB***	***4.17***	***0.035***	***Expressed in all (highest expression in Prostate)***	***Prostate[237], Other[238]***
12	***MLPH***	***4.09***	***0.045***	***Prostate, Other***	***Prostate[239]***
13	***NFIA***	***4.06***	***0.048***	***Expressed in all***	***Prostate[240], Other[241]***
14	***GLYATL1***	***4.00***	***0.049***	***Prostate, Kidney, Other***	***Other[242]***
15	***NFIB***	***3.98***	***0.048***	***Mixed***	***Prostate[243], Other[244]***
16	***CCDC88C***	***3.98***	***0.043***	***Expressed in all***	***Other[245]***
17	***HOXB13***	***3.97***	***0.042***	***Prostate***	***Prostate[246], Other[246]***
18	***PART1***	***3.95***	***0.035***	***Prostate****	***Prostate[247]***
19	***AZGP1***	***3.85***	***0.043***	***Prostate, Other***	***Prostate[248], Other[249,250]***
20	***TCEAL2***	***3.73***	***0.048***	***Tissue Enhanced (glands, reproductive including prostate and cerebral cortex)***	***Other[251]***

I researched the top twenty up-regulated transcripts in the two fractions to examine the link between the genes, prostate tissue and cancer (Table 6.33, Table 6.34). Information about normal tissue expression was usually acquired from 'protein atlas'[252] but when this was not available, data was instead acquired from 'Genecards'[253]. Known cancer associations were determined using a literature search using the gene ID and the words 'cancer' or 'prostate cancer'. 80% of the top 20 genes up-regulated in extracellular vesicles were associated with prostate tissue, compared with 25% from the cell fraction. 65% of the top 20 genes up-regulated in extracellular vesicles were linked with prostate cancer and 65% cancer generally. The equivalent figures for the cell fraction were 30% and 65%. This is a strong indication that the extracellular vesicles contain RNA from prostate cancer cells and it is a better source of biomarkers than the cell fraction.

Table 6.34 A list of the top 20 microarray detected gene-transcripts out of 116 that were found to be significantly more abundant in the cell sediment compared with extracellular vesicles from the same urine.

Rank	*Gene*	*Log_2(FC)*	*p-value*	*Tissue Expression*	*Known Cancer Associations*
1	***SCARNA9***	***7.86***	***0.035***		***None***
2	***SNORD58A, SNORD58B***	***7.77***	***0.043***		***None***
3	***ALOX5AP***	***7.16***	***0.042***	***Other***	***Prostate[254], Other[255]***
4	***LYZ***	***7.12***	***0.035***	***Other***	***Other[256]***
5	***FCER1G***	***7.00***	***0.035***	***Prostate, Other***	***Other[257]***
6	***FCGR2A***	***6.76***	***0.048***	***Prostate, Other***	***None***
7	***CYBB***	***6.72***	***0.045***	***Other***	***Prostate[258], Other[259]***
8	***TNFRSF1B***	***6.71***	***0.045***	***Other***	***Other[260]***
9	***SCARNA9***	***6.45***	***0.044***		***None***
10	***SRGN***	***6.43***	***0.045***	***Other***	***Other[261]***
11	***IL8***	***6.12***	***0.035***	***Other***	***Prostate[262]***
12	***EVI2B***	***6.03***	***0.043***	***Other***	***Other[263]***
13	***TREM1***	***5.97***	***0.044***	***Other***	***Other[264]***
14	***MIR21***	***5.67***	***0.049***	***Not found***	***Prostate[265], Other[266, 267]***
15	***SCARNA7***	***5.66***	***0.043***	***Not found***	***None***
16	***HNRNPK***	***5.62***	***0.050***	***Prostate, Other***	***Prostate[268], Other[269]***
17	***GNS***	***5.58***	***0.035***	***Prostate, Other***	***None***
18	***CBX3***	***5.32***	***0.045***	***Prostate, Other***	***Other[270, 271]***
19	***CTSS***	***5.32***	***0.045***	***Other***	***Prostate[272], Other[273, 274]***
20	***ERO1L***	***5.23***	***0.041***	***Other***	***Other[275]***

6.5.2 NanoString comparison of the global expression profile of Extracellular vesicle and Cell fractions.

6.5.2.1 Visualisation of expression differences between fractions

NanoString data (167 probes) from both extracellular vesicle and cell fractions were available for 92 patients. In this section NanoString internal positive control normalised data was used. A PCA plot (section 2.5.1) was produced to visualise the variance of the cell sediment expression against extracellular vesicles expression (Figure 6.1). The expression profiles for the fractions cluster together, indicating that fraction has a bigger influence on the expression profile than the patient.

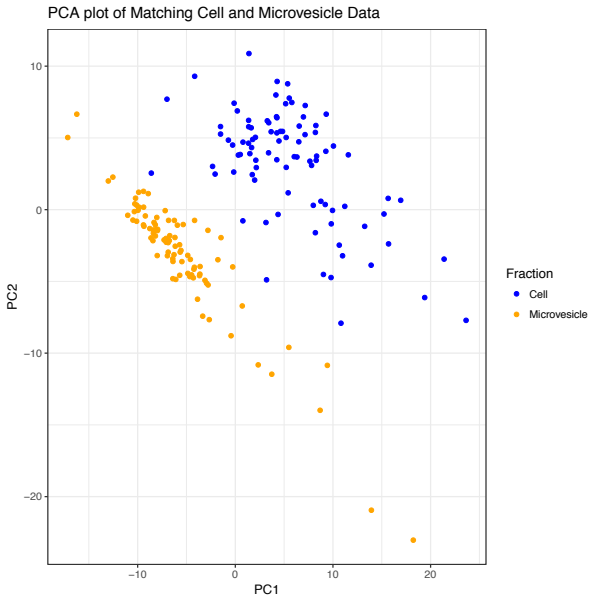


Figure 6.1 PCA plot of the expression levels for samples taken from the cell sediment and the extracellular vesicle fraction of urine.

6.5.2.2 Differentially expressed transcripts

Expression of 142/167 transcripts were significantly different between extracellular vesicle and cell fractions (adjusted $p < 0.05$, paired Mann Whitney U test). 100 were up-regulated in the extracellular vesicle fractions and 42 in the cell sediment fractions (Table 6.35, Table 6.36). *HOXC6* is a known PCa biomarker that can be identified in patient urine[276], it is therefore very interesting that it is found in abundance in EVs over whole urine. *PTPRC* is a positive regulator of T-cell coactivation and is found in immune cells[277].

Table 6.35 NanoString top twenty transcripts that were up-regulated in extracellular vesicle fractions compared to cell sediment fractions.

Transcript	*p*-value	Adjusted *p*-value	Log Fold change
HOXC6	$1.12x10^{-10}$	$1.07x10^{-08}$	0.81
SERPINB5	$9.78x10^{-14}$	$1.23x10^{-11}$	0.79
OR52A2	$3.57x10^{-13}$	$4.22x10^{-11}$	0.77
PTN	$1.17x10^{-15}$	$1.73x10^{-13}$	0.76
SChLAP1	$8.67x10^{-10}$	$7.89x10^{-08}$	0.67
P712P	$6.29x10^{-15}$	$8.69x10^{-13}$	0.67
PPFIA2	$2.07x10^{-12}$	$2.32x10^{-10}$	0.65
***SIM2* long**	$1.68x10^{-11}$	$1.74x10^{-09}$	0.65
***ERG3'* exons 4-5**	$2.31x10^{-05}$	0.0013	0.64
SMIM1	$1.69x10^{-13}$	$2.09x10^{-11}$	0.62
TMEM47	0.001	0.0434	0.61
CLU	$2.99x10^{-06}$	0.0002	0.61
Timp4	$6.74x10^{-11}$	$6.67x10^{-09}$	0.61
ARHGEF25	$6.56x10^{-10}$	$6.10x10^{-08}$	0.58
RNF157	$3.58x10^{-07}$	$2.47x10^{-05}$	0.58
PCA3	$7.66x10^{-14}$	$9.73x10^{-12}$	0.58
NKAIN1	$1.07x10^{-13}$	$1.34x10^{-11}$	0.57
DNAH5	$5.02x10^{-09}$	$4.22x10^{-07}$	0.57
KLK2	$8.19x10^{-16}$	$1.24x10^{-13}$	0.55
SYNM	$4.87x10^{-08}$	$3.55x10^{-06}$	0.54

Table 6.36 NanoString top twenty transcripts that were up-regulated in cell sediment fractions compared to extracellular vesicle fractions.

Transcript	*p*-value	Adjusted *p*-value	Log Fold change
PTPRC	$1.77x10^{-16}$	$2.94x10^{-14}$	-1.97
STOM	$2.93x10^{-16}$	$4.66x10^{-14}$	-1.73
SULF2	$2.48x10^{-16}$	$4.07x10^{-14}$	-1.69
MFSD2A	$3.24x10^{-16}$	$5.12x10^{-14}$	-1.66
NLRP3	$6.72x10^{-16}$	$1.03x10^{-13}$	-1.64
PSTPIP1	$3.96x10^{-16}$	$6.17x10^{-14}$	-1.44
MMP25	$2.17x10^{-16}$	$3.58x10^{-14}$	-1.43
CLIC2	$1.52x10^{-15}$	$2.20x10^{-13}$	-1.35
CCDC88B	$2.93x10^{-16}$	$4.66x10^{-14}$	-1.27
TMEM86A	$9.51x10^{-15}$	$1.27x10^{-12}$	-1.19
MKi67	$3.85x10^{-09}$	$3.27x10^{-07}$	-1.18
MAK	$1.39x10^{-14}$	$1.85x10^{-12}$	-1.15
MCTP1	$2.83x10^{-16}$	$4.59x10^{-14}$	-1.09
APOC1	$6.72x10^{-16}$	$1.03x10^{-13}$	-1.07
CP	$4.49x10^{-11}$	$4.54x10^{-09}$	-0.99
MIR146A	$1.74x10^{-15}$	$2.47x10^{-13}$	-0.96
NEAT1	$1.77x10^{-16}$	$2.94x10^{-14}$	-0.88
Met	$9.62x10^{-12}$	$1.01x10^{-09}$	-0.88
MIC1	$4.15x10^{-13}$	$4.85\ x10^{-11}$	-0.67
COL10A1	$2.09x10^{-11}$	$2.15x10^{-09}$	-0.59

6.6 Discussion

I found that the AUCs of the cell sediment models were marginally higher in the baseline normalised data for CB vs cancer models, CB vs high-risk cancer models and CB > L > I > H trend models (Table 6.30). However, these AUCs need to be taken with caution as the models have not been tested in a validation dataset and so overfitting may be occurring. There was a low number of samples used to build the cell predictive models and they all came from the same centre, so it is possible that the cell models are not as robust as one would desire. Comparing the transcripts identified via glm and those identified by Mann Whitney U, there were a large percentage of transcripts in common, suggesting a level of robustness when using these methods.

For the cell sediment, the transcripts identified as significantly different between the "No Evidence for Cancer" samples and the cancer samples differed depending on the normalisation (Supplementary Table 49). However, *CADPS* and *ERG3*' exons 6-7, *HOXC6, NAALADL2* and *TDRD* are present in all analyses. This shows a robustness of these

transcripts and indicates a level of importance when using cell sediment from urine samples. All are up-regulated in the cancer samples, with the exception of *NAALADL2*. *ERG3'* exons 6-7, *HOXC6* and *SIM2* long were the only probes that consistently distinguished high-risk cancer from CB samples. All three probes were up-regulated in the high-risk cancer samples. Looking at the trend across clinically benign, low-risk, intermediate-risk and high-risk cancer, *CADPS*, *ERG3*' exons 6-7, *HOXC6* and *NAALADL2* probes were again the common transcripts across all of the different normalisations. *CADPS* increases as risk increases, with lowest expression in CB and highest in high-risk cancer. *ERG* 3' exons 6-7 and *HOXC6* increase in low risk cancer but then have a decreased expression in intermediate and high-risk cancer with lowest expression found in CB. *NAALADL2* expression decreases in trend with advancement of cancer.

CADPS is a cytosolic and peripheral membrane protein required for vesicle docking and priming steps that precede vesicle exocytosis[278]. Down-regulation of *CADPS* has been associated with poor outcome in pancreatic ductal adenocarcinoma[279] and a genome wide molecular characterisation of central nervous system primitive neuroectodermal tumour and pineoblastoma found that the *CADPS* locus (3p14.2) was lost in 27.6% of cases and was also associated with poor prognosis[280]. Searching for "*CADPS* prostate cancer" yields no results during a literature search. *ERG* 3' is a proto-oncogene known to be associated with PCa, it is also involved in the *TMPRSS2:ERG* fusion but has been shown to be increased in PCa via alternate mechanisms to the fusion also[281]. The *TMPRSS2:ERG* fusion is identified in ~50% of PCa samples but has not been identified as a key biomarker for PCa prediction in cell data in this study. *HOXC6* is known to be associated with PCa, there is a urine based test that utilises the identification of *HOXC6* mRNA called the SelectMDx[276]. Therefore, our findings support other work showing its association with prostate cancer and its identification in PCa patient urine. *NAALADL2* is known to be overexpressed in PCa tissue compared to benign tissue using IHC. Expression of *NAALADL2* has been shown to impact on a number of pro-oncogenic pathways such as cell migration, invasion and colony-forming potential. Leading to the belief that *NAALADL2* is a useful biomarker for diagnosis

and prognosis[282]. In PCa cell urine, a lower expression has been associated with PCa, here. *TDRD* is a cancer/testis (CT) antigen that has previously been associated with liver cancer[283] and breast cancer[284] but not PCa. *SIM2* long has been found to be over-expressed in PCa tissue when compared with CB tissue[285] and it's up regulation has also been associated with biochemical recurrence post-radical prostatectomy[286]. However, it has not previously been identified as a PCa urine biomarker.

In the EV models, *ERG* 3' exon 4-5 was more highly selected as a biomarker over *ERG* 3' 6-7 like in the cell models. *HOXC6*, *TDRD* also appear in all of the EV models. However, *TMPRSS2:ERG*, *PCA3* also appear in EV models and not cell models. It has previously been observed that most of the RNA content in whole urine is actually coming from EVs and not from cells – this is shown by a comparison of the RNA yields from cells and EVs from the same urine samples (data not shown). Further to this, NanoString analysis was only performed on 95 cell RNA fractions out of the 756 Samples because amounts of cell RNA were on the whole so limiting that expression analysis was not deemed viable. Consistently higher EV RNA yields explains how *TMPRSS2:ERG* and *PCA3* are highly detectable in whole urine and EV fractions of urine. *CADPS* is not selected in EV models, which makes sense as *CADPS* is an EV making gene. However, many more transcripts are commonly selected such as *APOC1*, *KLK4* and *HPN*. Showing EVs are a good source of urinary biomarkers for PCa.

Comparing the expression of transcripts in the cell fraction to the EV fraction via microarray has shown that a high proportion of prostate, and PCa associated transcripts are more abundant in the EV fraction. It also showed that *PTPRC,* which is a blood immune associated transcript is more abundant in the cell fraction. This us to believe that many of the cells in the cell fraction of the PCa patient urine are actually immune cells and not prostate/PCa cells, which is in support of other literature[281].

Our findings support previous research that the genetic content of cell sediment and that of extracellular vesicles differs. Expression levels of many transcripts that were both expressed in the prostate tissue and known to be prostate cancer associated were found in

increased levels in the extracellular vesicle compared to the cell sediment. This highlights that the extracellular vesicle fraction is indeed of great interest to investigate further for PCa biomarkers.

7

Discussion

6.7 Summary

Prostate Cancer (PCa) is a major clinical problem worldwide with considerable variability in clinical outcome of patients. PCa diagnostics and prognostics currently lack specific and sensitive clinical biomarkers and treatment is not well individualised. The *PCA3* test, amongst others, highlights the utility of urine in PCa diagnostics and prognostics[214]. The extracellular vesicle (EV) fraction contains exosomes and is obtainable from urine. Exosome levels are known to be increased during malignancy and those produced by tumours contain nucleic material from malignant cells[104]. EVs from tumour cells have roles involved in tumourigenesis, metastasis, and response to therapy by triggering signalling cascades and transferring mRNA, miRNA and proteins between cancer cells and the tumour microenvironment[105]. Our aim was to interrogate PCa patient's urine samples, mostly the EV fraction to identify novel biomarkers or sets of biomarkers to aid in PCa management. This study was completed as part of the Movember GAP1 global PCa biomarker initiative, which involved multiple collaborators and samples collected from four different centres worldwide, for the identification of urinary biomarkers for the risk-stratification of PCa.

7.1.1. Chapter 3: NanoString Data Analysis 1: The Pilot Study

In a pilot study, NanoString technology was able to detect PCa specific markers in 196 samples, such as *TMPRSS2:ERG,* which was detected in 58% of all PCa samples and in 19% of samples from men with no clinical evidence of PCa (CB). Latent Process Decomposition unsupervised analysis clustered the EV expression data into four groups, which was associated with clinical risk categories ($p < 0.05$). Transcripts were identified that were differentially expressed and models were built that could distinguish between PCa and samples that showed no evidence of PCa (CB) with an AUC of 0.937, high-risk PCa and samples showing no evidence of PCa (CB) with an AUC of 0.852 and metastatic PCa (A) and samples showing no evidence of PCa (CB) with an AUC of 0.983. These findings highlight that the transcript data collected from urinary EVs in PCa patients comes, at least in part, from the prostate and holds clinically relevant structure.

6.8 Chapter 4: NanoString2 Analysis: The Movember GAP1 Project

Following on from the pilot study, further samples ($n = 756$) obtained from four centres worldwide were sent to NanoString for the quantification of 167 transcripts. The aims were to primarily identify optimal models capable of predicting PCa and to risk-stratify PCa without the need for biopsy. Models were built to answer four important clinical questions:

1) Determine which samples were from PCa and which were from samples with no evidence of Ca (AUC = 0.851).

2) Determine which samples were from high-risk PCa only and which were from samples with no evidence of cancer, (AUC = 0.897).

3) Determine if there was a trend in expression that corresponds to a trend in risk category (CB>L>I>H), (AUC = 0.709).

4) Determine if there was a trend in expression that corresponds to a trend in patient type (CB>Ca>Metastatic cancer), (AUC = 0.6469).

The data was stratified into training and test sets in the ratio 2:1, models were built with the training set and validated using the test set. I used four different normalisations of the data, which included using *KLK2* ratio, *KLK2* adjusting, *KLK3* adjusting, and *GAPDH* & *RPLP2* normalisation. Models built using the *GAPDH* & *RPLP2* normalised data generally had higher AUCs. These models are improvements on existing tests and have the potentially to be developed in to clinical tests.

6.8.1 Chapter 5: Response to treatment

Many cancers have benefitted from treatment stratification due to expression of certain genes, however with the exception of the DESNT poor prognosis expression group, this has not yet been done for PCa. With hormone therapy (HT) it is known that patients will inevitably progress to castration resistant prostate cancer (CRPC). How long each patient will last on HT varies widely from months to years. Samples from the advanced patients in the NanoString pilot study (n = 32) were used to identify a significant predictor of early progression in patients on HT: A signature of seven transcripts was identified that could optimally predict progression of patients on hormone therapy (cox-regression model; p = $2.3x10^{-05}$; HR = 0.04288). The transcripts in the predictor were *AGR2*, *DLX1*, *KLK2*, *NAALADL2*, *AR* exons 4-8, *PPAP2A* and *AMACR*. This model was an independent predictor of progression when established clinical variables initial PSA, age, Gleason score and initial bone scan result were taken into account (cox-regression model; p = 0.003; HR = 0.03). When the data was adjusted to *KLK2* levels, similar to *KLK3* adjustment used in the PCA3 test, an optimal model of three transcripts (*CAMKK2*, *PSGR* and *UPK)* was identified (cox-regression model; p = 0.007*,* *HR* = 1.0028). This model does not remain significant predictor when adding clinical factors (cox-regression model; p = 0.14; HR= 1.009). However when both of these models were applied to the second NanoString dataset but they were not validated. Despite this, I have shown the potential of using urine extracellular vesicles from prostate cancer patients with NanoString measurements of expression to

predict patient response to treatments. A larger cohort with longer follow up would be required to further develop these models in to something usable in the clinic.

6.8.2 Chapter 6: Analysis of Cell Fraction and comparison with EV fraction

The transcriptome profiles of cell sediment and EV fractions were compared from PCa patients and controls (taken from patients with no evidence of cancer (CB)). Data from microarray of samples collected from NNUH, Norwich and Royal Marsden Hospital, London was used for this comparison. 98 genes were found to be significantly ($p < 0.05$) up-regulated in the extracellular vesicles and 116 up-regulated in the cell sediment fraction. 92 samples from the NanoString 2 experiment were also EV and cell sediment matched and were also used to compare transcriptome profiles. 100 genes were found to be significantly ($p < 0.05$) up-regulated in the extracellular vesicles and 42 genes were up-regulated in the cell sediment fractions. The top twenty of each set of these genes were investigated for known prostate expression and PCa association. The EV fraction contained higher levels of prostate expressed and PCa associated transcripts. This is a strong indication that the EVs contain RNA from prostate cancer cells and it is a better source of biomarkers than the cell fraction.

The NanoString data from cell sediment was used to produce models able to predict PCa (low, intermediate and high-risk) from CB samples, high-risk PCa from CB samples and trend in expression across clinical category. These models had similar AUCs in the training set to the EV fractions but we were unable to validate them at this stage. The power of these cell fraction models is also reduced due to a much lower sample size.

6.9 Discussion

There is an urgent clinical need for biomarkers to determine which patients have PCa, which patients have disease that will progress rapidly, and to individualise treatment to optimise response. Lung cancer and breast cancer are already benefitting from individualised treatment based on expression levels[10,14]. For PCa, stratification models have been produced that include a number of clinical factors, these include D'Amico, and nomograms or points systems such as CAPRA[287], the Prostate Health Index[288], the European Randomised Study of Screening for Prostate Cancer (ERSPC) Risk Calculator[289], the Prostate Cancer Prevention Trial Risk Calculator (PCPT-RC)[212]. However, apart from D'Amico, none of these risk calculators are in general use in the clinic, and effectiveness varies with the cohort[290]. The production of the PCA3 test[214] has led to an increase in studies investigating urine as a source of PCa biomarkers for clinical tests that may prevent unnecessary biopsies. Further research has merged clnical data with urine expressin information such as MiPS[213], which built on the PCA3 urine test to include other urine expression data (*TMPRSS2:ERG*) and PSA. There is also a model for predicting high-grade PCa using *HOXC6* and *DLX1* urinary expression levels along with clinical factors such as prostate volume[215], the ExoDx Prostate (IntelliScore "EPI")[216] which can be used in conjunction with clinical data, and the PCRT-RC which is designed to incorporate future biomarker information as it becomes available. *TMPRSS2:ERG* fusions are only found in ~50% of PCa tumours and PCA3 is not expressed in all PCa tumours also. A panel of more transcripts may improve the diagnostic and prognostic abilities of these tests.

EVs have been investigated as a source of urinary biomarkers in renal cancer studies and it was found that the RNA profile was better preserved in urinary

microvesicles compared with whole cells[217]. It has been suggested that this is because the EV membrane may protect RNA from degradation in urine[217]. Identification of PCa biomarkers in EVs was subsequently observed[218] .

Biomarker discovery in urine of prostate cancer patients has so far focused on just a few gene targets. In this study we have taken a more holistic, but not transcriptome wide approach increasing the number of probes significantly. I have shown that NanoString is a viable technology to measure 100s of probes in urine efficiently and a viable solution for biomarker discovery and potential implementation in the clinic. I have produced potentially important combinations of biomarkers to predict prostate cancer, aggressive prostate cancer, and response to hormone therapy treatment. These gave AUCs up to 0.897, which is an improvement on published tests in the literature. The translational appeal of NanoString analysis can be seen in the ProSigna PAM50 test for aggressive breast cancer[291], which uses NanoString technology and is commercially available.

In our study I have shown that the transcriptome profile of whole cells and EVs differs and that EVs are a potential better source of PCa biomarkers as they contain more prostate derived transcripts as well as more PCa associated and cancer associated transcripts. This indicates that using EVs in biomarker discovery in urine will improve results, but it is likely that whole urine could be used in a final test. Biologically it is likely that EVs can find their way into the urine more easily than the bulky cell counterparts. We also observed that the whole urine includes many white blood cells. Recent research has shown that WBC can be utilised as prognostic markers in BCa showing capability for predicting distant metastasis preoperatively over a 65-month timeline. Increased platelet indices and decreased neutrophil numbers were associated with a poorer prognosis[292].

I have identified urinary EV models from NanoString data capable of predicting PCa, and PCa risk categories with AUCs similar to previously published urine models. Which include

both known PCa associated transcripts from whole urine and novel transcripts that may be EV specific. The AUCs of cell models and EV models are very similar and thus it may be that a combinatory model could be better to predict PCa and its prognosis.

The need for cancer specific biomarkers for assessing response to hormonal treatments in metastatic PCa has been acknowledged[293], yet very little work appears to have been completed in this area. I identified two signatures capable of distinguishing early relapse to HT in two different data normalisations. However these signatures were not validated in a second dataset.

A urine test would aid clinicians and patients for the management of PCa in a few areas. Firstly, there is a decision of whether a biopsy needs to be undertaken. Usually, this is based on serum PSA level and DRE findings. A urine test could help limit the amount of unnecessary biopsies conducted. Secondly, it is known that biopsies generally under grade the PCa, and higher Gleason scores are identified on whole prostates from radical prostatectomies. Therefore, a urine test may help to identify which patients can safely go on to active surveillance.

A third area where a urine test could aid in the clinic is alongside MRIs. MRIs have shown great potential in the diagnosis of PCa but does suffer from a high false positive rate (~50%)[294]. Introduction of a urine test alongside MRI could help to reduce the false positive rate especially for PIRADS ≤ 4[295] (Figure 0.1).

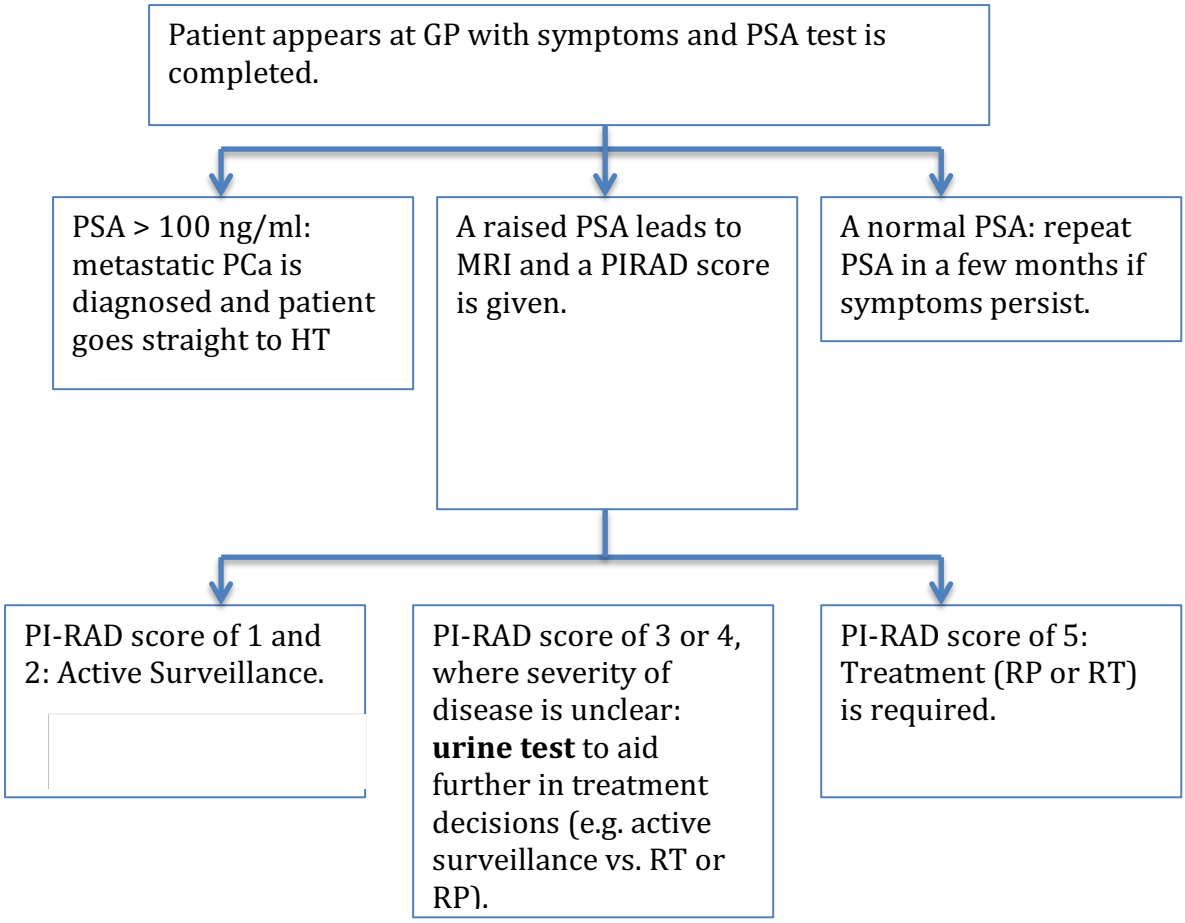


Figure 0.1 A flow diagram showing where the urine test would be best utilised in current diagnostic procedures.

6.10 Future Work

Further work needs to be performed on the models produced for determining between PCa and CB samples, high-risk PCa and CB samples as well as risk category (CB>L>I>H) and patient type (CB> cancer > metastatic cancer) trends. In particular, an immediate next step should be that they are incorporated with clinical factors to identify if they outperform clinical factors alone, as well as if they have a better prediction when including the clinical factors (including but not limited to prostate volume, age, family history, previous biopsy results and serum PSA). The next major step would be to validate these models in an independent large-scale trial such as PROMIS or PROTECT. If this was successful, then the

models would be evaluated in a large multi-centre prospective study. This is necessary to obtain FDA approval and translation in to a test used in the clinic.

Additional work to optimise the methodology used to collect urine, to standardise it, (simplify and make it more robust). Currently, samples have to be processed within 2 hours of collection, the introduction of urine preservatives could also streamline procedures. The models would need to be tested to see if they worked in whole urine and without DRE, which would make the collection and processing methodology a lot simpler. Comparing alternative methods for the quantification of transcripts from urinary EVs may also help to improve the reliability and clinical use of the models.

Further work needs to be completed to identify a robust and validated signature for the prediction of early relapse to CRPC. This is a vital area that needs improvement for the clinical management of PCa. A larger cohort with longer follow up is required. I would also like to look at the data from patients on active surveillance in the NanoString 2 data set. There is considerable potential to develop a predictor of time to treatment in these patients. Another response to treatment that should be investigated is biochemical recurrence (BCR) after radical prostatectomy or radiation therapy. Unfortunately, our follow up was not long enough to have sufficient numbers of patients that suffered from BCR to be able to perform any of these experiments at this time. I would also like to examine whether models that were developed for the prediction of aggressiveness could be applied to predict response to treatment. For example, could the optimal model for predicting risk category also be used to predict time to treatment for patients on active surveillance.

In this whole project I have been reliant on the 167 gene probes used in the NanoString assay. It is not clear whether these are the optimal probes to use,

although it is apparent that they are at least sufficient for some clinical questions. I would like to perform a similar scale project but using a global transcriptome approach using microarrays, or for an exon splice variant analysis then RNAseq would be ideal, though highly analysis intensive. This would allow us to identify the very best probes to use in a clinical test to answer the important clinical questions in prostate cancer.

Due to my work in this thesis re-funding has been awarded for the development of a clinically implementable Prostate Urine Risk test. This has resulted in two further PhD posts one for lab work and one for bioinformatics.

6.11 Conclusions

O1: To determine whether RNA expression from urine extracellular vesicles in prostate cancer patients are a viable target for the development of biomarkers through the use of NanoString technology.

I have shown that urine extracellular vesicles from prostate cancer patients contain information from tumours and are a viable area to investigate for non-invasive biomarkers. I have shown that NanoString technology is sensitive and specific enough to use as a semi-high throughput approach for discovery and potentially for clinical use.

O2: To determine an optimal combination of probes to predict cancer presence and aggression in prostate cancer patients.

I have determined a number of models that work extremely well in predicting both cancer presence and the aggressiveness of disease. These have the potential, with further work, to have an impact in the clinic. Models to accurately stratify patients' disease into D'Amico risk groupings were less satisfactory and may require alternative probes or other techniques.

O3: To determine whether an optimal combination of probes can predict response to hormone therapy treatment.

I have shown that there may be some information in urine extracellular vesicles to predict patient response to treatments. I have developed some potential tests, but for confidence in these a much bigger data set with longer follow up would be required.

O4: To evaluate the differences between urine fractions (extracellular vesicles and cell sediment) and determine whether cell sediment can be used to predict cancer presence and aggression in prostate cancer patients.

I have shown that there are considerable differences between the extracellular vesicles fraction and the cell sediment fraction of urine collected from prostate cancer patients. There is a strong indication that the EVs contain more RNA from prostate cancer cells and it is a better source of biomarkers than the cell fraction. Despite this, I was able to produce some models that were reasonable good at detecting the presence and aggressiveness of prostate cancer.

In this thesis, I have shown that by interrogating the EV faction of PCa patient's urine samples using NanoString technology that novel biomarkers or sets of biomarkers can be identified to aid in PCa management in a non-invasive test.

8

References

1. Leongamornlert, D. *et al.* Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease. **110,** 1663–1672 (2014).
2. Akin, O. *et al.* Transition zone prostate cancers: features, detection, localization, and staging at endorectal MR imaging. *Radiology* **239,** 784–792 (2006).
3. Olson, S., Robinson, S. & Giffin, R. *Accelerating the Development of Biomarkers for Drug Safety: Workshop Summary*. *Sciences-New York* (2009). doi:20464768
4. World Health Organisation. WHO International Programme on Chemical Safety Biomarkers and Risk Assessment: Concepts and Principles. (1993).
5. Strimbu, K. & Tavel, J. A. What are biomarkers? *Curr. Opin. HIV AIDS* **5,** 463–466 (2010).
6. Fuzery, A. K., Levin, J., Chan, M. M. & Chan, D. W. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. *Clin. Proteomics* **10,** 13 (2013).
7. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.* **72,** 1117–1130 (2003).
8. Markert, E. K., Mizuno, H., Vazquez, A. & Levine, A. J. Molecular classi

fi cation of prostate cancer using curated expression signatures. (2011). doi:10.1073/pnas.1117029108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1117029108

9. Heinemann, V., Stintzing, S., Kirchner, T., Boeck, S. & Jung, A. Clinical relevance of EGFR- and KRAS-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR. *Cancer Treat. Rev.* **35,** 262–271 (2009).
10. Rosell, R., Bivona, T. G. & Karachaliou, N. Genetics and biomarkers in personalisation of lung cancer treatment. *Lancet* **382,** 720–31 (2013).
11. Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti EGFR therapy in colorectal cancer. *Nature* **486,** 532–536 (2014).
12. Diamandis, E. P. Cancer Biomarkers: Can We Turn Recent Failures into Success? *JNCI J. Natl. Cancer Inst.* **102,** 1462–1467 (2010).
13. Sawyers, C. L. The cancer biomarker problem. *Nature* **452,** 548–552 (2008).
14. Eaton, A. A. *et al.* Estimating the OncotypeDX score : validation of an inexpensive estimation tool. *Breast Cancer Res. Treat.* (2016). doi:10.1007/s10549-016-4069-4
15. Marrone, M., Potosky, A. L., Penson, D. & Freedman, A. N. A 22 Gene-expression Assay, Decipher® (GenomeDx Biosciences) to Predict Five-year Risk of Metastatic Prostate Cancer in Men Treated with Radical Prostatectomy. *PLoS Curr.* **7,** ecurrents.eogt.761b81608129ed61b0b48d42c04f92a4 (2015).
16. Cuzick, J. *et al.* Prognostic value of a cell cycle progression signature for prostate cancer death in a conservatively managed needle biopsy cohort. 1095–1099 (2012). doi:10.1038/bjc.2012.39
17. Cancer Research UK. Molecular diagnostic provision in England for targeted cancer medicines (solid tumours) in the NHS. (2015).
18. Olzscha, H., New, M. & La Thangue, N. B. Personalised Cancer Medicine: Fulfilling the Promise. *Encycl. Life Sci.* 1–10 (2013). doi:10.1002/9780470015902.a0025180
19. Mitri, Z., Constantine, T. & O'Regan, R. The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. *Chemother. Res. Pract.* **2012,** 1–7 (2012).
20. Crown, J., O'Shaughnessy, J. & Gullo, G. Emerging targeted therapies in triple-negative breast cancer. *Ann. Oncol.* **23,** vi56-vi65 (2012).
21. Giaccone, G. & Rodriguez, J. A. EGFR inhibitors: what have we learned from the treatment of lung cancer? *Nat. Clin. Pract. Oncol.* **2,** 554–61 (2005).
22. Dziadziuszko, R. & Jassem, J. Epidermal growth factor receptor (EGFR) inhibitors and derived treatments. *Ann. Oncol.* **23 Suppl 1,** x193-6 (2012).
23. Velonas, V. M., Woo, H. H., dos Remedios, C. G. & Assinder, S. J. Current status of biomarkers for prostate cancer. *Int. J. Mol. Sci.* **14,** 11034–11060 (2013).
24. Ferlay, J. *et al.* Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *Eur. J. Cancer* **49,** 1374–1403

(2013).
25. Lloyd, T. *et al.* Lifetime risk of being diagnosed with, or dying from, prostate cancer by major ethnic group in England 2008–2010. *BMC Med.* **13,** 171 (2015).
26. UK, C. R. Cancer Research UK. *Prostate cancer statistics* (2014). Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer. (Accessed: 1st January 2015)
27. Oh WK, Hurwitz M, D. A. *Biology of Prostate Cancer. Hollan-Frei Cancer Medicine.* (BC Decker, 2003).
28. Cancer Research UK. Cancer Research UK. *Prostate Cancer: Types and grades* (2016). Available at: http://www.cancerresearchuk.org/about-cancer/prostate-cancer/types-grades. (Accessed: 1st January 2016)
29. UK, C. R. Cancer Research UK. *Prostate cancer statistics* (2014).
30. Vellekoop, A. & Loeb, S. More Aggressive Prostate Cancer in Elderly Men. *Rev. Urol.* **15,** 202–204 (2013).
31. Gann, P. H. Risk factors for prostate cancer. *Rev. Urol.* **4 Suppl 5,** S3–S10 (2002).
32. Madu, C. O. & Lu, Y. Novel diagnostic biomarkers for prostate cancer. *J. Cancer* **1,** 150–177 (2010).
33. Ito, K. Prostate cancer in Asian men. *Nat Rev Urol* **11,** 197–212 (2014).
34. Ito, K. Prostate cancer in Asian men. *Nat Rev Urol* **11,** 197–212 (2014).
35. Eeles, R. *et al.* The genetic epidemiology of prostate cancer and its clinical implications. *Nat Rev Urol* **11,** 18–31 (2014).
36. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46,** 1103–9 (2014).
37. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46,** 1103–9 (2014).
38. Catalona, W. J. *et al.* Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men. *J. Urol.* **151,** 1283–1290 (1994).
39. Humphrey, P. A. Gleason grading and prognostic factors in carcinoma of the prostate. *Mod. Pathol.* **17,** 292–306 (2004).
40. National Collaborating Centre for Cancer. Prostate Cancer : diagnosis and treatment. Clinical guideline. *Natl. Inst. Heal. Care Excell.* 1–480 (2014).
41. (Uk), N. C. C. for C. Prostate Cancer: Diagnosis and Treatment. *Natl. Inst. Heal. Clin. Excell. Guid.* 2007–2009 (2008).
42. Paller, C. J. & Antonarakis, E. S. Management of biochemically recurrent prostate cancer after local therapy: evolving standards of care and new directions. *Clin. Adv. Hematol. Oncol.* **11,** 14–23 (2013).
43. Paller, C. J. & Antonarakis, E. S. Management of biochemically recurrent prostate cancer after local therapy: evolving standards of care and new directions. *Clin. Adv. Hematol. Oncol.* **11,** 14–23 (2013).
44. Wan, X. *et al.* UHRF1 overexpression is involved in cell proliferation

and biochemical recurrence in prostate cancer after radical prostatectomy. *J. Exp. Clin. Cancer Res.* **35,** 34 (2016).
45. Raatikainen, S., Aaaltomaa, S., Karja, V. & Soini, Y. Increased Peroxiredoxin 6 Expression Predicts Biochemical Recurrence in Prostate Cancer Patients After Radical Prostatectomy. *Anticancer Res.* **35,** 6465–6470 (2015).
46. Qu, X. *et al.* Identification of Combinatorial Genomic Abnormalities Associated with Prostate Cancer Early Recurrence. *J. Mol. Diagn.* **18,** 215–224 (2016).
47. Ma, D. *et al.* Association of molecular biomarkers expression with biochemical recurrence in prostate cancer through tissue microarray immunostaining. *Oncol. Lett.* **10,** 2185–2191 (2015).
48. Meng, Y., Li, H., Xu, P. & Wang, J. Do tumor volume , percent tumor volume predict biochemical recurrence after radical prostatectomy ? A meta-analysis. **8,** 22319–22327 (2015).
49. Lorente, J. A., Morote, J., Raventos, C., Encabo, G. & Valenzuela, H. Clinical efficacy of bone alkaline phosphatase and prostate specific antigen in the diagnosis of bone metastasis in prostate cancer. *J. Urol.* **155,** 1348–1351 (1996).
50. (Uk), N. C. C. for C. Prostate Cancer: Diagnosis and Treatment. *Natl. Inst. Heal. Clin. Excell. Guid.* 2007–2009 (2008).
51. Perlmutter, M. A. & Lepor, H. Androgen deprivation therapy in the treatment of advanced prostate cancer. *Rev. Urol.* **9 Suppl 1,** S3-8 (2007).
52. Mostaghel, E. a. Abiraterone in the treatment of metastatic castration-resistant prostate cancer. *Cancer Manag. Res.* **6,** 39–51 (2014).
53. Schoenborn, J. R., Nelson, P. & Fang, M. Genomic profiling defines subtypes of prostate cancer with the potential for therapeutic stratification. *Clin. Cancer Res.* **19,** 4058–66 (2013).
54. Mostaghel, E. a. Abiraterone in the treatment of metastatic castration-resistant prostate cancer. *Cancer Manag. Res.* **6,** 39–51 (2014).
55. Miyamoto, H., Messing, E. M. & Chang, C. Androgen deprivation therapy for prostate cancer: current status and future prospects. *Prostate* **61,** 332–353 (2004).
56. Chen, C. D. *et al.* Molecular determinants of resistance to antiandrogen therapy. *Nat. Med.* **10,** 33–39 (2004).
57. Nice. Single Technology Appraisal (STA) Cabazitaxel for the second-line treatment of metastatic hormone refractory prostate cancer. 1–164 (2011).
58. Mikolajczyk, S. D. *et al.* A Precursor Form of Prostate-specific Antigen Is More Highly Elevated in Prostate Cancer Compared with Benign Transition Zone Prostate Tissue A Precursor Form of Prostate-specific Antigen Is More Highly Elevated in Prostate Cancer Compared with Benign Tra. 756–759 (2000).
59. Thompson, Ian M, Lucia, M. S. *et al.* new england journal. 2239–2246 (2004).
60. DeAntoni, E. P. *et al.* Age- and race-specific reference ranges for prostate-specific antigen from a large community-based study. *Urology* **48,** 234–239 (1996).

61. Romero Otero, J., Garcia Gomez, B., Campos Juanatey, F. & Touijer, K. a. Prostate cancer biomarkers: An update. *Urol. Oncol. Semin. Orig. Investig.* **32,** 252–260 (2014).
62. National Cancer Institute (NCI). Prostate-Specific Antigen (PSA) Test. *Prostate-Specific Antigen (PSA) Test* (2012). Available at: https://www.cancer.gov/types/prostate/psa-fact-sheet. (Accessed: 1st January 2015)
63. Tinzl, M., Marberger, M., Horvath, S. & Chypre, C. DD3PCA3 RNA analysis in urine--a new perspective for detecting prostate cancer. *Eur. Urol.* **46,** 182–6; discussion 187 (2004).
64. Catalona, W. J. *et al.* Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. *JAMA* **279,** 1542–1547 (1998).
65. Mikolajczyk, S. D., Marks, L. S., Partin, A. W. & Rittenhouse, H. G. Free prostate-specific antigen in serum is becoming more complex. *Urology* **59,** 797–802 (2002).
66. Mikolajczyk, S. D. *et al.* A Truncated Precursor Form of Prostate-specific Antigen Is a More Specific Serum Marker of Prostate Cancer A Truncated Precursor Form of Prostate-specific Antigen Is a More Specific Serum Marker of Prostate Cancer. 6958–6963 (2001).
67. Guazzoni, G. *et al.* Prostate-specific antigen (PSA) isoform p2PSA significantly improves the prediction of prostate cancer at initial extended prostate biopsies in patients with total PSA between 2.0 and 10 ng/ml: results of a prospective study in a clinical setting. *Eur. Urol.* **60,** 214–22 (2011).
68. Stephan, C. *et al.* A [-2]proPSA-based artificial neural network significantly improves differentiation between prostate cancer and benign prostatic diseases. *Prostate* **69,** 198–207 (2009).
69. National Cancer Institute (NCI). Prostate-Specific Antigen (PSA) Test. *Prostate-Specific Antigen (PSA) Test* (2012).
70. Makarov, D. V *et al.* Management for Prostate Cancer. **15,** 7316–7321 (2010).
71. Wang, Y., Liu, X.-J. & Yao, X.-D. Function of PCA3 in prostate tissue and clinical research progress on developing a PCA3 score. *Chin. J. Cancer Res.* **26,** 493–500 (2014).
72. Sokoll, L. J. *et al.* A multicenter evaluation of the PCA3 molecular urine test: pre-analytical effects, analytical performance, and diagnostic accuracy. *Clin. Chim. Acta.* **389,** 1–6 (2008).
73. Hessels, D. *et al.* Detection of TMPRSS2-ERG fusion transcripts and prostate cancer antigen 3 in urinary sediments may improve diagnosis of prostate cancer. *Clin. Cancer Res.* **13,** 5103–8 (2007).
74. van Gils, M. P. M. Q. *et al.* The time-resolved fluorescence-based PCA3 test on urinary sediments after digital rectal examination; a Dutch multicenter validation of the diagnostic performance. *Clin. Cancer Res.* **13,** 939–43 (2007).
75. Schmidt, U. *et al.* Quantitative Multi-Gene Expression Profiling of Primary Prostate Cancer. **1534,** (2006).
76. Gittelman, M. C. *et al.* PCA3 molecular urine test as a predictor of

repeat prostate biopsy outcome in men with previous negative biopsies: a prospective multicenter clinical study. *J. Urol.* **190,** 64–69 (2013).
77. Manuscript, A., Variants, A. R. S. & Tissues, M. P. Expression in Normal and Malignant Prostate Tissues. **77,** 1–12 (2012).
78. Sambasivarao, S. V. NIH Public Access. **18,** 1199–1216 (2013).
79. Ozgur, T., Atik, E., Hakverdi, S. & Yaldiz, M. The expressions of {AMACR} and {iNOS} in prostate adenocarcinomas. *Pakistan J. Med. Sci.* **29,** 610–613 (2013).
80. Lloyd, M. D., Darley, D. J., Wierzbicki, A. S. & Threadgill, M. D. Alpha-methylacyl-CoA racemase--an 'obscure' metabolic enzyme takes centre stage. *FEBS J.* **275,** 1089–102 (2008).
81. Rubin, M. a *et al.* Decreased alpha-methylacyl CoA racemase expression in localized prostate cancer is associated with an increased rate of biochemical recurrence and cancer-specific death. *Cancer Epidemiol. Biomarkers Prev.* **14,** 1424–1432 (2005).
82. Ware, K. E., Garcia-Blanco, M. a., Armstrong, A. J. & Dehm, S. M. Biologic and clinical significance of androgen receptor variants in castration resistant prostate cancer. *Endocr. Relat. Cancer* **21,** 87–103 (2014).
83. Geng, C. *et al.* Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 6997–7002 (2013).
84. An, J., Wang, C., Deng, Y., Yu, L. & Huang, H. Destruction of full-length androgen receptor by wild-type SPOP, but not prostate-cancer-associated mutants. *Cell Rep.* **6,** 657–669 (2014).
85. Schoenborn, J. R., Nelson, P. & Fang, M. Genomic profiling defines subtypes of prostate cancer with the potential for therapeutic stratification. *Clin. Cancer Res.* **19,** 4058–66 (2013).
86. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44,** 685–9 (2012).
87. Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44,** 685–9 (2012).
88. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487,** 239–43 (2012).
89. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487,** 239–43 (2012).
90. Hossain, D. & Bostwick, D. G. Significance of the TMPRSS2:ERG gene fusion in prostate cancer. *BJU Int.* **111,** 834–5 (2013).
91. Clark, J. P. & Cooper, C. S. ETS gene fusions in prostate cancer. *Nat. Rev. Urol.* **6,** 429–39 (2009).
92. Clark, J. P. & Cooper, C. S. ETS gene fusions in prostate cancer. *Nat. Rev. Urol.* **6,** 429–39 (2009).
93. Li, L. *et al.* Targeting Poly(ADP-Ribose) Polymerase and the c-Myb-Regulated DNA Damage Response Pathway in Castration-Resistant Prostate Cancer. *Sci. Signal.* **7,** ra47 (2014).
94. Bisen, A. & Claxton, D. F. Tyrosine kinase targeted treatment of

chronic myelogenous leukemia and other myeloproliferative neoplasms. *Adv. Exp. Med. Biol.* **779,** 179–196 (2013).
95. Tomlins, S. a. *et al.* Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer. *Neoplasia* **10,** 177-IN9 (2008).
96. Tomlins, S. a. *et al.* Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer. *Neoplasia* **10,** 177-IN9 (2008).
97. Shao, L. *et al.* Highly specific targeting of the TMPRSS2/ERG fusion gene using liposomal nanovectors. *Clin. Cancer Res.* **18,** 6648–57 (2012).
98. Goh, C. L. *et al.* Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J. Intern. Med.* **271,** 353–65 (2012).
99. Goh, C. L. *et al.* Genetic variants associated with predisposition to prostate cancer and potential clinical implications. *J. Intern. Med.* **271,** 353–65 (2012).
100. Levy-Lahad, E. & Friedman, E. Cancer risks among BRCA1 and BRCA2 mutation carriers. *Br. J. Cancer* **96,** 11–15 (2007).
101. Drake, Richard, Vogl, Wayne & Mitchell, A. *Gray's Anatomy for Students*. (Elsevier Inc., 2015).
102. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized , multifocal prostate cancer. (2015). doi:10.1038/ng.3315
103. Bátiz, L. F. *et al.* Exosomes as Novel Regulators of Adult Neurogenic Niches . *Frontiers in Cellular Neuroscience* **9,** 501 (2016).
104. Nilsson, J. *et al.* Prostate cancer-derived urine exosomes : a novel approach to biomarkers for prostate cancer. **100,** 1603–1607 (2009).
105. Lin, J. *et al.* Exosomes : Novel Biomarkers for Clinical Diagnosis. *Sci. World J.* **2015,** (2015).
106. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).
107. Azmi, A. S., Bao, B. & Sarkar, F. H. Exosomes in cancer development, metastasis, and drug resistance: A comprehensive review. *Cancer Metastasis Rev.* **32,** 623–642 (2013).
108. Hoshino, A. *et al.* Tumour exosome integrins determine organotropic metastasis. *Nature* **527,** 329–335 (2015).
109. Costa-silva, B. *et al.* Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. **17,** (2015).
110. Skog, J. *et al.* Glioblastoma microvesicles transport RNA and protein that promote tumor growth and provide diagnostic biomarkers. *Nat. Cell Biol.* **10,** 1470–1476 (2008).
111. Nilsson, J. *et al.* Prostate cancer-derived urine exosomes : a novel approach to biomarkers for prostate cancer. **100,** 1603–1607 (2009).
112. Dijkstra, S. *et al.* Prostate cancer biomarker profiles in urinary sediments and exosomes. *J. Urol.* **191,** 1132–1138 (2014).
113. Hu, L. *et al.* Fluorescence in situ hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomark. Res.* **2,** 3 (2014).
114. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.* **26,** 317–25 (2008).
115. Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression

with color-coded probe pairs. *Nat. Biotechnol.* **26,** 317–25 (2008).
116. NanoString Technologies®. nCounter® Gene Expression CodeSets.
117. Sanger, F., Nicklen, S. & Coulson, a R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).
118. National Human Genome Research Insitute (NIH). The Cost of Sequencing a Human Genome. (2016). Available at: https://www.genome.gov/sequencingcosts/. (Accessed: 1st January 2016)
119. National Human Genome Research Insitute (NIH). The Cost of Sequencing a Human Genome. (2016).
120. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016).
121. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016).
122. Holt, R. A. & Jones, S. J. M. The new paradigm of flow cell sequencing. 839–846 (2008). doi:10.1101/gr.073262.107.cell
123. Fullwood, M. J., Wei, C., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. 521–532 (2009). doi:10.1101/gr.074906.107.Freely
124. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
125. Voelkerding, K. V, Dames, S. a & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* **55,** 641–58 (2009).
126. Voelkerding, K. V, Dames, S. a & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* **55,** 641–58 (2009).
127. Lee, H. *et al.* Third-generation sequencing and the future of genomics. *bioRxiv* 048603 (2016). doi:10.1101/048603
128. Lee, H. *et al.* Third-generation sequencing and the future of genomics. *bioRxiv* 048603 (2016). doi:10.1101/048603
129. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. **59,** 5–15 (2013).
130. Warr, A. *et al.* Exome Sequencing: Current and Future Perspectives. *G3: Genes|Genomes|Genetics* **5,** 1543–1550 (2015).
131. Warr, A. *et al.* Exome Sequencing: Current and Future Perspectives. *G3: Genes|Genomes|Genetics* **5,** 1543–1550 (2015).
132. Wilhelm, B. T. & Landry, J. R. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48,** 249–257 (2009).
133. Zhao, S., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. **9,** (2014).
134. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12,** 87–98 (2011).
135. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing : recent advances and remaining challenges [version 1 ; referees : 2 approved] Referee Status : **5,** (2016).
136. Liu, E. T., Pott, S. & Huss, M. Q & A : ChIP-seq technologies and the study of gene regulation. 4–9 (2010).

137. Mamanova, L. *et al.* Target-enrichment strategies for next- generation sequencing. **7,** (2010).
138. Clark-langone, K. M. *et al.* assay. **18,** 1–18 (2007).
139. Knuutila, S. Biomarker analysis in human neoplasias : superior next-generation sequencing on frozen bone marrow cells and on formalin-fixed , paraffin- embedded tumor tissues. *BMC Proc.* **7,** K18 (2013).
140. Saini, S. PSA and beyond: alternative prostate cancer biomarkers. *Cell. Oncol. (Dordr).* **39,** 97–106 (2016).
141. Murphy, D. Gene expression studies using microarrays: principles, problems, and prospects. *Adv. Physiol. Educ.* **26,** 256–270 (2002).
142. Vermeeren, V. & Michiels, L. *Evolution Towards the Implementation of Point-Of-Care Biosensors*. (INTECH Open Access Publisher, 2011).
143. Kretschmer, A. & Tilki, D. Biomarkers in prostate cancer – Current clinical utility and future perspectives. *Crit. Rev. Oncol. Hematol.* **120,** 180–193 (2017).
144. Dalela, D. *et al.* Genomic Classifier Augments the Role of Pathological Features in Identifying Optimal Candidates for Adjuvant Radiation Therapy in Patients With Prostate Cancer: Development and Internal Validation of a Multivariable Prognostic Model. *J. Clin. Oncol.* **35,** 1982–1990 (2017).
145. Burgess, A., Shah, K., Hough, O. & Hynynen, K. HHS Public Access. **15,** 477–491 (2016).
146. Klein, E. A. *et al.* Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. *Urology* **90,** 148–152 (2016).
147. Graves, P. R. & Haystead, T. A. J. Molecular Biologist ' s Guide to Proteomics. **66,** 39–63 (2002).
148. Semmes, O. J. Defining the Role of Mass Spectrometry in Cancer Diagnostics Defining the Role of Mass Spectrometry in Cancer Diagnostics. 1555–1557 (2004).
149. Kakimoto, Y., Tsuruyama, T., Yamamoto, T., Furuta, M. & Kotani, H. Novel In Situ Pretreatment Method for Significantly Enhancing the Signal In MALDI-TOF MS of Formalin- Fixed Paraffin-Embedded Tissue Sections. **7,** 1–7 (2012).
150. Comparison, P. Validation Processes of Protein Biomarkers in Serum — A Cross Platform Comparison. 12710–12728 (2012). doi:10.3390/s120912710
151. Manuscript, A. targeted mass spectrometry. **10,** 28–34 (2014).
152. Bayani, J. & Squire, J. a. Application and interpretation of FISH in biomarker studies. *Cancer Lett.* **249,** 97–109 (2007).
153. Bayani, J. & Squire, J. a. Application and interpretation of FISH in biomarker studies. *Cancer Lett.* **249,** 97–109 (2007).
154. Hu, L. *et al.* Fluorescence in situ hybridization (FISH): an increasingly demanded tool for biomarker research and personalized medicine. *Biomark. Res.* **2,** 3 (2014).
155. Dunstan, R. W., Wharton, K. a, Quigley, C. & Lowe, A. The use of immunohistochemistry for biomarker assessment--can it compete with other technologies? *Toxicol. Pathol.* **39,** 988–1002 (2011).
156. Dunstan, R. W., Wharton, K. a, Quigley, C. & Lowe, A. The use of immunohistochemistry for biomarker assessment--can it compete

with other technologies? *Toxicol. Pathol.* **39,** 988–1002 (2011).
157. Ferrier, C. M. *et al.* Comparison of immunohistochemistry with immunoassay (ELISA) for the detection of components of the plasminogen activation system in human tumour tissue. *Br. J. Cancer* **79,** 1534–1541 (1999).
158. Sant, K. E., Nahar, M. S. & Dolinoy, D. C. DNA methylation screening and analysis. *Methods Mol. Biol.* **889,** 385–406 (2012).
159. Shen, L. & Waterland, R. A. Methods of DNA methylation analysis. *Curr. Opin. Clin. Nutr. Metab. Care* **10,** 576–581 (2007).
160. Jin, B., Li, Y. & Robertson, K. D. DNA Methylation : Superior or Subordinate in the Epigenetic Hierarchy ? 607–617 (2011). doi:10.1177/1947601910393957
161. Hoque, M. O. & Surgery, N. HHS Public Access. **9,** 243–257 (2015).
162. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel).* **5,** 3 (2016).
163. Habeeb, N. M. A. W. *et al.* - ntegrated analysis of epigenomic and genomic changes by DNA methylation dependent mechanisms provides potential novel biomarkers for prostate cancer. **5,**
164. Yang, M. & Park, J. Y. DNA methylation in promoter region as biomarkers in prostate cancer. *Methods Mol. Biol.* **863,** 67–109 (2012).
165. Murphy, T. M., Perry, A. S. & Lawler, M. The emergence of DNA methylation as a key modulator of aberrant cell death in prostate cancer. (2008). doi:10.1677/ERC-07-0208
166. Goering, W., Kloth, M. & Schulz, W. A. DNA methylation changes in prostate cancer. *Methods Mol. Biol.* **863,** 47–66 (2012).
167. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel).* **5,** 3 (2016).
168. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer US, 2011).
169. Hahne, F., Huber, W., Gentleman, R. & Falcon, S. *Bioconductor Case Studies*. (2008).
170. Knuutila, S. Biomarker analysis in human neoplasias: superior next-generation sequencing on frozen bone marrow cells and on formalin-fixed, paraffin-embedded tumor tissues. *BMC Proc.* **7,** K18–K18 (2013).
171. Protocol, Q. Ovation Pico WTA System ®. *Ovation* 4–7 (2010).
172. Waggott, D. *et al.* NanoStringNorm: an extensible R package for the pre-processing of NanoString mRNA and miRNA data. *Bioinformatics* **28,** 1546–8 (2012).
173. Nickles, A. D., Sandmann, T., Ziman, R. & Bourgon, R. Package ' NanoStringQCPro '. (2018).
174. Waggott, D. *et al.* NanoStringNorm: an extensible R package for the pre-processing of NanoString mRNA and miRNA data. *Bioinformatics* **28,** 1546–8 (2012).
175. Jung, S.-H. & Sohn, I. Statistical Issues in the Design and Analysis of nCounter Projects. *Cancer Inform.* **13,** 35–43 (2014).
176. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods.

Biostatistics **8,** 118–127 (2007).
177. Hart, A. Mann-Whitney test is not just a test of medians: differences in spread can be important. *Bmj* **323,** 391–393 (2001).
178. Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24,** 69–71 (2012).
179. Howell, D. C. Chi-Square Test - Analysis of Contingency Tables. *Test* 1–4 (2000). doi:10.1007/978-3-642-04898-2_174
180. Bland, J. M. & Altman, D. G. The logrank test. *BMJ* **328,** 1073 (2004).
181. Royston, P. Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *J. R. Stat. Soc. Ser. C (Applied Stat.* **44,** 547–551 (1995).
182. Yokoyama, M., Nishi, Y., Yoshii, J., Okubo, K. & Matsubara, K. Identification and cloning of neuroblastoma-specific and nerve tissue-specific genes through compiled expression profiles. *DNA Res.* **3,** 311–320 (1996).
183. Carey, M. V. Package ' ROC '. (2018).
184. Bendix, A. & Carstensen, M. B. Package ' Epi '. (2018).
185. Jeong, D. H., Ziemkiewicz, C., Ribarsky, W. & Chang, R. Understanding Principal Component Analysis Using a Visual Analytics Tool. *Proc. UKC 2009, Math. Fundam. Appl. 2009* 1–10 (2009). doi:10.1.1.157.1469
186. Legendre, P. & Legendre, L. Numerical Ecology, Volume 24. *(Developments Environ. Model.* **24,** 870 (1988).
187. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542 (2006).
188. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Applied Stat.* **28,** 100–108 (1979).
189. Singh, K., Malik, D. & Sharma, N. Evolving limitations in K-means algorithm in data mining and their removal. *IJCEM Int. J. Comput. Eng. Manag. ISSN* **12,** 2230–7893 (2011).
190. Singh, K., Malik, D. & Sharma, N. Evolving limitations in K-means algorithm in data mining and their removal. *IJCEM Int. J. Comput. Eng. Manag. ISSN* **12,** 2230–7893 (2011).
191. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. **NbClust** : An *R* Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **61,** (2014).
192. Rogers, S., Girolami, M., Campbell, C. & Breitling, R. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2,** 143–156 (2005).
193. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).
194. Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B* **58,** 267–288 (1994).
195. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
196. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
197. Zhang, Z. Variable selection with stepwise and best subset

approaches. *Ann. Transl. Med.* **4,** 136 (2016).
198. Zhang, Z. Variable selection with stepwise and best subset approaches. *Ann. Transl. Med.* **4,** 136 (2016).
199. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).
200. Crawley, M. J. *The R book*. (Second edition. Chichester, West Sussex, United Kingdom : Wiley, 2013., 2013).
201. Lumley, T. & S-, R. Package ' survival '. (2018).
202. Schloerke, B. Package ' GGally ' R topics documented : (2018).
203. Ahmed, H. U. *et al.* Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389,** 815–822 (2018).
204. Malkov, V. a *et al.* Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter Assay System. *BMC Res. Notes* **2,** 80 (2009).
205. David, A. *et al.* Unusual alternative splicing within the human kallikrein genes KLK2 and KLK3 gives rise to novel prostate-specific proteins. (2002).
206. Gleason, D. F. & Mellinger, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J. Urol.* **111,** 58–64 (1974).
207. Gordetsky, J. & Epstein, J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn. Pathol.* **11,** 25 (2016).
208. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 8418–23 (2003).
209. Rosell, R. *et al.* Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **13,** 239–246 (2012).
210. Bethune, G., Bethune, D., Ridgway, N. & Xu, Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J. Thorac. Dis.* **2,** 48–51 (2010).
211. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21,** 171–178 (2005).
212. Thompson, I. M. *et al.* Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J. Natl. Cancer Inst.* **98,** 529–534 (2006).
213. Tomlins, S. A. *et al.* Urine TMPRSS2:ERG Plus PCA3 for Individualized Prostate Cancer Risk Assessment. *Eur. Urol.* **70,** 45–53 (2016).
214. Hessels, D. *et al.* DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur. Urol.* **44,** 6–8 (2003).
215. Van Neste, L. *et al.* Detection of High-grade Prostate Cancer Using a Urinary Molecular Biomarker–Based Risk Score. *Eur. Urol.* **70,** 740–748 (2016).
216. McKiernan, J. *et al.* A Novel Urine Exosome Gene Expression Assay to

Predict High-grade Prostate Cancer at Initial Biopsy. *JAMA Oncol.* **2,** 882–889 (2016).
217. Miranda, K. C. *et al.* Nucleic acids within urinary exosomes/microvesicles are potential biomarkers for renal disease. *Kidney Int.* **78,** 191–199 (2010).
218. Dijkstra, S. *et al.* Prostate cancer biomarker profiles in urinary sediments and exosomes. *J. Urol.* **191,** 1132–1138 (2014).
219. Mootha, V. K. *et al.* PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34,** 267–273 (2003).
220. Dhaese, S. *et al.* Functional and profiling studies prove that prostate cancer upregulated neuroblastoma thymosin beta is the true human homologue of rat thymosin beta15. *FEBS Lett.* **581,** 4809–4815 (2007).
221. Yokoyama, M., Nishi, Y., Yoshii, J., Okubo, K. & Matsubara, K. Identification and cloning of neuroblastoma-specific and nerve tissue-specific genes through compiled expression profiles. *DNA Res.* **3,** 311–320 (1996).
222. Theunissen, W. *et al.* Thymosin Beta 4 and Thymosin Beta 10 Expression in Hepatocellular Carcinoma. *Eur. J. Histochem.* **58,** 2242 (2014).
223. Darb-Esfahani, S. *et al.* Thymosin beta 15A (TMSB15A) is a predictor of chemotherapy response in triple-negative breast cancer. *Br. J. Cancer* **107,** 1892–1900 (2012).
224. Lenka, G., Weng, W.-H., Chuang, C.-K., Ng, K.-F. & Pang, S.-T. Aberrant expression of the PRAC gene in prostate cancer. *Int. J. Oncol.* **43,** 1960–1966 (2013).
225. Koestler, D. C. *et al.* Distinct patterns of DNA methylation in conventional adenomas involving the right and left colon. *Mod. Pathol. an Off. J. United States Can. Acad. Pathol. Inc* **27,** 145–155 (2014).
226. Lose, F. *et al.* Genetic association of the KLK4 locus with risk of prostate cancer. *PLoS One* **7,** e44520 (2012).
227. Vaananen, R.-M. *et al.* Association of transcript levels of 10 established or candidate-biomarker gene targets with cancerous versus non-cancerous prostate tissue from radical prostatectomy specimens. *Clin. Biochem.* **46,** 670–674 (2013).
228. Bostwick, D. G., Pacelli, A., Blute, M., Roche, P. & Murphy, G. P. Prostate specific membrane antigen expression in prostatic intraepithelial neoplasia and adenocarcinoma: a study of 184 cases. *Cancer* **82,** 2256–2261 (1998).
229. Vainio, P. *et al.* Arachidonic acid pathway members PLA2G7, HPGD, EPHX2, and CYP4F8 identified as putative novel therapeutic targets in prostate cancer. *Am. J. Pathol.* **178,** 525–536 (2011).
230. Lorenzo, Y. *et al.* Differential genetic and functional markers of second neoplasias in Hodgkin's disease patients. *Clin. Cancer Res.* **15,** 4823–4828 (2009).
231. Figueroa, J. D. *et al.* Bladder cancer risk and genetic variation in AKR1C3 and other metabolizing genes. *Carcinogenesis* **29,** 1955–1962 (2008).

232. Brenner, D. R. *et al.* Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. *Hum. Genet.* **132,** 579–589 (2013).
233. Tomatis, S. *et al.* Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model. *Phys. Med. Biol.* **57,** 1399–1412 (2012).
234. Hendriksen, J. *et al.* RanBP3 enhances nuclear export of active (beta)-catenin independently of CRM1. *J. Cell Biol.* **171,** 785–797 (2005).
235. Polakis, P. Wnt signaling and cancer. *Genes Dev.* **14,** 1837–1851 (2000).
236. Liguori, L. *et al.* The metallophosphodiesterase Mpped2 impairs tumorigenesis in neuroblastoma. *Cell Cycle* **11,** 569–581 (2012).
237. Chen, W.-Z., Pang, B., Yang, B., Zhou, J.-G. & Sun, Y.-H. Differential proteome analysis of conditioned medium of BPH-1 and LNCaP cells. *Chin. Med. J. (Engl).* **124,** 3806–3809 (2011).
238. Loo, J. M. *et al.* Extracellular metabolic energetics can promote cancer progression. *Cell* **160,** 393–406 (2015).
239. Schumacher, F. R. *et al.* Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum. Mol. Genet.* **20,** 3867–3875 (2011).
240. Grabowska, M. M. *et al.* NFI Transcription Factors Interact with FOXA1 to Regulate Prostate-Specific Gene Expression. *Mol. Endocrinol.* **28,** 949–964 (2014).
241. Lee, J. S. *et al.* A novel tumor-promoting role for nuclear factor IA in glioblastomas is mediated through negative regulation of p53, p21, and PAI1. *Neuro. Oncol.* **16,** 191–203 (2014).
242. Matsuo, M. *et al.* Designation of enzyme activity of glycine-N-acyltransferase family genes and depression of glycine-N-acyltransferase in human hepatocellular carcinoma. *Biochem. Biophys. Res. Commun.* **420,** 901–906 (2012).
243. Grabowska, M. M. *et al.* NFI Transcription Factors Interact with FOXA1 to Regulate Prostate-Specific Gene Expression. *Mol. Endocrinol.* **28,** 949–964 (2014).
244. Mirabello, L. *et al.* A Genome-Wide Scan Identifies Variants in NFIB Associated with Metastasis in Patients with Osteosarcoma. *Cancer Discov.* **5,** 920–931 (2015).
245. Gosenca, D. *et al.* Identification and functional characterization of imatinib-sensitive DTD1-PDGFRB and CCDC88C-PDGFRB fusion genes in eosinophilia-associated myeloid/lymphoid neoplasms. *Genes. Chromosomes Cancer* **53,** 411–421 (2014).
246. Bhatlekar, S., Fields, J. Z. & Boman, B. M. HOX genes and their role in the development of human cancers. *J. Mol. Med. (Berl).* **92,** 811–823 (2014).
247. Henderson, D. J. P. *et al.* The cAMP phosphodiesterase-4D7 (PDE4D7) is downregulated in androgen-independent prostate cancer cells and mediates proliferation by compartmentalising cAMP at the plasma membrane of VCaP prostate cancer cells. *Br. J. Cancer* **110,** 1278–1287 (2014).
248. Severi, G. *et al.* A three-protein biomarker panel assessed in

diagnostic tissue predicts death from prostate cancer for men with localized disease. *Cancer Med.* **3,** 1266–1274 (2014).
249. Ji, D. *et al.* Prognostic role of serum AZGP1, PEDF and PRDX2 in colorectal cancer patients. *Carcinogenesis* **34,** 1265–1272 (2013).
250. Huang, C. *et al.* Decreased expression of AZGP1 is associated with poor prognosis in primary gastric cancer. *PLoS One* **8,** e69155 (2013).
251. Kim, Y.-S., Do Hwan, J., Bae, S., Bae, D.-H. & Ahn Shick, W. Identification of differentially expressed genes using an annealing control primer system in stage III serous ovarian carcinoma. *BMC Cancer* **10,** 576 (2010).
252. Protein Atlas. No Title. Available at: https://www.proteinatlas.org/.
253. GeneCards.
254. Amirian, E. S., Ittmann, M. M. & Scheurer, M. E. Associations between arachidonic acid metabolism gene polymorphisms and prostate cancer risk. *Prostate* **71,** 1382–1389 (2011).
255. White, K. L. *et al.* Ovarian cancer risk associated with inherited inflammation-related variants. *Cancer Res.* **72,** 1064–1069 (2012).
256. Sanchez-Espiridion, B. *et al.* Immunohistochemical markers for tumor associated macrophages and survival in advanced classical Hodgkin's lymphoma. *Haematologica* **97,** 1080–1084 (2012).
257. Huang, H., Hara, A., Homma, T., Yonekawa, Y. & Ohgaki, H. Altered expression of immune defense genes in pilocytic astrocytomas. *J. Neuropathol. Exp. Neurol.* **64,** 891–901 (2005).
258. Lu, J. P. *et al.* Androgens induce oxidative stress and radiation resistance in prostate cancer cells though NADPH oxidase. *Prostate Cancer Prostatic Dis.* **13,** 39–46 (2010).
259. Kikuchi, H., Hikage, M., Miyashita, H. & Fukumoto, M. NADPH oxidase subunit, gp91(phox) homologue, preferentially expressed in human colon epithelial cells. *Gene* **254,** 237–243 (2000).
260. Speeckaert, M. M., Speeckaert, R., Laute, M., Vanholder, R. & Delanghe, J. R. Tumor necrosis factor receptors: biology and therapeutic potential in kidney diseases. *Am. J. Nephrol.* **36,** 261–270 (2012).
261. He, L. *et al.* Serglycin (SRGN) overexpression predicts poor prognosis in hepatocellular carcinoma patients. *Med. Oncol.* **30,** 707 (2013).
262. Baron, V. T., Pio, R., Jia, Z. & Mercola, D. Early Growth Response 3 regulates genes of inflammation and directly activates IL6 and IL8 expression in prostate cancer. *Br. J. Cancer* **112,** 755–764 (2015).
263. Huang, M.-Y. *et al.* EVI2B, ATP2A2, S100B, TM4SF3, and OLFM4 as potential prognostic markers for postoperative Taiwanese colorectal cancer patients. *DNA Cell Biol.* **31,** 625–635 (2012).
264. Kim, S.-K. *et al.* A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol. Oncol.* **8,** 1653–1666 (2014).
265. Leite, K. R. M. *et al.* Controlling RECK miR21 Promotes Tumor Cell Invasion and Is Related to Biochemical Recurrence in Prostate Cancer. *J. Cancer* **6,** 292–301 (2015).
266. Mao, B., Xiao, H., Zhang, Z., Wang, D. & Wang, G. MicroRNA21 regulates the expression of BTG2 in HepG2 liver cancer cells. *Mol. Med. Rep.* **12,** 4917–4924 (2015).

267. Maachani, U. B., Tandle, A., Shankavaram, U., Kramp, T. & Camphausen, K. Modulation of miR-21 signaling by MPS1 in human glioblastoma. *Oncotarget* (2015).
268. Ciarlo, M. *et al.* Regulation of neuroendocrine differentiation by AKT/hnRNPK/AR/beta-catenin signaling in prostate cancer cells. *Int. J. Cancer* **131,** 582–590 (2012).
269. Chauhan, S. S. *et al.* Prediction of recurrence-free survival using a protein expression-based risk classifier for head and neck cancer. *Oncogenesis* **4,** e147 (2015).
270. Han, S.-S. *et al.* RNA sequencing identifies novel markers of non-small cell lung cancer. *Lung Cancer* **84,** 229–235 (2014).
271. Saini, V. *et al.* Identification of CBX3 and ABCA5 as putative biomarkers for tumor stem cells in osteosarcoma. *PLoS One* **7,** e41401 (2012).
272. Fernandez, P. L. *et al.* Expression of cathepsins B and S in the progression of prostate carcinoma. *Int. J. Cancer* **95,** 51–55 (2001).
273. Liu, P., Jiang, W., Ren, H., Zhang, H. & Hao, J. Exploring the Molecular Mechanism and Biomakers of Liver Cancer Based on Gene Expression Microarray. *Pathol. Oncol. Res.* **21,** 1077–1083 (2015).
274. Tsai, J.-Y. *et al.* Effects of novel human cathepsin S inhibitors on cell migration in human cancer cells. *J. Enzyme Inhib. Med. Chem.* **29,** 538–546 (2014).
275. Kutomi, G. *et al.* Human endoplasmic reticulum oxidoreductin 1-alpha is a novel predictor for poor prognosis of breast cancer. *Cancer Sci.* **104,** 1091–1096 (2013).
276. Fujita, K. & Nonomura, N. Urinary biomarkers of prostate cancer. *Int. J. Urol.* **25,** 770–779 (2018).
277. NCBI: Gene. (2018). Available at: https://www.ncbi.nlm.nih.gov/gene/5788.
278. Petrie, M. *et al.* The Vesicle Priming Factor CAPS Functions as a Homodimer via C2 Domain Interactions to Promote Regulated Vesicle Exocytosis. *J. Biol. Chem.* **291,** 21257–21270 (2016).
279. Haider, S. *et al.* A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. *Genome Med.* **6,** 105 (2014).
280. Miller, S. *et al.* Genome-wide molecular characterization of central nervous system primitive neuroectodermal tumor and pineoblastoma. *Neuro. Oncol.* **13,** 866–879 (2011).
281. Pal, R. P. *et al.* Immunocytochemical detection of ERG expression in exfoliated urinary cells identifies with high specificity patients with prostate cancer. *BJU Int.* **117,** 686–696 (2016).
282. Whitaker, H. C. *et al.* N-acetyl-L-aspartyl-L-glutamate peptidase-like 2 is overexpressed in cancer and promotes a pro-migratory and pro-metastatic phenotype. *Oncogene* **33,** 5274–5287 (2014).
283. Yoon, H. *et al.* Tudor domain-containing protein 4 as a potential cancer/testis antigen in liver cancer. *Tohoku J. Exp. Med.* **224,** 41–46 (2011).
284. Jiang, Y., Liu, L., Shan, W. & Yang, Z.-Q. An integrated genomic analysis of Tudor domain-containing proteins identifies PHD finger protein 20-like 1 (PHF20L1) as a candidate oncogene in breast cancer. *Mol.*

Oncol. **10,** 292–302 (2016).
285. Halvorsen, O. J. *et al.* Increased expression of SIM2-s protein is a novel marker of aggressive prostate cancer. *Clin. Cancer Res.* **13,** 892–897 (2007).
286. Long, Q. *et al.* Protein-coding and microRNA biomarkers of recurrence of prostate cancer following radical prostatectomy. *Am. J. Pathol.* **179,** 46–54 (2011).
287. Cooperberg, M. R. *et al.* Multiinstitutional validation of the UCSF cancer of the prostate risk assessment for prediction of recurrence after radical prostatectomy. *Cancer* **107,** 2384–2391 (2006).
288. Roobol, M. J. *et al.* Prediction of prostate cancer risk: the role of prostate volume and digital rectal examination in the ERSPC risk calculators. *Eur. Urol.* **61,** 577–583 (2012).
289. Foley, R. W. *et al.* European Randomised Study of Screening for Prostate Cancer (ERSPC) risk calculators significantly outperform the Prostate Cancer Prevention Trial (PCPT) 2.0 in the prediction of prostate cancer: a multi-institutional study. *BJU Int.* **118,** 706–713 (2016).
290. De Nunzio, C. *et al.* External validation of Chun, PCPT, ERSPC, Kawakami, and Karakiewicz nomograms in the prediction of prostate cancer: A single center cohort-study. *Urol. Oncol.* **36,** 364.e1-364.e7 (2018).
291. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics* **8,** 54 (2015).
292. Mantas, D., Kostakis, I. D., Machairas, N. & Markopoulos, C. White blood cell and platelet indices as prognostic markers in patients with invasive ductal breast carcinoma. 1610–1614 (2016). doi:10.3892/ol.2016.4760
293. Schalken, J., Dijkstra, S., Baskin-Bey, E. & Van Oort, I. Potential utility of cancer-specific biomarkers for assessing response to hormonal treatments in metastatic prostate cancer. *Ther. Adv. Urol.* **6,** 245–252 (2014).
294. Ahmed, H. U. *et al.* Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389,** 815–822 (2017).
295. Purysko, A. S., Rosenkrantz, A. B., Barentsz, J. O., Weinreb, J. C. & Macura, K. J. PI-RADS Version 2: A Pictorial Update. *Radiographics* **36,** 1354–1372 (2016).

9

Appendices

Supplementary Table 1 Probe list for NanoString2 (n = 167).

Transcript	*Accession*	*Capture Probe*	*Reporter Probe*	*Biomarker Type*	*Source*
AATF	***NM_012138.3:1175***	***TCATCATCTTCACTAGAAATCTCCTCA CTTCCCGCATTGGGCTTTGTCCC***	***CTCTTTGCAGGGACCCTTCTTCGTTGCT GCTTCTTCTCTTCTACCAGC***	***test***	***Cooper NGS***
ABCB9	***NM_001243013.1:488***	***GGGCCCCAGCGCACTGTTCTTGGCCAC ACCAATGGTGG***	***ACGAAGAGGCACACGAGGGTGATGACC AGCCACGAGGCCCGCAGCCGCCG***	***test***	***Cooper NGS***
ACTR5	***NM_024855.3:1840***	***CAAGGCATGGCGTGCAGGGCAGTCTC TCTGGAGGG***	***GGCAGGTACATCTAGCACAATCACAGT CCTGTCACACTGCCAACGTGGCC***	***test***	***Cooper NGS***
AGR2	***NM_006408.2:1365***	***TGCCTCATCAACACGTCACCACCCTTT GCTCTTCTTCCAATTAGTCACAT***	***TGCCACAGCCTTTCACGTTTCCTAAACC CTAGTAACCTCTGATCTCCATC***	***test***	***Mills***
ALAS1	***NM_000688.4:1615***	***AGTGTTCCAGAAATGATGTCCATTTTT GGCATGACTCCATCCCGATCCCC***	***GAGAACTCGTGCTGGCGATGTACCCTC CAACACAACCAAAGGCTTTGCCA***	***housekeeper***	***Cooper***
AMACR	***NM_014324.4:2145***	***TGGAATCTACCCCTTCCTCACATGCCT TTAGGAAGTTGAGTCCAGGGAAG***	***CAACATCCATTCTCTACTCCCTCTACTC TGATGGCACCCGGATTAGATTG***	***PCa positive control***	***Cooper***
AMH	***NM_000479.3:1626***	***TTGGCCTGGTAGGTCTCGGGGATGAG TACGGAGCG***	***CGGACTGAGGCCAGCCGCACACGCCCT GGCAATTG***	***test***	***Sanda***
ANKRD34B	***NM_001004441.2:1460***	***TTTATAGGATAGTTCTTCCTCTGGTGT AATATCCTGGAGCTCCTCTTGCA***	***ATGCTTTGGTGCCTAGTGATGAACCGC TTGGAAAGTGCCAGCCCATTGGT***	***test***	***Sanda***
ANPEP	***NM_001150.1:2670***	***GTAATGCTGATGATGGTAGAGGTGGC GTCCTGCTTCCGGATTAAGTC***	***AGTTGCTCTGGACAAAGTCCCAGACCA GACCTTGCCCAATGACGTTGTTG***	***test***	***Mills***
APOC1	***NM_001645.3:32***	***CGGAGGGGCACTCTGAATCCTTGCTG GAGGGCTTGGTTGGGAGGTC***	***CAGAACCACCACCAGGACCGGGAGCGA CAGGAAGAGCCTCATGGCGAGGC***	***test***	***Sanda***
ARexon9	***NM_000044.2:3401***	***GACTTGTGCATGCGGTACTCATTGAAA ACCAGATCAGGGGCGAAGTAGAG***	***CAAACTCTTGAGAGAGGTGCCTCATTC GGACACACTGGCTGTACATCCGG***	***test***	***Cooper***
ARexons4-8	***ENST00000514029.1:3171***	***TTTGAAGAGAGGGGTTGGCTGGCTTCT TCTCCTGGAGAAGCAGAAATCTG***	***CAGTAAGGCTAGATGTAAGAGGGAAAG TCGGACTGTAGTCTCTCAGTGTG***	***test***	***Cooper***
ARHGEF25	***NM_001111270.2:1102***	***CAGCGCTTGGGCACAAAGCACATGAC CTCCACAGCTTG***	***CTCAAATCCCCGCAATCTCCCCAGCGT CATCATATCGTTG***	***test***	***Cooper***
AURKA	***NM_003600.2:405***	***AAGGAAATTGCTGAGTCACGAGAACAC GTTTTGGACCTCCAACTGGAGCT***	***ACACAAGACCCGCTGAGCCTGGCCACT ATTTACAGGTAATGGATTCTGAC***	***test***	***Cooper***
B2M	***NM_004048.2:25***	***CACGGAGCGAGACATCTCGGCCCGAA TGCTGTCAGCTT***	***CAGGCCAGAAAGAGAGAGTAGCGCGA GCACAGCTAAGGC***	***housekeeper***	***Cooper***
B4GALNT	***NM_178537.4:492***	***TCCCTCGCCGGGTGGATGAAACCAAAA***	***CAGAACTCCGAGTTGTCGTCTGAGGCC***	***test***	***Sanda***

4		***ATACGGAGTCCATAGTTCTTCCA***	***ACAGAAAACTGGACGTCTCCG***		
BRAF	***NM_004333.3:565***	***AGTGCTTTCTTTAGACTGTCTCGGACT GTAACTCCACACCTTGCAGGTAC***	***CCTGAATTCTGTAAACAGCACAGCACT CTGGGATTAGACCTCTCATCATC***	***test***	***Cooper***
BTG2	***NM_006763.2:1700***	***CAAGGAATACATGCAAGGCTGACTAGC CAGCCATCATCCCAAGGAGAG***	***ACAAGAATACCAAGTAGTCTTGCAGAA CATGGGGCACTCTCCCATTCAGC***	***test***	***Sanda***
CACNA1D	***NM_000720.3:6044***	***GTACTTCTGGGCTTTACTTGAATCTAG GCCGGCAACTGCCATGATCTGTT***	***GTTGCTGGAGGGGTGGCCCACGACCG GGTCGAGTGACTCGGTGA***	***test***	***Sanda***
CADPS	***NM_183394.2:1870***	***TTGAGGCTTATCCATTCGGACAGCAAG TTTGATTTTGAGATCTTGGTCGG***	***TTCCAGACATTCTTACCGATGGCCCATA AATACCCAGAATGCTTCATGTT***	***test***	***Sanda***
CAMK2N2	***NM_033259.2:908***	***AAATACAAATGTGCTGAGGAAGTCCCT TAGAAAGAGGCTGAGGCTGGGGT***	***GGGAGGGCAGGAACCATGAGCAGAGC CAGTAAACAAAGAGTCGGATATAA***	***test***	***Sanda***
CAMKK2	***NM_006549.3:1710***	***GGTGGATGATCTTCTGGTAGTGTAAGT ACTCGATGCCTTTGATCAGATCC***	***CTTGATGTGCCCATCTTCTCCGACCAG GAGGTTGGAAGGTTTGATGTCAC***	***test***	***Cooper***
CASKIN1	***NM_020764.3:1664***	***ACCTTGTAGTACTGGGCCAGGCCGATC ATGGACAG***	***AGGTGATGTCGGTGATGAAATCAATGT TCTCGTAGCCATTGTCCACCAAC***	***test***	***Sanda***
CCDC88B	***NM_032251.5:400***	***TCCACCGCTTCTTCTGAGAGAGGGTCA AATCCCAATGTCTG***	***TGACGCTCCCAACAGTAGCCGAAGAAC GCCTTCCAGCTGC***	***test***	***Cooper - NGS***
CDC10	***NM_000902.2:5059***	***TAGGGCTGGAACAAGGACTCTTTTCTC TGGACAGCTTGCACCTACAATCC***	***CCAAAGGAATATTGCAAATACCCAAGG TCACCCTGTCAGGAGTGGCAGAA***	***test***	***Whitaker***
CDC20	***NM_001255.2:430***	***CCTCTACATCAAAACCGTTCAGGTTCA AAGCCCAGGCTTTCTGATGTTCC***	***ACCCTCTGGCGCATTTTGTGGTTTTCCA CTGAGCCGAAGGATCTTGGCTT***	***test***	***Mills***
CDC37L1	***NM_017913.2:1146***	***TCATCTTCTTTATGTACCACCGAGTTTA AGCTGCAGAGAGCTGTACTGAT***	***GGCCTCAGCAGTCTTAACCAAATTATA CAGTGTCCATCATTTTGGGTTCA***	***test***	***Sanda***
CDKN3	***NM_005192.3:510***	***AGACAAGATCTCCCAAGTCCTCCATAG CAGTGTATTAAGGTTTTTCGGTA***	***CTCTGGTGATATTGTGTCAGACAGGTA TAGTAGGAGACAAGCAGCTACA***	***test***	***Mills***
CKAP2L	***NM_152515.3:1120***	***TGAGGTATACAAACTTGGCTGGACTTC TGATCTTGCTTGATGTTTGGATG***	***AATTAGGCCTCTGGCTTATGGCTTTTGA CTTTTGCAGTACACATGATGTC***	***test***	***Cooper***
CLIC2	***NM_001289.4:50***	***CCAGTCTCTTCTCTCAAGAGGTGTGAC GCAGAAAATTCTAGATGCTTAAG***	***TGCTTTAAGAAGACCGTCTAGCTTGTA GTGGACTGAGTCAGACCTGGAG***	***test***	***Cooper***
CLU	***NM_203339.1:2460***	***GCCTGTGGTCCAGGGAAAGGTATGAA GATCATATAAACCGGCGGTGGACA***	***AGCGTAGGGTACTGCAGCCCAGCTATG GTTCAGACTAAAAGCCGAGAAAC***	***test***	***Cooper***
COL10A1	***NM_000493.3:135***	***CCTGTGGGCATTTGGTATCGTTCAGCG TAAAACACTCCATGAACCAAGTT***	***TGTAGGGAATGAAGAACTGTGTCTTGG TGTTGGGTAGTGGGCCTTTTATG***	***test***	***Sanda***
COL9A2	***NM_001852.3:795***	***CGATAGCGCCCACCATGCCTTTATATC***	***CCTAGGACCTTCCTCACCCGGTGGCCC***	***test***	***Sanda***

		CATGAGGGCCCGTCTCTCCCTTG	AGTGGCAC		
CP	NM_000096.3:1110	CTTGCCCGTGAAAGAAAGCTGCGTGCA CATCAACTTCATTACCCATACCA	AGCAGGAAAGAGGTTGATTGTGTCAAT ACGGTAGTTCTTGTTAGTCAGTG	test	Cooper – NGS
CTA-211A9.5/MIATNB	CTA_211A95.1:407	CTGGAGGTATCCAAGAGTCTGCCGAG GGACTTCAAGTATTCAGGAAGGGG	GAAGAGCCCAAACCTGCCTGGCTTCAA AACAGGTGGTGAGCTCCCCATTG	test	Cooper – NGS
DLX1	NM_001038493.1:1335	CAGCCTCAGGCGAAGTCCATTTCTCAA TAAATAAAACCCCCTCCCTCCAA	CGTTTGAACAGTGCGTTCCTTGCGCCC AGCAGAACCCTGAATTGGCAAA	test	Schalken
DNAH5	NM_001369.2:12374	GGCGGAACGCATCATGTACAAGCTCA GTTTCTATGATTATGTCCATCAGC	CTGAAGGAGTGTAATGGGAAACTGCTT ATGAGCCTCGGTGGTCATCCAGA	test	Sanda
DPP4	NM_001935.3:2700	AAATCCACTCCAACATCGACCAGGGCT TTGGAGATCTGAGCTGACTGCTG	CTGCTAGCTATTCCATGGTCTTCATCAG TATACCACATTGCCTGG	test	
EIF2D	NM_006893.2:1600	GCTCTTGTCCGGGAAGGGTCACTTGAT AGGCAGGCTGTAATTTTTCCAAA	TTGTGCTAGGGTGATGTCAATTGGACA GATTCTCCCTTTCTTCACAATGG	test	Sanda
EN2	NM_001427.3:2576	AAGGTAGCCACATGTTTCAGAACTGTG GACTCAAACACGCCTGGTGTGTG	CTTTCTTCCTTCTTCTAGATCCTGGAGG ATTCTGAGTTCTTTTGAAAGAC	test	Pandha
ERG 3' ex 4-5	NM_001243428.1:177	CCATCTTTTTTCTCTGTGAGTCATTTGT CTTGCTTTTGGTCAACACGGCT	CCATCTACCAGCTGTTCAGAACCTGAC GGCTTTAGTTGCCCTTGGTTCTG	test	Cooper
ERG3' ex 6-7	NM_004449.4:477	TGAGCCATTCACCTGGCTAGGGTTACA TTCCATTTTGATGGTGACCCTGG	CCACCATCTTCCCGCCTTTGGCCACACT GCATTCATCAGGAGAGTTCCT	test	Cooper
ERG5'	NM_182918.3:697	ACATCATCTGAAGTCAAATGTGGAAGA GGAGTCTCTCTGAGGTAGTGGAG	CTGTGTTTCTAGCATGCATTAACCGTG GAGAGTTTTGTAAGGCTTTATCA	PCa positive control	Cooper
FDPS	NM_001135822.1:404	CATCCTGTTTCCTTGGCTCCACCAGCT CCCGGAATGCTACTAC	CCAGCCCACAGTCCAGGCCCGCTGGAG ACTATCAG	test	Sanda
FOLH1	NM_004476.1:695	TGAAAGGTGGTACAATATCCGAAACAT TTTCATATCCTGGAGGAGGTGGT	GTTAACATACACTAGATCGCCCTCTGG CATTCCTTGAGGAGAGAAAGCAC	test	Mills
GABARAPL2	NM_007285.6:340	GGGACTGTCTTATCCACAAACAGGAAG ATCGCCTTTTCAGAAGGAAGCTG	CTTCATCTTTTTCCTTCTCGTAAAGCTG TCCCATAGTTAGGCTGGACTGT	test	Clark
GAPDH	NM_002046.3:972	AAGTGGTCGTTGAGGGCAATGCCAGC CCCAGCGTCAAAG	CCCTGTTGCTGTAGCCAAATTCGTTGTC ATACCAGGAAATGAGCTTGACA	housekeeper	Cooper
GCNT1	NM_001097633.1:394	TTTCAAACAATAATCAGGGATTTCCTT TGTGAAGGGCAGTCTTCTATGCT	GTATTTGGTGGGATAAGAAAAAAGTCT CCTTCGCAGCAACGTCCTCAGCA	Test	Sanda
GJB1	NM_000166.5:190	TGAAGATGAAGATGACCGAGAGCCAT ACTCGGCCAATGGCAGTAGAATGC	TTTCTCATCACCCCACACACTCTCTGCA GCCACCACCAGCACCATGATTC	test	Sanda

GOLM1	***NM_016548.3:508***	***GGATGAGCCTCTCACCTGTGGTGATGTTATTCACCAAAACCGC***	***TAATTCCTCTGCAGGGTCTTTAACTGGTCTTGCAGCACTC***	***test***	***Cooper***
HIST1H1C	***NM_005319.3:401***	***CTTGGCTGCCCCAACTGGCTTCTTAGGTTTGGTTCCGCCCGCCTTTTTAA***	***TTCGGAGTTGCGCCGCCAGCCGCCTTCTTGGGCTT***	***test***	***Sanda***
HIST1H1E	***NM_005321.2:172***	***GCGCTCCTTGGAGGCGGCAACAGCTTTAGTAATGAGCTCGG***	***CTGCCAGCGCTTTCTTGAGAGCGGCCAAAGATACGCCGCT***	***test***	***Sanda***
HIST1H2BF	***NM_003522.3:313***	***CTTGGTGACGGCCTTGGTGCCCTCTGACACGGCGTG***	***AGCCTTTGGGATTGGGTATGAAGACGTTAGAATTACTTAGAGCTGGTGTA***	***test***	***Cooper – NGS***
HIST1H2BG	***NM_003518.3:318***	***TATACTTGGTGACAGCCTTGGTACCTTCGGACACTGCGTGCTTGG***	***AAGAGCCTTTGAGTTTTAAAGCACCTAAGCACACATTTACTTGGAGCTTG***	***test***	***Sanda***
HIST3H2A	***NM_033445.2:114***	***CGGAGCAACCGGTGCACGCGGCCCACGGGGAACTG***	***CGCCGGCGCCCACGCGCTCCGAATAGTTGCCCTTG***	***test***	***Sanda***
HMBS	***NM_000190.3:1020***	***GCTGGGCAGGGACATGGATGGTAGCCTGCATGGTC***	***AGTGATGCCTACCAACTGTGGGTCATCCTCAGGGCCATCTTCAT***	***test***	***Clark***
HOXC4	***NM_014620.4:1058***	***TGAATTTTTTTCATCCATGGGTAGACTATGGGTTGCTTGCTGGCGGCG***	***CGCTTGGGTTCCCCTCCGTTATAATTGGGGTTCACCGTGCTAACG***	***test***	***Schalken***
HOXC6	***NM_153693.3:570***	***GGTCGAGAAATGCCTCACTGGATCATAGGCGGTGGAATTGAGGGCGACGT***	***GAATAAAAGGGAGTCGAGTAGATCCGGTTCTGGGCAACGGCCGCTCCATA***	***test***	***Schalken***
HPN	***NM_182983.1:1870***	***CCGAGAGATGCTGTCCTCACACACAAAGGGACCACCGCTG***	***CCAACTCACAATGCCACACAGCCGCCAACGTGGCGT***	***test***	***Cooper***
HPRT	***NM_000194.1:240***	***TGAGCACACAGAGGGCTACAATGTGATGGCCTCCCATCTCCTTCATCACA***	***CAGTGCTTTGATGTAATCCAGCAGGTCAGCAAAGAATTTATAGCCCCCCT***	***housekeeper***	***Cooper***
IFT57	***NM_018010.2:790***	***AATCGTGACTTTCAGTTGCGGTAGTACACGTTCCACTTCTAGGCTCCATT***	***TGCTGGTGCATTTGGTCAACATGGATTCTCCAATCCTTATTGTCAGTCCT***	***test***	***Sanda***
IGFBP3	***NM_000598.4:1255***	***CGGGCGCATGAAGTCTGGGTGCTGTGCTCGAGTCTCTGAATATTTTGATA***	***TGGTCGGCCGCTTCGACCAACATGTGGTGAGCATTCCA***	***test***	***Sanda***
IMPDH2	***NM_000884.2:545***	***TCTTTGAGAAAATCAATGTCCCTGGAGGAGATGATGCCCACCAAGCGGCT***	***TCCCTCTTTGTCATTATCTCTTCCAAGAAACAGTCATGTTCCTCC***	***test***	***Mills***
ISX	***NM_001008494.1:3140***	***ATCTGGCATTTTTAAGATGGCAAAGCACTTTTGCATCCTGTGGGCTGTTG***	***TGCTAGAGACCTGGTGTTGATATCCACATTCATAGGCTCTGAGTG***	***test***	***Sanda***
ITGBL1	***NM_004791.2:1317***	***AGACCACACCATCGAGGTCTTCACAGCGGCGATCATCACACTCACAAGTC***	***TCCTCTCTCACAAACACAGCGACCACAGGAACATGTGCCGTGGCCTCCAC***	***test***	***Sanda***
ITPR1	***NM_001099952.1:6775***	***GACAATCTCTATCTGCGCCGTGTGCTTGGCATAAAACTCCAGGGC***	***CATATGCTGGGCACGGGAAAGACTATCTGTTCCATTGTTCGGTCTAATCT***	***test***	***Sanda***

KLK2	***NM_005551.3:1820***	***CTTGGACACTAAGGATCAGGTGAGCTT CCTCAGTTGGAATTACTTTGTAC***	***GTCAATTATTCAAGTACTCCATACTCGT CCTACAGACCCCCAGTAAAAAC***	***test***	***Cooper***
KLK3/PSA (exons1-2	***NM_001030048.1:16***	***TGAGGAAGACAACCGGGACCCACATG GTGACACAGCTCTCCGGGTG***	***AATCCGAGACAGGATGAGGGGTGCAGC ACCAATCCACGTCACGGACAGGG***	***Prostate control***	***Doll***
KLK3/PSA (exons2-3	***NM_001648.2:209***	***ATCACGCTTTTGTTCCTGATGCAGTGG GCAGCTGTG***	***CCTGTGTCTTCAGGATGAAACAGGCTG TGCCGACCCAGCAAG***	***Prostate control***	***Cooper***
KLK4	***NM_004917.3:410***	***CCCAGCCAGAAACGAGGCAAGAGTTC CCCGCGGTAG***	***CAGCACGGTAGGCATTCTGCCGTTCGC CAGCAGAC***	***test***	***Cooper***
LASS1	***NM_198207.2:1918***	***GCATCTCGCACCTCCCGTTCCAAAAAA CGTCACGGAGCTCTGAG***	***CTGCCTGGCTACAGCCCCGGATGTGTT AAATGTCT***	***test***	***Sanda***
LBH	***NM_030915.3:2340***	***GAGAGTATGGATGAACCACTCTCTGCA GCCAAAACAGAACGAAGCGGGGA***	***ACAGGAATTGAAAAGGCAAGACCCCCG TCCACAAGGGGAGGCGAGGGAAT***	***test***	***Sanda***
MAK	***NM_005906.3:1395***	***TATCTCCAGACTTGAAGATAGTCTGAC CCCAACGCCTCCTACCACTTTTA***	***CTTCTTGGAATGGGAGGCTCCGAAATC ATAGTCCTCCAACTCTTCCCAGC***	***test***	***Sanda***
MAPK8IP 2	***NM_012324.2:1885***	***CTCTCGCTCCTCGCCGTTGACCAGACA GGAGAAAAGGCCAAAGGACTCG***	***CCGCGGGATGAACCTGAACACAGCCCG GTGAGTCTG***	***test***	***Sanda***
MARCH5	***NM_017824.4:2136***	***TGTGCTGAAACTAGACTGTCAACTCTG TAAGAGCTTGGACCAAGTCTGTC***	***AAACAAAGAGCTCAAGGCCTCACCTTG GTTTATTCACTGCTGGTTTTCTA***	***test***	***Sanda***
MCM7	***NM_182776.1:1325***	***TGTGTTCTCTCCTTCTACCAGCACCGT GATACTACGAGGGATATT***	***CAAGAAAATACCAGTGACGCTGACGTG GTCTCCAGGCTGGGCAATCCT***	***test***	***Perry***
MCTP1	***NM_024717.4:1005***	***AACTCCAATTGTGTCAGATCCAGAAAG GCTGAGCCCATAAAGTCATCCTG***	***GATAATGAGGATCTTTCAGAGTAAGGG TCACATCTGTGGGCCTGTTT***	***test***	***Cooper – NGS***
MDK	***NM_001012334.1:71 1***	***CGAGCAGACAGAAGGCACTGGTGGGT CACATCTCGGGC***	***GGGGCTGGGGAGTGAGAGGGACAAGG CAGGGCATGATTGATTAAAGCTAA***	***test***	***Cooper***
MED4	***NM_001270629.1:32 4***	***TCTTGCTTTTTCTATTGACTTGAGTTTC TCCTTCGCTTGGTAAACAGCTG***	***CTGATCCTATGTGCATACTTAATTATTT CTTCAGAGGAGATAGCACCTTT***	***test***	***Sanda***
MEMO1	***NM_001137602.1:11 92***	***GAATGTGCAGGTGGCATCCCTGAGGA TTCAGAGCT***	***TATCGTGGTAAAGGCTAGGCTGGGACC CCGGACAGAGTATGA***	***test***	***Sanda***
Met	***NM_001127500.1:19 25***	***AAATTTATTATTCCTCCGAAATCCAAA GTCCCAGCCACATATGGTCAGCC***	***GTCAAGGTGCAGCTCTCATTTCCAAGG AGAACTCTAGTTTTCTTTAAATC***	***test***	***Cooper***
MEX3A	***NM_001093725.1:20 90***	***GATCTATGCAACTTCTGATAGGACTCC AACTCCCTTACACTGCTGGAAAC***	***CCTTTCAGCCACAGAAACGATTGACAT GCTTCTCTCCCCAACCCCTAGAA***	***test***	***Sanda***
MFSD2A	***NM_032793.4:592***	***AAGAGGCAATAGAAAAGCAGGTACCA ATAGGTCTGGCCGTGTGGGAAGTC***	***ACATGGTGAGAGCCGAGTAGGGAACAT GGAAACACGTGACCATTGTTTCA***	***test***	***Cooper – NGS***

MGAT5B	***NM_144677.2:3392***	***GGTTGGAACAAGCAGGAGAGAGAAAC AATTCAACCAGGGTCTGGGTGGTC***	***CAGGTCATGCCAGGATGGGTTTTGGGA GAAGCCCAGAGTGAAAAG***	***test***	***Sanda***
MIC1	***NM_004864.2:180***	***CCTGGTTAGCAGGTCCTCGTAGCGTTT CCGCAACTC***	***GTGTTCGAATCTTCCCAGCTCTGGTTG GCCCGCAG***	***test***	***Whitak er***
MIR146A/ DQ658414	***ENST00000517927.1 :1642***	***CGGTTGAGATTTCACCAAGGTTCTGGT TCTGGAATGAGTCACTGGCTAAG***	***TTCTGGATTTTCTCCATCAGTCTAGGAC TGAAGACACCGATCTCTGGTGT***	***test***	***Cooper – NGS***
MIR4435-1HG/lOC541471	***ENST00000409569b.1:45***	***AAAGCAGCGACCATCCAGTCATTTATT TCCCTCCATTCCCAATGATGTAC***	***CAGGCACGGGCTCAGGCACCGCTTGTC TGGAATGTCAATTTGAAACTTAA***	***test***	***Cooper – NGS***
MKi67	***NM_002417.2:4020***	***CTGATGGCATTAGATTCCTGCACGCTA AGAGTTCTCCCTCTACATCTG***	***GTCTTTCTCTTCACCTACTGATGGTTTA GGCGTGTGCATGGCTTTGCCTG***	***test***	***Cooper***
MMP11	***NM_005940.3:702***	***TCAGTGGGTAGCGAAAGGTGTAGAAG GCGGACATCAGGGCCTTGG***	***ATATAGGTGTTGAACGCCCCTGCAGTC ATCTGGGCTGAGAC***	***test***	***Sanda***
MMP25	***NM_022468.4:2955***	***CATTTAGATCCTAAAACTGTGGGGAGT GGGGACAGGGTGAACGAGGTGCC***	***CCCAGTGATTCTGATGTGGGATAGTCT AGAAGAATAGTTCCAGAGGCAAT***	***test***	***Cooper – NGS***
MMP26	***NM_021801.3:515***	***CAGGATTTCCAGAATTTGGTAAAAAGG CATGGCCTAAGATACCACCTGGC***	***TCCAGTGTCTGAAGCTGACCAGTGTTC ATTCTTGTCAAAATGGACAACTC***	***test***	***Cooper***
MNX1	***NM_005515.3:1680***	***TTTCTTGAAGAGCAGGTGAGGCGCCCT TGCTTAAAAGGGAAGCGCCCAGG***	***TTAAAAGAACCAGAGTTCAAGTTTCAG CCCCCTGGGTCTCCCTCTCGCTG***	***test***	***Sanda***
MSMB	***NM_002443.2:295***	***TTTTTGGGTCCTTCTTCTCCACCACGA TATACTTGCAGTCCTCCTTCTTG***	***GTGCCTACTAGAAGCACATTAGATTAT CCATTCACTGACAGAACAGGTCT***	***test***	***Whitak er***
MXI1	***NM_001008541.1:61 5***	***GAAGTGAATGAAAGTTTGACACTGGCA CTGGAGTAACCCTCGTCACTCCC***	***TGGCCCAGTGAATATTTTGCCCTGCAC TGTTATGTCATGCTGGGTTCTAT***	***test***	***Sanda***
MYOF	***NM_013451.3:5805***	***ATGATCGTGTGACGCAAGTCAAGTTCT AGGAAACCCAAGTAGTCATCCAG***	***TGAGGTCCGGAATCATGTCCAATCTGC ATTTCTCTGGTGATTTTGCAGGA***	***test***	***Cooper – NGS***
NAALADL2	***NM_207015.2:250***	***ATTCTCAGCACCGTCTAGCTGGAATTG GTCAAAACCAGACTCCTCTAGTT***	***TGAATGGAATCAAGATTGAGGTCTATA GTCTCTGAATGCCCTAGGTTCTG***	***test***	***Mills***
NEAT1	***NR_028272.1:1850***	***TTTCTCACACACAGATTTAGGAATGAC CAACTTGTACCCTCCCAGCGTTT***	***TTCTCCTAGTAATCTGCAATGCAATCAC AATGCCCAAACTAGACCTGCCA***	***test***	***Sanda***
NKAIN1	***NM_024522.2:1620***	***CACTGTGTTCAAGGCCCACTTCCACCA AAAATCTAGCTGTGTGGCCTCAA***	***GAACTCAGAGAGCAGACACTGGGTTTT ACAGTCAGAAACTGCAGAAAGTA***	***test***	***Sanda***
NLRP3	***NM_001079821.2:41 5***	***CTGGCATATCACAGTGGGATTCGAAAC ACGTGCATTATCTGAACCCCACT***	***CTCGAAAGGTACTCCAGTAAACCCATC CACTCCTCTTCAATGCTGTCTTC***	***test***	***Cooper – NGS***
OGT	***NM_181672.1:1080***	***CTTTGAGAGCATTGGCTAGGTTGCAGT***	***ACGGAGAGCTGTATTATAACAATCTTCT***	***test***	***Cooper***

		AAGCATCAGGGAAATGTGGTTGT	*GCTTCAGCAACACTGCCCTTCT*		
PSGR	*NM_030774.2:360*	*GAGCGTGCAGGCTGCGTTCCGTCCTTA CGATGAAGACCACGATGCAGTTT*	*GGATAAGGCCAGGTCAATGGCTGCAAG CATGCAGAGAAAGAGGTACATCG*	*test*	*Doll*
PALM3	*NM_001145028.1:234*	*AGCTGGGACTGGAGTGTGAACAAACT GTCTTCCAGGTTCCG*	*GCTGGGCACCTGTGGAAGCACTTTGCA ACAGTTGC*	*test*	*Cooper – NGS*
PCA3	*NR_015342.1:362*	*TAAGGAACACATCAATTCATTTTCTAA TGTCCTTCCCTCACAAGCGGGAC*	*TCCCGTTCAAATAAATATCCACAACAG GATCTGTTTTCCTGCCCATCCTT*	*test*	*Cooper*
PCSK6	*NM_138320.1:1112*	*ACATCGCCGTCCAGCATGCGGATGCCT CCTATTTTGGCATTGTACGCTAT*	*CGATGTAGTTGGGTCTGATGCCCAGCG ACTTTGCCTCGACCACATCTGTG*	*test*	*Sanda*
PDLIM5	*NR_046186.1:120*	*CTCAAAGTCCAATGACAGAAAATGAAA TATGCTCGGGTCCGGCGCGGCGC*	*GGCCAACCAGTGACACACTGTAGTTGC TCATGGTTCTAATGG*	*test*	*Sanda*
PECI	*NM_006117.2:940*	*GAAAACTTCAGTAACAAGTCCTTGAGC ACATGCCTCTCCCGCTGTTAACT*	*CAAATGCCTTCAGCCTGGTCCAGACTT CTTTCTGAAAAGTGCTATCAGG*	*test*	*Mills*
PPAP2A	*NM_176895.1:1215*	*GTGATTGCTCGGATAGTGATTCCCAGT TGTTGGTGTTTCATGCAGAGTTG*	*TTAGAAAACAGGCCAGCTTCACCTGGG CACCCTGCTGCCTTTCAAGGCTG*	*test*	*Mills*
PPFIA2	*NM_003625.2:3670*	*CACTTTCATCCAGTCGCCTTTCAGTTC CCAGGGCCAAGAGGTTATTGTAT*	*AGGAGGAAACTGCCTTCTCCAGGTTGA TCCACGTCTGAAGTTCTTGTCAT*	*test*	*Sanda*
PPP1R12B	*NM_001167857.1:1305*	*TGCTCTGTGATACTACTCTTGCTTTCA GAGTTGGAATGATTGACAAAGGC*	*CTAGCAGAAGAGGCAGAGAAGGTATTT TGAGCTGGTGCTGGTATC*	*test*	*Cooper – NGS*
PSTPIP1	*XM_006720737.1:352*	*TCAAAGGAGGCCCTCAGGGAGTTGAT CTCCGTCTG*	*AGCTGCCCACATTCTCCATTTGCTGCTT CAAGGAG*	*test*	*Cooper – NGS*
PTN	*NM_002825.5:418*	*TTTCTTCCCTGCTTCAGCAGTATCCAC AGCTGCCAGTATGAAAATGAATG*	*CCATTCTCCACAGTCAGACTTCTTCACT TTTTTTTCTGGTTTCTC*	*test*	*Sanda*
PTPRC	*NM_080923.2:154*	*CAAGAGTTTAAGCCACAAATACATGGT CATATCTGGAAGTCAGCCGTGTC*	*CTTTGCCCTGTCACAAATACTTCTGTGT CCAGAAAGGCAAAGCCAAATGC*	*Blood control*	*Cooper*
PVT1	*NR_003367.2:0*	*AAAATACTTGAACGAAGCTCCATGCAG CTGACAGGCACAGCCATCTTGAG*	*AGCGTTATTCCCCAGACCACTGAAGAT CACTGTAAATCCATCAGGCTCAG*	*test*	*Sanda*
RAB17	*NR_033308.1:1310*	*ACAGCACTTTCCTGGGAGCCATGTGAC GCCAGATCTTCCTCTGGCAGTTC*	*GGAACAGGCACAGGCATCGGGGAATCA GATGGTATCAGTGGGGATAGGGC*	*test*	*Sanda*
RIOK3	*NM_003831.3:1920*	*CTGGAAAAACTGCGAGACATTCCTGCA GTCCCGGAACAAGAACTCCAGGC*	*ACAGCATTGAAGAGTTCTCGTTCACTA AGGGCTTCCTTGACTCCTCCTTT*	*test*	*Clark*
RNF157	*NM_052916.2:618*	*ACTAGAGGGTAAACTTCTCGGTCTAAA TCAAAGCCAAGCTCCTCTTCGGC*	*CATGGCAATGGCCAAAATACTCGTCTC CTTCATCCACCACGGCATGTACC*	*test*	*Sanda*
RP11-	*ENST00000561140.1*	*TTGCCAGTCGCTGGTTTTCATCCAGAG*	*CAGCAATATATCCTGTTCATCTTCTTCA*	*test*	*Cooper*

97O12.7	***:110***	***CACGAAGCTCGTGGTCTGAATAC***	***TCATGAAGGTCAGCTTTCTTCT***		***– NGS***
RPL18A	***NM_000980.3:177***	***GAGATACAAAGTACCAGAAGCGGGAC TTGGCGACGACATGATTAGGCGCA***	***CTGCCCACAGTAGACAATCTCCCCTGA AGACTTCTTCATCTTCTTTAACT***	***test***	***Cooper – NGS***
RPL23AP53	***NR_003572.2:3226***	***AAATCCGAAAGGATCTCATCCCATTAG GACCCTTGTCTCCTTTTCTGTTG***	***CATTTATGGCTGTCAACCCGCCAGTTCT CAGGAGTTTGTATAAAAGCCT***	***test***	***Sanda***
RPLP2	***NM_001004.3:186***	***CTGATAACCTTGTTGAGCCGGTCGTCG TCCGCCTCGATAC***	***TGCCAATACCCTGGGCAATGACGTCTT CAATGTTTTTTCCATTCAGCTCA***	***housekeeper***	***Whitaker***
RPS10	***NM_001014.3:219***	***GAAATGTCTCCAGGCAAACTGTTCCTT CACGTAGCCTCGGGACTTGAGAG***	***TGAAGGTAATCACGGAGATACTGGATA CCCTCATTGGTAAGGTACCAGTA***	***test***	***Cooper – NGS***
RPS11	***NM_001015.3:105***	***CAGCAGGACCCTCTTCTTGTTTTGAAA GATGGTCGGCTGCTTTTGGTAGG***	***AGACCGATGTTCTTGTAGTACCGCGGG AGCTTCTCCTTGCCAGTTTCTCC***	***test***	***Cooper – NGS***
SACM1L	***NM_014016.3:685***	***AGAAAGTTCTCTTAGAAGATGACCATT CCATACAAACCGCTGATCTGCCC***	***ATAAAGCCATGTAACACTGGAAGGGCA AACCGATGAACCTCTGGCTGTGC***	***test***	***Sanda***
SChLAP1	***NR_104320.1:359***	***CCAGGTACATGGTGAAAGTGCCTTATA CAGGTTGAATAAAAATCACTGCC***	***ACCTTGTGTCCCCAGCATCTAGATTGCT GAAAAAGATGTAGATGTTGCTT***	***test***	***Sanda***
SEC61A1	***NM_013336.3:2245***	***CTCTAAGCCCAACCAGAAGAGTCAGCT AGAAGAGCCAATAGGTGCACAGA***	***GAGCTGATGACCCAAGTGGACTAAACA CGGAGCTAGCAGAAACAGGCAGA***	***test***	***?***
SERPINB5	***NM_002639.4:90***	***CGGGCCTGGAGTCACAGTTATCCTGGA AAATGCGTGGAAAAGGAACAGGC***	***GAACAGATCAACGGCAAAAGCCGAATT TGCTAGTTGCAGGGCATCCATTG***	***test***	***Cooper***
SFRP4	***NM_003014.2:1060***	***CAGCCTCTCTTCCCACTGTATGGATCT TTTACTAAGCTGATCTCTCCATT***	***CCCGGCTGTTTTCTTCTTGTCCTGAACT GTTCTCCGCTGTTCCTG***	***test***	***Sanda***
SIM2.long	***NM_005069.3:2099***	***TTAATGTAGGTCGTGCGCATTTGCCGG GCTCGGTGGCGCCGCAGCC***	***ATCCGCAAGTCGGCGGCGGGGTCCAAT TCAAACAGCTGTCTCTGCATAAA***	***test***	***Sanda***
SIM2.short	***NM_009586.3:2220***	***CTGCCACCCACCGCCATGGCTGCTTCG GCTCCCGG***	***GAAGCAGAAAGAGGGCAAGTTTGCCCA AAGCGTGAGGGTTCTGTCTCCAT***	***test***	***Sanda***
SIRT1	***NM_012238.4:1595***	***GGTGTGGGTGGCAACTCTGACAAATAA GCCAATTCTTTTTGTGTTCGTGG***	***CTGGTGGTGAAGTTCTTTCTGGTGAAC TTGAGTCTTCTGAAACATGAAGA***	***test***	***Sanda***
SLC12A1	***NM_000338.2:3380***	***CCATATACAACAAATCCGATATGGATC CCTTTCTTGCCACGGGAAGGCTC***	***TCTAACTAGTAAGACAGGTGGGAGGTT CTTTGTGAGGATTTCCAACCAAG***	***Kidney control***	***Cooper, Mills***
SLC43A1	***NM_003627.5:925***	***TTGACTTCCTCAGGGGCAGGAAAGGCT TCGATGGGCCAGTTGAGGGTGCA***	***CTTGTGGTCCAGGGCCAGCCCACTCAG CTTGATCTTCTTCGTGTAA***	***test***	***Sanda***
SLC4A1 S	***NM_000342.3:2770***	***CATCATCAGCATCCAGACACTGAAGCT CCACGTTCCTGAAGATGAGCGG***	***CACTTCGTCGTATTCATCCCGACCTTCC TCCTCATCAAAGGTTGCCTTGG***	***test***	***Clark***
SMAP1, ex	***NM_021940.3:1075***	***GAGTACTTTGCTGTTGAATGGTTCCTG***	***TGGTGCTTGTGAGGTAAATGGTATATT***	***test***	***Cooper***

7-8		***TGCCATACAGAGATAAGATGGAG***	***TGTGGGTCCCATAAATACACCAG***		
SMIM1	***ENST00000444870.1:353***	***TTCATGGCGATGCCCAGCTTGCCCGTGCACAGCCTCTGGGAGAT***	***GGTAGCCCAGGATGAAGATGATCCAGAAGAGGGCCACGCCGCCCAGCACC***	***test***	***Cooper – NGS***
SNCA	***NM_007308.2:568***	***ACTGGGAGCAAAGATATTTCTTAGGCTTCAGGTTCGTAGTCTTGATACCC***	***GGAACTGAGCACTTGTACAGGATGGAACATCTGTCAGCAGATCTCAAGAA***	***test***	***Clark***
SNORA20	***NR_002960.1:2***	***CGTATAACTGCTCGTATCACTGTGAGACTACAAGCAGCAAATAAATGGGA***	***ATGGTTACTTCATCTCAATTTACAGTGGCCCAATGTTATTTTATCCCATG***	***test***	***Sanda***
SPINK1	***NM_003122.2:65***	***AAGTTCTGCGTCCAGAGGTCAGTTGAAAACTGCACCGCACTTACCACGTC***	***CAACAGGGCCAAGGCACTGAGAAGAAAGATGCCTGTTACCTTCATGGCTG***	***test***	***Cooper***
SPON2	***NM_012445.1:1680***	***CATTTATTCACTTCTCAAGTGGCCCCCGCTTGGATGCGCCCTCG***	***AACGCAGAGAGATCCATAACATGGAACACTGACGCTTCCGAAACCGCCC***	***test***	***Whitaker***
SRSF3	***NM_003017.4:2640***	***TAAAGTAACTGCCAACTGGGACTGTATGTCACCTAAGTCAGGATAACTCC***	***CCATGTTCTAAAGTTTCTAAGAGTCTTGAGGTTATGCTAGGGCTCCTGGT***	***test***	***Sanda***
SSPO	***NM_198455.2:7270***	***CCACAAGGCAGGGAGAGAAGGGAGCCACATAAGTAGATTCCTGGCG***	***ATGGTAGGCATCATGAAGGGCACAGTGCTCGCTGC***	***test***	***Sanda***
SSTR1	***NM_001049.2:2575***	***TCCGACCCCGCAATCTTATAAAAACTCCTCATTCGGCTTGTTCTCAGCTC***	***GGTCTTTGAAAACGCGCAGTAGGAGGGTGATTCCTATTACGCGCCCACAC***	***test***	***Sanda***
ST6GALNAC1	***ENST00000592042.1:1036***	***TTTTTCCTCAAAATCCCACCGAGGCTCAGATTTGAAGTTGGCGGCCTTCA***	***TTCACAGAGTCAGGGCAAGTCGTCTGAAGGCCTCCTATTTCGAAGCTGTA***	***test***	***Cooper – NGS***
STEAP2	***NM_152999.2:845***	***ATATATAAACCTGCCGGCTGGCATCCTTAGGTCCTAACTGAAGTGCCCAA***	***CTGGCGGGCAAGTTCAATAACCTGTTGTCGCGCTTGAATATTGTTGCTGC***	***test***	***Mills***
STEAP4	***NM_024636.2:3555***	***ATCAAAGATAAGTTGAAGGAGCGTGTGTTCTGTGTACCTTTGCAACCAGT***	***CCATGACTCTACTCAATGTCGTCCAACTTTTGTATCCTTGCTTGGGTTT***	***test***	***Mills***
STOM	***NM_004099.5:120***	***GAGTCGGGGAGCCGCTGGGCTTCGGAGTCCCGTGT***	***CCAAAATCCATCCGCAAGGTCCAAGGCCCTTACTGGGGCTGTCCTTGAAG***	***test***	***Clark***
SULF2	***NM_001161841.1:1206***	***ATGAGGTCTGTGAGGTAATCCTTGGAGTAGTCGGAGC***	***GTACATCTTCTTGGACGTGCGGAAGAAGCTCACGCTGTCATTGGTG***	***test***	***Cooper – NGS***
SULT1A1	***NM_177534.2:1393***	***CCCTCAATTCATATTTTATTCTTGAGCCGCTTGGTCAGGTTTGATTCGCA***	***TCAGCCTCCAAATTGCTGGGATTACAGACATGACCTACCGTCCCGGG***	***test***	***Mills***
SYNM	***NM_015286.4:2460***	***AATGTGACATCGCTTTCTCCATAACCTTCCTCCTCCTTAACCAACCCCCA***	***TCGTGTTCTCCTGAGGCTGCTTGGTCCTTCGATGCTGATTAACTGAG***	***test***	***Sanda***
TBP	***NM_001172085.1:587***	***GCACGAAGTGCAATGGTCTTTAGGTCAAGTTTACAACCAAGATTCACTGT***	***TCCTCATGATTACCGCAGCAAACCGCTTGGGATTATATTCGGCGTTTCGG***	***housekeeper***	***Cooper***
TDRD	***NM_198795.1:2615***	***TGTTTCTAGACTGTATATCTGCTAACT***	***CCCAGCAACACACATCTGGAATCTTGT***	***test***	***Schalke***

		GGCACCGTATTCCCTGAAAGGGA	TATGGCTTCTTCAGACCAATGTT		n
TERF2IP	NM_018975.3:1100	GCCTGTGTAACTGTTGATAGATCCAAG TTAAACTTCTCCATTAACTGCCG	ACGCTAAGAAGGCGGAAGTAGCCTCCA GCTCACCACTATTTTTTAGGAAG	test	Clark
TERT	NM_198253.1:2570	CGCAAGACCCCAAAGAGTTTGCGACG CATGTTCCTCCCAGCCTTGAAGCC	TCTGGAGGCTGTTCACCTGCAAATCCA GAAACAGGCTGTGACACTTCAGC	test	Cooper
TFDP1	NM_007111.4:551	TTCCTCTGCACCTTCTCGCAGACCTTC ATGGAGAAATGCCGTAGGCCCTT	TGAACTCCGCAACCAGCTCGTCTGCCA CTTCGTTGTAGGAAGTGGTCCCT	test	Clark
Timp4	NM_003256.2:1000	TCTGCAGGGAAGGAGAACTGGCTTGA TCTTCAGGACTCTTGAAGGGATGT	GGCACTTCTTATTAGCTGGCAGCAAGA GGTCAGGTGGTAATGGCCAAAGC	test	Cooper
TMCC2	NM_014858.3:1312	ACGTTGCTGCCGTCGGCCAGCAGCAG AGCAGTGTCGGTG	CCCCGATGCCTTCGGCCTCCTCAGCCA GGAGGTAC	test	Clark
TMEM45B	NM_138788.3:469	GCATACAGCAGGAGTGAGTGGATGTG CTGGTCCAGCGGAGGCCGG	GGTCCCGGAAGATCACCTCTAGGGAGA TACTAACACACCCTCCGAACAGA	test	Sanda
TMEM47	NM_031442.3:1215	AGCAAATAACCAACAGCCAATGTAGTC ATTGGGTAGGATAAGCAGGCGGT	CCCATTAGATGCTGAAGGGCAGTTCAT TTTTCAAGGGCTCACTCA	test	Cooper – NGS
TMEM86A	NM_153347.1:2320	AATGAATCAGCCAATCTAATCCCATTG CTCCCAGCTGTTCAACTAAGCCC	GCTCCTGGAGCAGAGTGATGTATTATT CTGCCAGGGCTTTACAACTAATG	test	Cooper – NGS
TMPRSS2: ERG fusion	Fusion_0120.1:0	CTGCCGCGCTCCAGGCGGCGCTCCCC GCCCCTCGC	TAGGCACACTCAAACAACGACTGGTCC TCACTCACAACTGATAAGGCTTC	PCa positive control	Schalken, Cooper
TRPM4	NM_001195227.1:2800	CTTCCAGTAGAGATCGCTGTTGCCCTG TACTTTGCCGAATGTGTAACTGA	GCCAGCGCGGGCCGAGAGTGGAATTCC CGGATGAGGCGGTAACGCTGCGC	test	Sanda
TWIST1	NM_000474.3:393	CTCGGCGGCTGCTGCCGGTCTGGCTCT TCCTCGCTG	TGCTGCTGCGCCGCTTGCGTCCCCCGC GCTTGCCG	test	Sanda
UPK2	NM_006760.3:332	ACGAGGTTTGTCACCTGGTATGCACTG AGCCGAGTGACTG	TCCCCTTCTTCACTAGGTAGGAAATGTA GAATTTGGTTCCTGGC	Bladder control	Cooper
VAX2	NM_012476.2:871	TCACAGGGTGGGAGTCTTAAGTGTTAG CTTTCTTGCAG	ACAGGAGACTGGGAAGGTGCTGTGCTC GGGACTCAGTG	test	Sanda
VPS13A	NM_033305.2:8260	TAAAGGGCTTTGGTGCTGAATCCATGG TGACCGACTTTGGAGGTTTAACA	ACGTGATATCTGGGAATGTCCTGCAGA TCTCATGACAATACTGACATCTG	Test	?
ZNF577	NM_032679.2:268	TCTCTCTTCTGTCTATTCTGGGCCTTCC CAGAAGTGGTGGTCAG	GCCTTGCCCATTTCGTTCAACTCTTAGG GGCTAGCAACTCTAGTATGTTC	Test	Sanda

Supplementary Table 2 Samples flagged by quality checks on Nanostring2 data set

	Samples flagged by Quality Checks		
Samples detected by NanoStringNorm where normalization parameters extended beyond 100% from the mean	***M_83_7***	***M_26_6***	***pc145***
	M_84_2	***M_27_1***	***pc1008_0***
	M_84_5	***M_31_1***	***pc017***
	M_84_6	***M_31_3***	***a293***
	M_85_1	***M_73_1***	***a316***
	M_85_2	***M_76_6***	***a303***
	M_86_1	***M_77_1***	***a1316***
	M_86_2	***M_77_2***	***a1319***
	M_86_3	***M_78_3***	***a1329***
	M_133_7	***M_78_5***	***a138a***
	M_142_7	***M_78_6***	***C113_1***
	M_120_5	***M_78_7***	***C118_4***
	M_122_2	***M_78_9***	***C110_1***
	M_127_6	***M_79_2***	***C112_4***
	M_129_3	***M_79_4***	***C107_2***
	M_131_4	***M_81_1***	***C111_1***
	M_75_3	***M_81_2***	***C109_4***
	M_42_7	***M_81_4***	***C107_1***
	M_80_3	***M_81_5***	***C118_3***
	M_73_7	***M_68_8***	***C106_8***
	M_129_5	***M_92_5***	***C116_5***
	M_131_8	***M_54_7***	***C116_2***
	M_132_2	***M_58_5***	
	M_132_5	***M_67_5***	
Samples detected by NanoStringQCPro which were found to have overlapping barcodes	***M_91-6***	***pc135***	
	M_97-3	***pc137***	
	M_97-4	***pc139***	
	M_142-7	***pc145***	
	M_120-5	***pc146***	
	M_122-2	***pc1008***	
	M_127-6	***pc1013***	
	M_129-3	***pc1029***	
	M_131-4	***pc1043***	
	pc105-5	***pc118***	
	pc130-2	***pc119***	
	pc140-20	***pc121***	
	pc140-5	***a1316***	
	pc140-2	***a1331***	

Supplementary Table 3 Differentially expressed probes for each LPD group determined in the Nanostring2 data set. A) LPD groups 1-3. B) LPD groups 4-5.

A

LPD Group 1			LPD Group 2			LPD Group 3		
	Adjusted *p*-value	Log2(FC)		Adjusted *p*-value	Log2(FC)		Adjusted *p*-value	Log2(FC)
CAMKK2	4.80X10-14	-1.17	***IFT57***	9.50X10-14	0.18	***KLK2***	1.97X10-12	-0.20
CACNA1D	1.10X10-13	-0.49	***OGT***	1.26X10-13	0.27	***DPP4***	2.20X10-12	-0.23
GABARAPL2	1.10X10-13	-0.32	***GABARAPL2***	1.31X10-13	0.17	***CASKIN1***	1.29X10-10	-0.21
RPS11	3.13X10-13	-0.13	***DPP4***	1.56X10-13	0.19	***MSMB***	1.34X10-10	-0.08
RPL23AP53	3.74X10-13	-1.35	***IMPDH2***	1.65X10-13	0.26	***CACNA1D***	1.55X10-10	-0.20
PPAP2A	3.94X10-13	-0.33	***HPRT***	1.68X10-13	0.30	***GABARAPL2***	1.71X10-10	-0.14
CTA.211A9.5.MIATNB	4.44X10-13	-2.43	***EIF2D***	1.69X10-13	0.25	***TERT***	2.02X10-10	-0.24
STEAP2	5.07X10-13	-0.60	***MXI1***	2.05X10-13	0.22	***ZNF577***	2.69X10-10	-0.26
IFT57	8.73X10-13	-0.33	***PECI***	2.09X10-13	0.25	***SSPO***	3.12X10-10	-0.20
MIC1	8.77X10-13	-1.20	***RP11.97O12.7***	2.10X10-13	0.28	***CAMK2N2***	3.32X10-10	-0.52
CASKIN1	1.05X10-12	-0.41	***CACNA1D***	2.14X10-13	0.22	***IFT57***	5.68X10-10	-0.15
HMBS	1.25X10-12	-0.84	***FDPS***	3.25X10-13	0.19	***FOLH1***	5.86X10-10	-0.32
MED4	1.25X10-12	-0.83	***MYOF***	4.00X10-13	0.26	***MNX1***	6.38X10-10	-0.22
RPLP2	1.32X10-12	-0.16	***BRAF***	4.31X10-13	0.30	***MXI1***	1.56X10-09	-0.16
HIST1H1C	1.72X10-12	-0.31	***ALAS1***	4.49X10-13	0.21	***RP11.244H18.1.P712P***	3.02X10-09	-0.28
PCSK6	1.97X10-12	-0.47	***MARCH5***	4.73X10-13	0.24	***STEAP2***	4.33X10-09	-0.24
MMP11	1.98X10-12	-0.66	***KLK2***	4.87X10-13	0.20	***TWIST1***	4.66X10-09	-0.21
Timp4	2.11X10-12	-1.43	***RIOK3***	6.19X10-13	0.35	***RPLP2***	8.02X10-09	-0.08
SIM2.short	2.17X10-12	-0.76	***ZNF577***	6.44X10-13	0.27	***HPRT***	8.93X10-09	-0.19
SLC43A1	2.17X10-12	-2.98	***HIST1H1C***	6.62X10-13	0.17	***MED4***	9.74X10-09	-0.21
SSTR1	2.28X10-12	-0.45	***MED4***	6.69X10-13	0.29	***RPS11***	2.70X10-08	-0.05
SYNM	2.48X10-12	-1.69	***TWIST1***	6.69X10-13	0.32	***PCSK6***	3.33X10-08	-0.27
RPS10	2.52X10-12	-0.18	***TFDP1***	8.58X10-13	0.28	***MMP26***	4.93X10-08	-0.90
MEX3A	2.59X10-12	-1.06	***STEAP2***	9.68X10-13	0.27	***PSGR***	7.51X10-08	-0.33
HIST1H2BG	2.78X10-12	-1.35	***GAPDH***	1.04X10-12	0.18	***NKAIN1***	1.44X10-07	-0.31
VAX2	3.63X10-12	-0.55	***LBH***	1.23X10-12	0.46	***ARexons4.8***	1.46X10-07	-0.19
HOXC4	3.91X10-12	-1.20	***SSPO***	1.33X10-12	0.38	***SSTR1***	1.75X10-07	-0.18

RPL18A	4.06X10-12	-0.33	***TERF2IP***	1.33X10-12	0.18	***EN2***	1.76X10-07	-0.31
PALM3	4.82X10-12	-0.82	***HIST1H2BF***	1.50X10-12	0.22	***PPAP2A***	1.97X10-07	-0.17
FDPS	5.40X10-12	-0.35	***MEMO1***	1.67X10-12	0.28	***CDC20***	2.70X10-07	-0.34
TWIST1	5.54X10-12	-0.51	***NAALADL2***	1.67X10-12	0.30	***MGAT5B***	3.52X10-07	-0.25
EN2	7.01X10-12	-0.56	***RPL18A***	1.67X10-12	0.15	***TFDP1***	3.65X10-07	-0.19
MXI1	1.01X10-11	-0.36	**CASKIN1**	1.86X10-12	0.33	***SLC4A1.S***	7.77X10-07	-0.63
STEAP4	1.11X10-11	-0.64	***ITPR1***	1.86X10-12	0.28	***KLK3 exons 2-3***	7.99X10-07	-0.25
ISX	1.17X10-11	-1.67	***RPLP2***	2.27X10-12	0.10	***EIF2D***	8.19X10-07	-0.19
KLK3 exons 2-3	1.26X10-11	-0.69	***HMBS***	2.34X10-12	0.34	***KLK4***	1.21X10-06	-0.14
TMEM86A	1.36X10-11	-1.51	***PPAP2A***	2.34X10-12	0.18	***ARHGEF25***	1.36X10-06	-0.60
KLK4	1.49X10-11	-0.36	***SLC4A1.S***	2.46X10-12	0.79	***COL9A2***	1.47X10-06	-0.79
MNX1	1.54X10-11	-0.43	***MMP11***	2.71X10-12	0.44	***HIST1H2BF***	2.43X10-06	-0.21
AMH	1.82X10-11	-0.59	***SIM2.short***	2.81X10-12	0.45	***SNCA***	4.94X10-06	-0.28
AR exons 4-8	2.58X10-11	-0.67	***Ar exons 4-8***	2.85X10-12	0.27	***MIR146A.DQ658414***	5.41X10-06	-0.26
SMIM1	2.64X10-11	-1.17	***RP11.244H18.1.P712P***	3.07X10-12	0.25	***COL10A1***	5.98X10-06	-0.30
RIOK3	3.77X10-11	-0.83	***MNX1***	3.52X10-12	0.42	***CADPS***	6.01X10-06	-0.59
BRAF	3.82X10-11	-0.46	***HIST1H2BG***	3.66X10-12	0.25	***VAX2***	6.99X10-06	-0.21
SChLAP1	4.04X10-11	-1.22	***STEAP4***	3.67X10-12	0.33	***HIST1H1C***	7.08X10-06	-0.11
DLX1	4.75X10-11	-1.94	***AMH***	4.12X10-12	0.41	***PTN***	7.92X10-06	-0.39
PVT1	4.81X10-11	-0.68	***SMAP1 exons 7-8***	4.43X10-12	0.36	***PSTPIP1***	8.67X10-06	-0.57
IMPDH2	4.92X10-11	-0.59	***FOLH1***	5.02X10-12	0.27	***MARCH5***	9.65X10-06	-0.13
MGAT5B	5.09X10-11	-0.68	***RPS10***	5.90X10-12	0.13	***SIM2.short***	9.74X10-06	-0.26
CD10	5.79X10-11	-0.54	***GCNT1***	7.08X10-12	0.39	***AMH***	1.07X10-05	-0.21
HPRT	6.13X10-11	-0.66	***BTG2***	7.11X10-12	0.31	***MMP11***	1.57X10-05	-0.20
ITGBL1	6.13X10-11	-0.86	***CCDC88B***	7.11X10-12	0.78	***RPS10***	1.62X10-05	-0.07
EIF2D	7.83X10-11	-0.46	***CDC37L1***	8.55X10-12	0.39	***STEAP4***	1.62X10-05	-0.13
DPP4	8.53X10-11	-0.44	***TMCC2***	8.55X10-12	0.65	***IMPDH2***	1.67X10-05	-0.16
TFDP1	9.69X10-11	-0.88	***DNAH5***	9.37X10-12	0.67	***SPINK1***	2.08X10-05	0.29
TERF2IP	1.01X10-10	-0.31	***NLRP3***	9.81X10-12	0.80	***ISX***	3.99X10-05	-0.55
ZNF577	1.22X10-10	-0.84	***RPS11***	1.05X10-11	0.08	***ACTR5***	4.01X10-05	-0.45
ARHGEF25	1.39X10-10	-1.39	***PDLIM5***	1.13X10-11	0.24	***MMP25***	5.23X10-05	-0.58
SIM2.long	1.42X10-10	-1.29	***SSTR1***	1.72X10-11	0.42	***CAMKK2***	5.77X10-05	-0.39
RP11.97O12.7	1.48X10-10	-0.97	***SRSF3***	2.05X10-11	0.51	***MYOF***	6.98X10-05	-0.15
SRSF3	1.51X10-10	-1.16	***ABCB9***	2.15X10-11	0.77	***NAALADL2***	7.35X10-05	-0.22
SFRP4	1.55X10-10	-0.89	***PPP1R12B***	2.45X10-11	0.33	***TERF2IP***	8.59X10-05	-0.09

PTN	1.58X10-10	-0.71	***B2M***	2.51X10-11	0.27	***CD10***	8.75X10-05	-0.15
COL10A1	1.95X10-10	-0.59	***MDK***	3.09X10-11	0.27	***SChLAP1***	9.41X10-05	-0.49
GOLM1	1.95X10-10	-2.39	***MGAT5B***	3.56X10-11	0.33	***MEMO1***	0.000124295	-0.15
PECI	2.39X10-10	-0.38	***CD10***	4.43X10-11	0.25	***GCNT1***	0.000127837	-0.35
SSPO	2.47X10-10	-0.38	***KLK4***	4.53X10-11	0.16	***KLK3 exons 1-2***	0.000138743	-0.22
MSMB	2.62X10-10	-0.15	***VPS13A***	6.07X10-11	0.37	***MAPK8IP2***	0.00014341	-0.57
SNCA	2.88X10-10	-0.88	***HIST1H1E***	6.14X10-11	0.32	***DNAH5***	0.00014901	-0.39
TBP	2.91X10-10	-2.06	***SEC61A1***	6.14X10-11	0.61	***ST6GALNAC1***	0.000152737	-1.26
PPP1R12B	3.16X10-10	-0.95	***TERT***	6.15X10-11	0.31	***CTA.211A9.5.MIAT NB***	0.000157375	-0.65
ANKRD34B	4.26X10-10	-1.57	***PSGR***	8.66X10-11	0.29	***SLC43A1***	0.00017283	-0.98
TMPRSS2:ERG	4.88X10-10	-1.95	***SMIM1***	9.84X10-11	0.26	***NLRP3***	0.000178965	-0.46
SACM1L	5.03X10-10	-0.62	***COL9A2***	1.20X10-10	0.91	***SRSF3***	0.000264469	-0.36
CADPS	5.12X10-10	-1.63	***AMACR***	1.23X10-10	0.32	***ANKRD34B***	0.000277869	-0.61
KLK2	5.12X10-10	-0.46	***CKAP2L***	1.41X10-10	0.82	***FDPS***	0.00030588	-0.11
COL9A2	5.17X10-10	-1.60	***PSTPIP1***	1.52X10-10	0.78	***SIM2.long***	0.000332184	-0.43
TMEM47	5.25X10-10	-2.48	***PTN***	1.52X10-10	0.36	***UPK2***	0.000344195	0.53
MARCH5	5.48X10-10	-0.58	***VAX2***	1.52X10-10	0.43	***PDLIM5***	0.000389202	-0.12
PPFIA2	5.70X10-10	-1.30	***MMP25***	1.67X10-10	0.78	***ERG5***	0.000408807	-0.47
LASS1	6.33X10-10	-0.69	***SACM1L***	1.91X10-10	0.35	***OGT***	0.000408807	-0.13
PDLIM5	8.66X10-10	-0.44	***CLIC2***	1.93X10-10	0.83	***PVT1***	0.000436352	-0.33
FOLH1	8.98X10-10	-0.62	***GOLM1***	1.99X10-10	0.43	***TMEM47***	0.000458492	-1.08
SNORA20	8.98X10-10	-2.62	***PCSK6***	2.05X10-10	0.38	***RPL18A***	0.000499086	-0.07
GCNT1	9.05X10-10	-1.97	***CDC20***	2.19X10-10	0.73	***PCA3***	0.00052407	-0.19
CLIC2	9.17X10-10	-1.42	***KLK3 exons 2-3***	2.19X10-10	0.19	***SNORA20***	0.000536406	-1.10
GAPDH	9.19X10-10	-0.23	***ACTR5***	3.15X10-10	0.38	***LBH***	0.000583164	-0.34
HOXC6	1.14X10-09	-1.38	***SNCA***	3.25X10-10	0.23	***HIST1H2BG***	0.000703794	-0.15
RP11.244H18.1.P712 P	1.14X10-09	-0.52	***HPN***	3.31X10-10	0.38	***MEX3A***	0.000862201	-0.30
SMAP1 exons 7-8	1.44X10-09	-1.35	***MEX3A***	3.33X10-10	0.68	***PECI***	0.000862201	-0.12
TMEM45B	1.51X10-09	-0.53	***RPL23AP53***	4.01X10-10	0.53	***Timp4***	0.000976934	-0.68
MDK	1.56X10-09	-1.02	***TMEM45B***	4.21X10-10	0.44	***TMEM86A***	0.001450492	-0.56
TERT	2.17X10-09	-0.36	***CAMKK2***	4.60X10-10	0.33	***CKAP2L***	0.001943582	-0.42
CDC20	2.41X10-09	-0.75	***SYNM***	4.93X10-10	0.54	***RNF157***	0.002251894	-0.52
MMP26	2.47X10-09	-1.45	***SIRT1***	5.07X10-10	0.54	***RIOK3***	0.002581105	-0.15
AGR2	2.87X10-09	-1.52	***MFSD2A***	5.26X10-10	0.78	***AGR2***	0.002794445	-0.32
OR52A2.PSGR	2.97X10-09	-1.37	***COL10A1***	6.30X10-10	0.44	***CCDC88B***	0.003237426	-0.54

MYOF	3.29X10-09	-0.36	***MCM7***	6.67X10-10	0.47	***GAPDH***	0.003237426	-0.07
SLC4A1.S	3.42X10-09	-2.05	***SIM2.long***	6.92X10-10	0.50	***SMAP1 exons 7-8***	0.003237426	-0.26
KLK3 exons 1-2	3.66X10-09	-0.93	***NKAIN1***	7.55X10-10	0.25	***SACM1L***	0.003611413	-0.15
AATF	3.68X10-09	-1.94	***MIR146A.DQ658414***	8.31X10-10	0.44	***LASS1***	0.003758809	-0.32
NLRP3	3.68X10-09	-1.21	***KLK3 exons 1-2***	9.53X10-10	0.21	***MCM7***	0.003892978	-0.36
SPON2	3.93X10-09	-0.79	***EN2***	1.00X10-09	0.34	***GOLM1***	0.004104272	-0.29
ACTR5	4.55X10-09	-0.87	***GJB1***	1.00X10-09	0.26	***MIC1***	0.004950076	-0.36
HPN	5.48X10-09	-0.99	***TMEM86A***	1.11X10-09	1.06	***SFRP4***	0.005187855	-0.43
MAK	5.71X10-09	-2.15	***ANKRD34B***	1.13X10-09	0.91	***SMIM1***	0.005244137	-0.16
B4GALNT4	5.82X10-09	-1.13	***TBP***	1.34X10-09	0.42	***MAK***	0.005275769	-0.75
CDKN3	5.82X10-09	-0.94	***AGR2***	1.39X10-09	0.45	***IGFBP3***	0.005353681	0.36
GJB1	5.82X10-09	-1.53	***MMP26***	1.39X10-09	1.02	***SYNM***	0.005479266	-0.45
TRPM4	5.82X10-09	-1.59	***PVT1***	1.83X10-09	0.45	***TMCC2***	0.006177628	-0.31
ITPR1	5.88X10-09	-0.64	***ANPEP***	1.91X10-09	0.38	***DLX1***	0.006513656	-0.51
CKAP2L	6.54X10-09	-0.97	***Timp4***	2.60X10-09	0.40	***SERPINB5***	0.007855423	-0.32
NAALADL2	9.10X10-09	-0.81	***RAB17***	2.66X10-09	0.45	***TRPM4***	0.009529009	-0.41
AMACR	9.36X10-09	-1.31	***AATF***	2.69X10-09	0.46	***RP11.97O12.7***	0.010060703	-0.14
HIST3H2A	1.00X10-08	-1.05	***CTA.211A9.5.MIATNB***	2.91X10-09	0.36	***CLU***	0.010087038	-0.45
MIR146A.DQ658414	1.04X10-08	-0.74	***SERPINB5***	3.68X10-09	0.47	***SLC12A1***	0.010326711	0.41
SEC61A1	1.15X10-08	-1.99	***NEAT1***	3.68X10-09	0.41	***MDK***	0.010838865	-0.16
ERG3 exons 4-5	1.50X10-08	-1.34	***B4GALNT4***	4.93X10-09	0.87	***CDKN3***	0.011131753	-0.39
MCM7	1.50X10-08	-1.48	***CADPS***	7.75X10-09	1.22	***BRAF***	0.011920795	-0.11
HIST1H2BF	1.69X10-08	-0.53	***SChLAP1***	8.52X10-09	0.63	***HMBS***	0.011920795	-0.22
SIRT1	1.85X10-08	-1.41	***MIC1***	8.65X10-09	0.58	***MIR4435.1HG.lOC541471***	0.013229338	0.15
ERG3 exons 6-7	1.93X10-08	-1.96	***PALM3***	8.65X10-09	0.46	***RPL23AP53***	0.013779284	-0.23
CDC37L1	2.09X10-08	-0.71	***HIST3H2A***	9.40X10-09	0.28	***AATF***	0.023593676	-0.23
ARexon9	2.38X10-08	-1.50	***ARHGEF25***	1.09X10-08	0.73	***B4GALNT4***	0.023593676	-0.28
HIST1H1E	2.50X10-08	-0.58	***MSMB***	1.31X10-08	0.06	***HIST3H2A***	0.023593676	-0.22
STOM	2.86X10-08	-2.23	***ISX***	2.15X10-08	0.95	***MKi67***	0.023593676	-0.61
SERPINB5	3.66X10-08	-1.39	***TRPM4***	2.75X10-08	0.54	***B2M***	0.024443101	-0.10
LBH	3.98X10-08	-1.22	***CLU***	2.93X10-08	0.84	***AR exon 9***	0.027370444	-0.57
TMCC2	4.17X10-08	-1.84	***HOXC6***	2.93X10-08	0.32	***CLIC2***	0.029401344	-0.28
MMP25	4.69X10-08	-1.22	***RNF157***	2.93X10-08	0.51	***SPON2***	0.029401344	-0.10
ERG5	5.45X10-08	-1.67	***ST6GALNAC1***	2.93X10-08	0.65	***ABCB9***	0.047180253	-0.18

CAMK2N2	7.63X10-08	-0.79	***AR exon 9***	3.18X10-08	1.02
VPS13A	8.47X10-08	-0.88	***HOXC4***	4.42X10-08	0.63
MFSD2A	8.52X10-08	-1.44	***ERG5***	5.96X10-08	0.79
ST6GALNAC1	9.99X10-08	-2.46	***SLC43A1***	7.19X10-08	0.70
AURKA	1.35X10-07	-2.28	***MAPK8IP2***	9.88X10-08	0.93
CLU	2.46X10-07	-0.87	***PCA3***	9.92X10-08	0.25
MEMO1	3.02X10-07	-0.58	***CAMK2N2***	1.00X10-07	0.43
PCA3	3.02X10-07	-0.31	***MKi67***	1.33X10-07	1.01
ALAS1	3.42X10-07	-0.26	***CP***	2.10X10-07	0.75
CCDC88B	3.75X10-07	-1.09	***ITGBL1***	2.10X10-07	0.45
OGT	3.75X10-07	-0.38	***SNORA20***	2.84X10-07	0.56
SULF2	3.75X10-07	-1.39	***PPFIA2***	2.90X10-07	0.60
Met	3.77X10-07	NA	***TDRD***	3.25X10-07	1.05
TDRD	4.18X10-07	-1.95	***STOM***	3.73X10-07	0.94
B2M	9.21X10-07	-0.38	***CDKN3***	4.26X10-07	0.68
RNF157	9.31X10-07	-1.40	***SULF2***	6.30X10-07	0.66
RAB17	9.77X10-07	-0.67	***AURKA***	6.74X10-07	0.49
SULT1A1	2.23X10-06	-1.48	***SFRP4***	6.74X10-07	0.73
MAPK8IP2	3.59X10-06	-0.96	***MIR4435.1HG.lOC 541471***	2.25X10-06	0.26
DNAH5	3.84X10-06	-1.08	***LASS1***	2.34X10-06	0.64
MCTP1	4.54X10-06	-1.97	***SLC12A1***	1.18X10-05	0.56
MKi67	5.94X10-06	-3.13	***SPON2***	1.18X10-05	0.18
BTG2	6.71X10-06	-1.07	***TMEM47***	1.26X10-05	0.75
NKAIN1	6.71X10-06	-1.27	***ERG3 exons 4-5***	2.02X10-05	0.73
CP	6.85X10-06	-1.89	***ERG3 exons 6-7***	2.02X10-05	0.92
ANPEP	7.42X10-06	-1.24	***MAK***	2.57X10-05	1.04
PSTPIP1	7.42X10-06	-0.98	***DLX1***	2.70X10-05	1.20
NEAT1	2.28X10-05	-1.31	***SULT1A1***	6.39X10-05	0.23
ABCB9	0.000117259	-0.68	***MCTP1***	6.98X10-05	0.48
APOC1	0.000171548	-1.63	***Met***	9.24X10-05	0.91
SLC12A1	0.000171548	-1.10	***PTPRC***	9.24X10-05	0.90
MIR4435.1HG.lOC54 1471	0.000399737	-1.01	***TMPRSS2:ERG***	9.24X10-05	1.19
UPK2	0.001159053	-1.38	***APOC1***	0.001191952	0.41
PTPRC	0.001376488	-1.44	***SPINK1***	0.004238366	0.17
IGFBP3	0.009734383	-1.56	***IGFBP3***	0.005829551	0.41

SPINK1	0.039101303	-0.22

B

LPD Group 4			LPD Group 5		
	Adjusted *p*-value	Log2(FC)		Adjusted *p*-value	Log2(FC)
VPS13A	3.38X10-06	-0.11	***GABARAPL2***	2.26X10-22	0.07
TERF2IP	3.79X10-06	-0.05	***CACNA1D***	2.71X10-21	0.09
ABCB9	1.47X10-05	-0.21	***STEAP2***	3.26X10-17	0.09
X05.Mar	1.64X10-05	-0.08	***KLK2***	4.09X10-17	0.07
MMP25	1.89X10-05	-0.25	***MED4***	2.31X10-16	0.09
TMEM45B	1.92X10-05	-0.14	***CASKIN1***	1.66X10-15	0.13
RPLP2	1.93X10-05	-0.03	***DPP4***	7.40X10-15	0.07
PECI	2.41X10-05	-0.06	***IFT57***	8.66X10-15	0.07
CASKIN1	2.64X10-05	-0.10	***RPS11***	8.75X10-14	0.03
MEMO1	3.23X10-05	-0.08	***MARCH5***	9.67X10-14	0.09
AMACR	4.96X10-05	-0.10	***MMP25***	1.56X10-13	0.33
SLC4A1.S	5.73X10-05	-0.24	***STEAP4***	2.42X10-13	0.09
GABARAPL2	6.12X10-05	-0.04	***TWIST1***	2.71X10-13	0.12
TWIST1	6.76X10-05	-0.08	***MMP26***	4.57X10-13	0.47
FDPS	8.19X10-05	-0.04	***SYNM***	5.46X10-13	0.31
CACNA1D	0.00010192	-0.06	***TERF2IP***	6.23X10-13	0.05
CP	0.000320555	-0.30	***FDPS***	6.52X10-13	0.05
RPS11	0.000453771	-0.02	***PCSK6***	8.32X10-13	0.13
TMEM86A	0.000537454	-0.27	***SSTR1***	1.15X10-12	0.12
PPP1R12B	0.00056124	-0.08	***MNX1***	1.79X10-12	0.15
TDRD	0.000623636	-0.71	***HPRT***	2.47X10-12	0.10
TMCC2	0.001057497	-0.18	***FOLH1***	2.60X10-12	0.11
BTG2	0.001079391	-0.07	***CDC20***	2.84X10-12	0.21
BRAF	0.001466681	-0.07	***SLC4A1.S***	3.09X10-12	0.32
ITPR1	0.00149178	-0.07	***COL10A1***	9.32X10-12	0.20
MFSD2A	0.001912749	-0.24	***TDRD***	1.16X10-11	0.89
EN2	0.001944902	-0.10	***TERT***	1.77X10-11	0.14
SLC12A1	0.001954045	-0.20	***EN2***	3.39X10-11	0.14
CDC37L1	0.002355197	-0.09	***ZNF577***	4.49X10-11	0.08
PSTPIP1	0.002736615	-0.18	***SSPO***	4.81X10-11	0.12
COL10A1	0.002804912	-0.14	***VAX2***	5.56X10-11	0.16
ITGBL1	0.00290803	-0.16	***MGAT5B***	8.97X10-11	0.13
ALAS1	0.003470784	-0.04	***RPL23AP53***	1.42X10-10	0.29
DPP4	0.003974	-0.04	***CAMK2N2***	1.56X10-10	0.26
CCDC88B	0.004126726	-0.22	***ERG5***	2.20X10-10	0.36
SPINK1	0.004287134	-0.15	***MXI1***	2.44X10-10	0.06
HOXC4	0.004817751	-0.20	***HIST1H2BG***	3.78X10-10	0.11
IGFBP3	0.006132992	-0.25	***PPAP2A***	4.44X10-10	0.06
UPK2	0.006309542	-0.30	***TMEM86A***	6.42X10-10	0.36
RP11.97O12.7	0.00684069	-0.05	***MEMO1***	9.10X10-10	0.09
GJB1	0.007619494	-0.09	***RP11.244H18.1.P712P***	9.28X10-10	0.09
SSTR1	0.007727274	-0.08	***ARHGEF25***	1.20X10-09	0.39
EIF2D	0.008645702	-0.06	***RPLP2***	1.25X10-09	0.03
MED4	0.009155071	-0.05	***SFRP4***	1.34X10-09	0.34
OGT	0.010622928	-0.06	***HIST1H1C***	1.50X10-09	0.05
MIR4435.1HG.lOC541471	0.011952383	-0.13	***COL9A2***	1.82X10-09	0.38
AMH	0.012837851	-0.08	***PECI***	1.82X10-09	0.06
MGAT5B	0.012837851	-0.09	***BRAF***	2.21X10-09	0.09
RIOK3	0.013178355	-0.07	***CAMKK2***	3.30X10-09	0.16
MXI1	0.01326158	-0.05	***SIM2.short***	3.59X10-09	0.15
PPAP2A	0.01326158	-0.04	***SChLAP1***	3.59X10-09	0.36
STEAP4	0.01326158	-0.06	***RIOK3***	3.88X10-09	0.11
PTPRC	0.017081934	-0.45	***AMH***	4.07X10-09	0.13
VAX2	0.017081934	-0.09	***LBH***	4.60X10-09	0.20
SSPO	0.021717641	-0.06	***SACM1L***	4.74X10-09	0.11
SACM1L	0.022028919	-0.08	***PDLIM5***	8.54X10-09	0.08

SULF2	0.02375717	-0.22	***ERG3 exons 4-5***	9.30X10-09	0.51
TMPRSS2:ERG	0.023967171	-0.54	***LASS1***	9.35X10-09	0.30
CKAP2L	0.029163389	-0.14	***GCNT1***	1.20X10-08	0.15
KLK2	0.035832883	-0.04	***MIR146A.DQ658414***	1.48X10-08	0.16
HIST1H2BG	0.041078353	-0.07	***MMP11***	1.90X10-08	0.12
			MEX3A	1.90X10-08	0.19
			ANKRD34B	2.35X10-08	0.27
			EIF2D	2.38X10-08	0.07
			OGT	2.54X10-08	0.07
			PSTPIP1	2.79X10-08	0.27
			TFDP1	3.91X10-08	0.07
			DLX1	4.64X10-08	0.65
			B4GALNT4	5.18X10-08	0.31
			TMPRSS2:ERG	5.95X10-08	0.93
			MSMB	7.42X10-08	0.03
			CADPS	1.01X10-07	0.37
			CKAP2L	1.23X10-07	0.24
			SNORA20	1.42X10-07	0.35
			MAK	1.67X10-07	0.72
			ERG3 exons 6-7	1.95X10-07	0.66
			ITPR1	3.32X10-07	0.09
			RP11.97O12.7	4.08X10-07	0.06
			AMACR	5.11X10-07	0.12
			ISX	5.11X10-07	0.36
			PVT1	5.11X10-07	0.16
			HOXC4	6.68X10-07	0.23
			AR exons 4-8	6.96X10-07	0.07
			MKi67	7.96X10-07	0.44
			TMEM47	7.96X10-07	0.42
			HMBS	8.12X10-07	0.14
			IMPDH2	8.35X10-07	0.05
			NAALADL2	8.35X10-07	0.07
			AR exon 9	9.27X10-07	0.58
			TMCC2	9.91X10-07	0.22
			HIST1H2BF	1.56X10-06	0.07
			SMAP1 exons 7-8	1.62X10-06	0.12
			CLIC2	1.66X10-06	0.34
			TMEM45B	2.10X10-06	0.12
			KLK4	2.14X10-06	0.06
			ABCB9	2.50X10-06	0.25
			GJB1	2.50X10-06	0.10
			MYOF	2.52X10-06	0.06
			DNAH5	2.53X10-06	0.28
			CDC37L1	3.63X10-06	0.09
			PTN	6.39X10-06	0.11
			PPP1R12B	6.95X10-06	0.08
			NKAIN1	8.02X10-06	0.10
			PPFIA2	1.11X10-05	0.28
			SRSF3	1.31X10-05	0.14
			CP	1.41X10-05	0.29
			TBP	1.64X10-05	0.10
			RNF157	1.79X10-05	0.22
			CCDC88B	1.99X10-05	0.29
			NLRP3	2.01X10-05	0.21
			ACTR5	2.28X10-05	0.13
			GOLM1	2.46X10-05	0.14
			VPS13A	2.47X10-05	0.08
			CD10	2.99X10-05	0.06
			MAPK8IP2	2.99X10-05	0.29
			PCA3	2.99X10-05	0.11
			CTA.211A9.5.MI	4.66X10-05	0.17

	ATNB		
	SLC43A1	4.66X10-05	0.27
	RPL18A	5.11X10-05	0.04
	SNCA	5.11X10-05	0.08
	MIC1	8.46X10-05	0.19
	Timp4	9.62X10-05	0.15
	MFSD2A	0.000117774	0.28
	RPS10	0.000141176	0.04
	CLU	0.000315682	0.32
	MCM7	0.000317762	0.16
	SMIM1	0.000387115	0.09
	SEC61A1	0.000429125	0.20
	CDKN3	0.000442848	0.19
	PALM3	0.000442848	0.13
	SIRT1	0.000556554	0.13
	TRPM4	0.000941039	0.18
	AATF	0.00110258	0.09
	ALAS1	0.001662622	0.03
	ST6GALNAC1	0.002493816	0.35
	ITGBL1	0.003929639	0.12
	KLK3 exons 2-3	0.004011089	0.04
	AGR2	0.00404698	0.11
	SERPINB5	0.00406767	0.13
	MDK	0.004201206	0.05
	SULF2	0.005040354	0.13
	GAPDH	0.010197511	0.02
	KLK3 exons 1-2	0.010197511	0.05
	MCTP1	0.010197511	0.21
	PSGR	0.010197511	0.07
	HOXC6	0.015699677	0.06
	STOM	0.015699677	0.24
	AURKA	0.016070673	0.14
	PTPRC	0.019655708	0.35
	SIM2.long	0.021194067	0.17
	NEAT1	0.030247935	0.09
	HIST3H2A	0.030856327	0.07
	SPON2	0.043657569	0.04
	B2M	0.050443121	0.04
	RAB17	0.050443121	0.10
	BTG2	0.05175315	0.03
	SPINK1	0.08344823	-0.09
	ANPEP	0.131696047	0.07
	HIST1H1E	0.221627267	0.03
	HPN	0.261795514	0.05
	SULT1A1	0.298725188	0.03

6.12 Binomial Testing between CB and Ca

Supplementary Table 4 Glm binomial tests – significant probes between CB and Ca (L I H)

KLK2 Ratio data				*KLK2 adjusted data*			
Transcript	***p-value***	***Log$_2$(FC)***	***Adjusted p-value***	***Transcript***	***p-value***	***Log$_2$(FC)***	***Adjusted p-value***
ERG3' exons 4-5	***1.54x10-09***	***1.582***	***2.55x10-07***	***PCA3***	***4.49 x10-07***	***0.192***	***7.46 x10-05***
TMPRSS2:ERG	***8.73x10-09***	***NA***	***1.44x10-06***	***HPN***	***4.82 x10-06***	***0.180***	***0.001***
PCA3	***1.10x10-08***	***0.321***	***1.81x10-06***	***SIM2.short***	***6.21 x10-05***	***0.124***	***0.010***
ERG3' exons 6-7	***2.44x10-08***	***2.808***	***3.97x10-06***	***AMACR***	***6.40 x10-05***	***0.124***	***0.010***
HOXC6	***8.04x10-07***	***0.295***	***0.0001***	***ERG3' exons 4-5***	***0.0001***	***0.103***	***0.018***
TDRD	***2.14x10-06***	***3.683***	***0.0003***	***SMIM1***	***0.0003***	***0.142***	***0.048***
DLX1	***2.76x10-05***	***4.219***	***0.004***	***ERG3' exons 6-7***	***0.0003***	***0.101***	***0.056***
ERG5	***0.0002***	***NA***	***0.025***	***HOXC6***	***0.0004***	***0.130***	***0.058***
ISX	***0.0002***	***2.227***	***0.028***	***GJB1***	***0.0004***	***0.129***	***0.061***
HOXC4	***0.0002***	***0.900***	***0.031***	***TMPRSS2:ERG***	***0.0004***	***0.098***	***0.061***
TRPM4	***0.0002***	***0.652***	***0.032***	***CAMKK2***	***0.001***	***0.098***	***0.079***
PPFIA2	***0.0002***	***0.613***	***0.032***	***GAPDH***	***0.001***	***0.119***	***0.116***
HPN	***0.0003***	***0.270***	***0.046***	***MMP11***	***0.001***	***0.083***	***0.132***
GJB1	***0.0003***	***0.234***	***0.050***	***TRPM4***	***0.001***	***0.103***	***0.143***
APOC1	***0.001***	***1.001***	***0.093***	***AMH***	***0.001***	***0.112***	***0.164***
AMACR	***0.001***	***0.261***	***0.099***	***SIM2.long***	***0.001***	***0.121***	***0.216***
DNAH5	***0.001***	***0.485***	***0.105***	***RAB17***	***0.002***	***0.164***	***0.286***
MCTP1	***0.001***	***1.025***	***0.116***	***IMPDH2***	***0.002***	***0.101***	***0.291***
SIM2.long	***0.001***	***0.132***	***0.121***	***DNAH5***	***0.002***	***0.086***	***0.330***
ANKRD34B	***0.001***	***5.060***	***0.219***	***TDRD***	***0.003***	***0.061***	***0.390***

SLC12A1	***0.002***	***0.759***	***0.280***	***RIOK3***	***0.003***	***0.065***	***0.445***
MEX3A	***0.002***	***0.782***	***0.286***	***RP11.97O12.7***	***0.004***	***0.104***	***0.578***
PVT1	***0.002***	***0.361***	***0.328***	***ISX***	***0.004***	***0.075***	***0.596***
CDKN3	***0.002***	***0.275***	***0.331***	***TWIST1***	***0.005***	***0.055***	***0.656***
RP11.97O12.7	***0.002***	***0.045***	***0.345***	***CLU***	***0.005***	***0.042***	***0.690***
SSTR1	***0.003***	***0.249***	***0.411***	***DLX1***	***0.007***	***0.067***	***0.974***
NAALADL2	***0.003***	***0.077***	***0.417***	***ANKRD34B***	***0.007***	***0.082***	***0.994***
CAMKK2	***0.003***	***0.141***	***0.429***	***RNF157***	***0.007***	***0.067***	***0.994***
SMIM1	***0.003***	***0.143***	***0.434***	***KLK4***	***0.008***	***-0.058***	***0.994***
RAB17	***0.004***	***0.207***	***0.542***	***ERG5***	***0.009***	***0.089***	***0.994***
NEAT1	***0.004***	***0.116***	***0.554***	***MYOF***	***0.009***	***-0.091***	***0.994***
RIOK3	***0.004***	***0.081***	***0.576***	***EN2***	***0.010***	***0.069***	***0.994***
SIM2.short	***0.005***	***0.375***	***0.625***	***SULT1A1***	***0.012***	***0.090***	***0.994***
ST6GALNAC1	***0.005***	***0.389***	***0.648***	***CASKIN1***	***0.013***	***0.056***	***0.994***
GOLM1	***0.005***	***0.193***	***0.662***	***PVT1***	***0.013***	***0.121***	***0.994***
RPL23AP53	***0.005***	***0.348***	***0.696***	***APOC1***	***0.016***	***0.095***	***0.994***
SULT1A1	***0.005***	***0.123***	***0.700***	***RPS11***	***0.016***	***-0.026***	***0.994***
MIC1	***0.006***	***0.395***	***0.781***	***MNX1***	***0.016***	***0.051***	***0.994***
IMPDH2	***0.006***	***0.074***	***0.831***	***GABARAPL2***	***0.016***	***-0.078***	***0.994***
RNF157	***0.007***	***0.439***	***0.885***	***SLC12A1***	***0.018***	***0.080***	***0.994***
SYNM	***0.008***	***0.250***	***0.946***	***PSGR***	***0.018***	***0.067***	***0.994***
COL9A2	***0.008***	***-2.292***	***0.967***	***ITGBL1***	***0.022***	***0.068***	***0.994***
AMH	***0.008***	***0.246***	***0.981***	***SSPO***	***0.022***	***0.077***	***0.994***
CLU	***0.008***	***0.406***	***0.981***	***MIC1***	***0.024***	***0.094***	***0.994***
MMP11	***0.009***	***0.295***	***0.992***	***HMBS***	***0.024***	***0.067***	***0.994***
MKi67	***0.010***	***-1.896***	***0.992***	***IGFBP3***	***0.024***	***-0.016***	***0.994***
MMP26	***0.010***	***0.438***	***0.992***	***RPLP2***	***0.025***	***-0.069***	***0.994***
SULF2	***0.010***	***1.538***	***0.992***	***MFSD2A***	***0.026***	***0.078***	***0.994***
MCM7	***0.010***	***0.290***	***0.992***	***SYNM***	***0.026***	***0.066***	***0.994***
MIR146A.DQ65	***0.010***	***0.324***	***0.992***	***NEAT1***	***0.027***	***0.061***	***0.994***

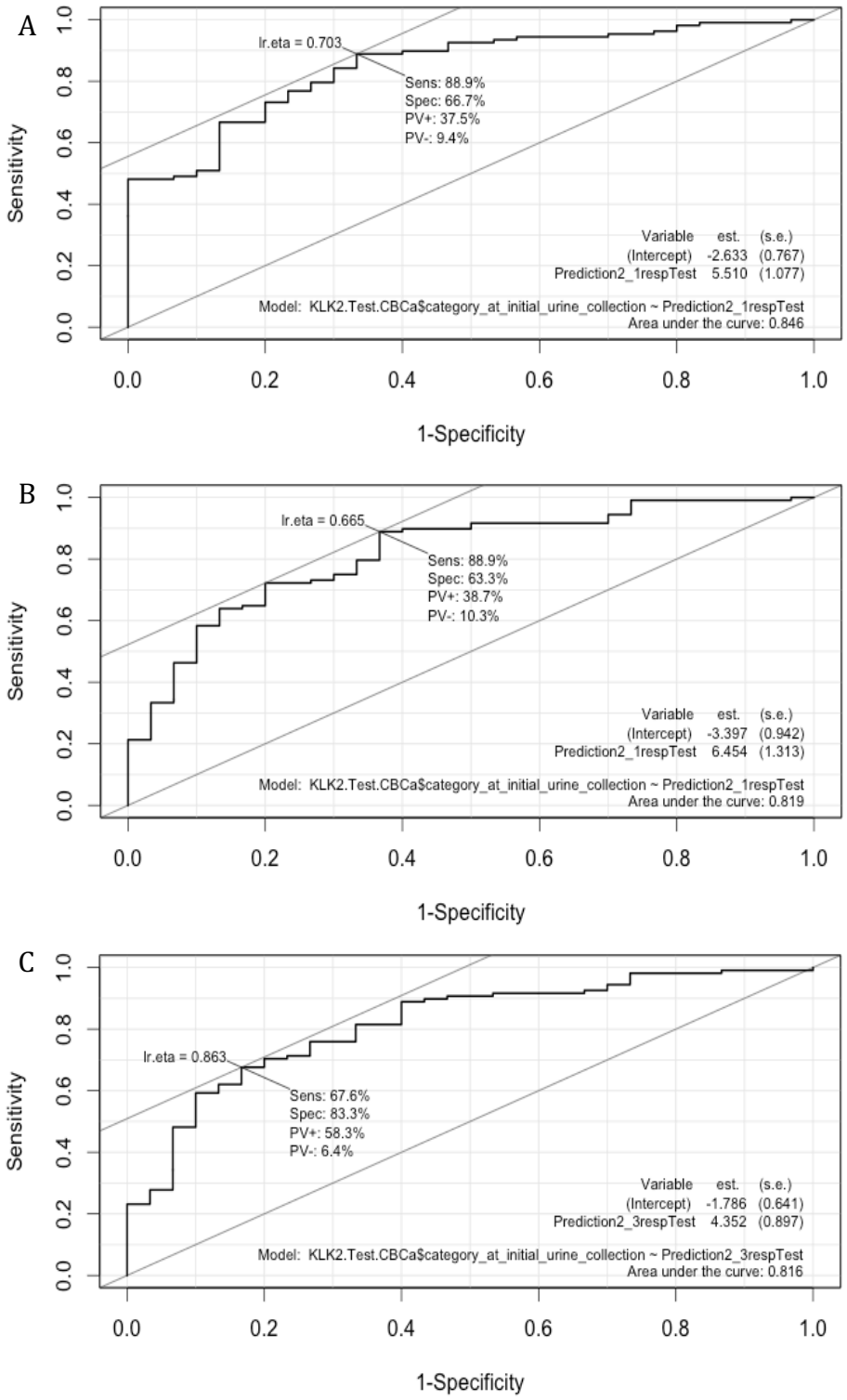
8414							
EN2	***0.011***	***0.239***	***0.992***	***CD10***	***0.027***	***-0.046***	***0.994***
TMCC2	***0.011***	***4.237***	***0.992***	***CDKN3***	***0.028***	***0.045***	***0.994***
ITGBL1	***0.011***	***0.388***	***0.992***	***SSTR1***	***0.029***	***0.061***	***0.994***
PECI	***0.014***	***0.018***	***0.992***	***TMCC2***	***0.031***	***0.020***	***0.994***
MMP25	***0.015***	***0.889***	***0.992***	***MIR146A.DQ65 8414***	***0.031***	***0.076***	***0.994***
LASS1	***0.015***	***0.199***	***0.992***	***ST6GALNAC1***	***0.032***	***0.030***	***0.994***
CASKIN1	***0.016***	***0.107***	***0.992***	***MMP26***	***0.033***	***0.047***	***0.994***
PALM3	***0.016***	***0.123***	***0.992***	***HIST1H1E***	***0.034***	***0.071***	***0.994***
HPRT	***0.017***	***0.060***	***0.992***	***TBP***	***0.036***	***0.063***	***0.994***
TMEM45B	***0.018***	***0.272***	***0.992***	***MKi67***	***0.038***	***0.070***	***0.994***
TMEM86A	***0.018***	***0.898***	***0.992***	***STOM***	***0.041***	***0.054***	***0.994***
MIR4435.1HG.lOC541471	***0.019***	***0.102***	***0.992***	***CADPS***	***0.048***	***0.032***	***0.994***
SChLAP1	***0.019***	***0.452***	***0.992***	***PTN***	***0.049***	***-0.048***	***0.994***
STOM	***0.021***	***NA***	***0.992***				
SFRP4	***0.022***	***0.456***	***0.992***				
FOLH1	***0.024***	***0.077***	***0.992***				
MNX1	***0.025***	***0.127***	***0.992***				
TWIST1	***0.026***	***0.103***	***0.992***				
CLIC2	***0.027***	***NA***	***0.992***				
VAX2	***0.034***	***0.170***	***0.992***				
PCSK6	***0.036***	***0.210***	***0.992***				
ACTR5	***0.036***	***0.153***	***0.992***				
CAMK2N2	***0.042***	***0.163***	***0.992***				
ABCB9	***0.042***	***NA***	***0.992***				
EIF2D	***0.042***	***0.054***	***0.992***				
HMBS	***0.043***	***0.107***	***0.992***				
B4GALNT4	***0.046***	***NA***	***0.992***				

Met	***0.046***	***1.819***	***0.992***
HIST3H2A	***0.047***	***0.065***	***0.992***
COL10A1	***0.048***	***0.191***	***0.992***

KLK3 Adjusted data				*HK normalised data*			
Transcript	***p-value***	***Log$_2$(FC)***	***Adjusted p-value***	***Transcript***	***p-value***	***Log$_2$(FC)***	***Adjusted p-value***
PCA3	***1.61x10-06***	***0.14***	***0.0003***	***ERG3' exons 4-5***	***4.58x10-09***	***0.699***	***7.64x10-07***
HPN	***3.27x10-05***	***0.13***	***0.01***	***PCA3***	***1.40x10-08***	***0.191***	***2.32x10-06***
SIM2.short	***0.0002***	***0.091***	***0.029***	***TMPRSS2:ERG***	***4.02x10-08***	***1.006***	***6.63x10-06***
ERG3' exons 4-5	***0.0002***	***0.080***	***0.031***	***ERG3' exons 6-7***	***4.79x10-07***	***1.130***	***7.86x10-05***
HOXC6	***0.001***	***0.113***	***0.084***	***HOXC6***	***3.71x10-06***	***0.178***	***0.001***
ERG3' exons 6-7	***0.001***	***0.062***	***0.117***	***TDRD***	***2.70x10-05***	***0.848***	***0.004***
AMACR	***0.001***	***0.086***	***0.118***	***HPN***	***0.0002***	***0.123***	***0.028***
TMPRSS2:ERG	***0.001***	***0.062***	***0.119***	***HOXC4***	***0.0003***	***0.200***	***0.046***
SMIM1	***0.001***	***0.124***	***0.217***	***DLX1***	***0.0004***	***0.424***	***0.057***
KLK4	***0.001***	***-0.099***	***0.219***	***APOC1***	***0.0004***	***0.390***	***0.057***
GJB1	***0.002***	***0.103***	***0.324***	***ERG5'***	***0.001***	***0.175***	***0.157***
TRPM4	***0.004***	***0.079***	***0.572***	***GJB1***	***0.001***	***0.129***	***0.182***
IMPDH2	***0.004***	***0.102***	***0.627***	***MCTP1***	***0.001***	***0.333***	***0.183***
MYOF	***0.005***	***-0.098***	***0.732***	***ISX***	***0.002***	***0.190***	***0.247***
RAB17	***0.005***	***0.106***	***0.738***	***SSTR1***	***0.002***	***0.035***	***0.252***
SIM2.long	***0.005***	***0.105***	***0.754***	***PPFIA2***	***0.002***	***0.312***	***0.255***
AMH	***0.005***	***0.075***	***0.787***	***TRPM4***	***0.002***	***0.294***	***0.326***
PTN	***0.007***	***-0.071***	***0.998***	***RAB17***	***0.002***	***0.111***	***0.366***
CAMKK2	***0.008***	***0.052***	***0.998***	***SIM2.long***	***0.003***	***0.079***	***0.432***
ISX	***0.008***	***0.062***	***0.998***	***SLC12A1***	***0.004***	***0.233***	***0.560***
DLX1	***0.009***	***0.042***	***0.998***	***SIM2.short***	***0.004***	***0.064***	***0.566***
MMP11	***0.009***	***0.070***	***0.998***	***AMACR***	***0.004***	***0.139***	***0.591***
TDRD	***0.010***	***0.042***	***0.998***	***MMP11***	***0.005***	***0.066***	***0.665***

ERG5	***0.012***	***0.055***	***0.998***	***ANKRD34B***	***0.005***	***0.099***	***0.724***
GAPDH	***0.012***	***0.078***	***0.998***	***DNAH5***	***0.006***	***0.266***	***0.795***
SULT1A1	***0.013***	***0.063***	***0.998***	***AMH***	***0.007***	***0.046***	***0.994***
IGFBP3	***0.014***	***-0.033***	***0.998***	***RP11_97O12.7***	***0.008***	***0.050***	***0.994***
RIOK3	***0.015***	***0.063***	***0.998***	***MEX3A***	***0.009***	***0.184***	***0.994***
TWIST1	***0.016***	***0.032***	***0.998***	***PVT1***	***0.011***	***0.079***	***0.994***
RP11.97O12.7	***0.016***	***0.066***	***0.998***	***SMIM1***	***0.011***	***0.082***	***0.994***
ANKRD34B	***0.016***	***0.069***	***0.998***	***EN2***	***0.011***	***0.059***	***0.994***
DNAH5	***0.017***	***0.052***	***0.998***	***CASKIN1***	***0.013***	***0.035***	***0.994***
CD10	***0.017***	***-0.040***	***0.998***	***KLK4***	***0.014***	***-0.031***	***0.994***
MARCH5	***0.018***	***-0.077***	***0.998***	***ITGBL1***	***0.015***	***0.092***	***0.994***
GABARAPL2	***0.019***	***-0.073***	***0.998***	***NEAT1***	***0.015***	***0.112***	***0.994***
APOC1	***0.019***	***0.051***	***0.998***	***SULT1A1***	***0.017***	***0.068***	***0.994***
SLC12A1	***0.022***	***0.049***	***0.998***	***CDKN3***	***0.019***	***0.080***	***0.994***
SSTR1	***0.022***	***0.025***	***0.998***	***RIOK3***	***0.022***	***0.023***	***0.994***
CLU	***0.025***	***0.033***	***0.998***	***MIR146A***	***0.023***	***0.079***	***0.994***
ITGBL1	***0.025***	***0.063***	***0.998***	***TMEM45B***	***0.023***	***0.038***	***0.994***
EN2	***0.026***	***0.049***	***0.998***	***NAALADL2***	***0.025***	***0.061***	***0.994***
RPS11	***0.026***	***-0.083***	***0.998***	***TWIST1***	***0.029***	***0.003***	***0.994***
RNF157	***0.026***	***0.043***	***0.998***	***RPL23AP53***	***0.030***	***0.181***	***0.994***
MNX1	***0.026***	***0.018***	***0.998***	***PALM3***	***0.033***	***0.061***	***0.994***
PVT1	***0.035***	***0.054***	***0.998***	***SULF2***	***0.035***	***0.055***	***0.994***
MIC1	***0.043***	***0.057***	***0.998***	***COL9A2***	***0.035***	***0.140***	***0.994***
CASKIN1	***0.044***	***0.038***	***0.998***	***RNF157***	***0.035***	***0.180***	***0.994***
MIR146A.DQ658414	***0.044***	***0.072***	***0.998***	***CLU***	***0.037***	***0.078***	***0.994***
STOM	***0.046***	***0.019***	***0.998***	***MIR4435_1HG***	***0.039***	***0.086***	***0.994***
				MMP25	***0.040***	***0.030***	***0.994***
				MIC1	***0.040***	***0.102***	***0.994***
				RPS11	***0.040***	***-0.007***	***0.994***
				IMPDH2	***0.041***	***0.055***	***0.994***

	MKi67	***0.042***	***0.371***	***0.994***
	TMCC2	***0.043***	***0.029***	***0.994***

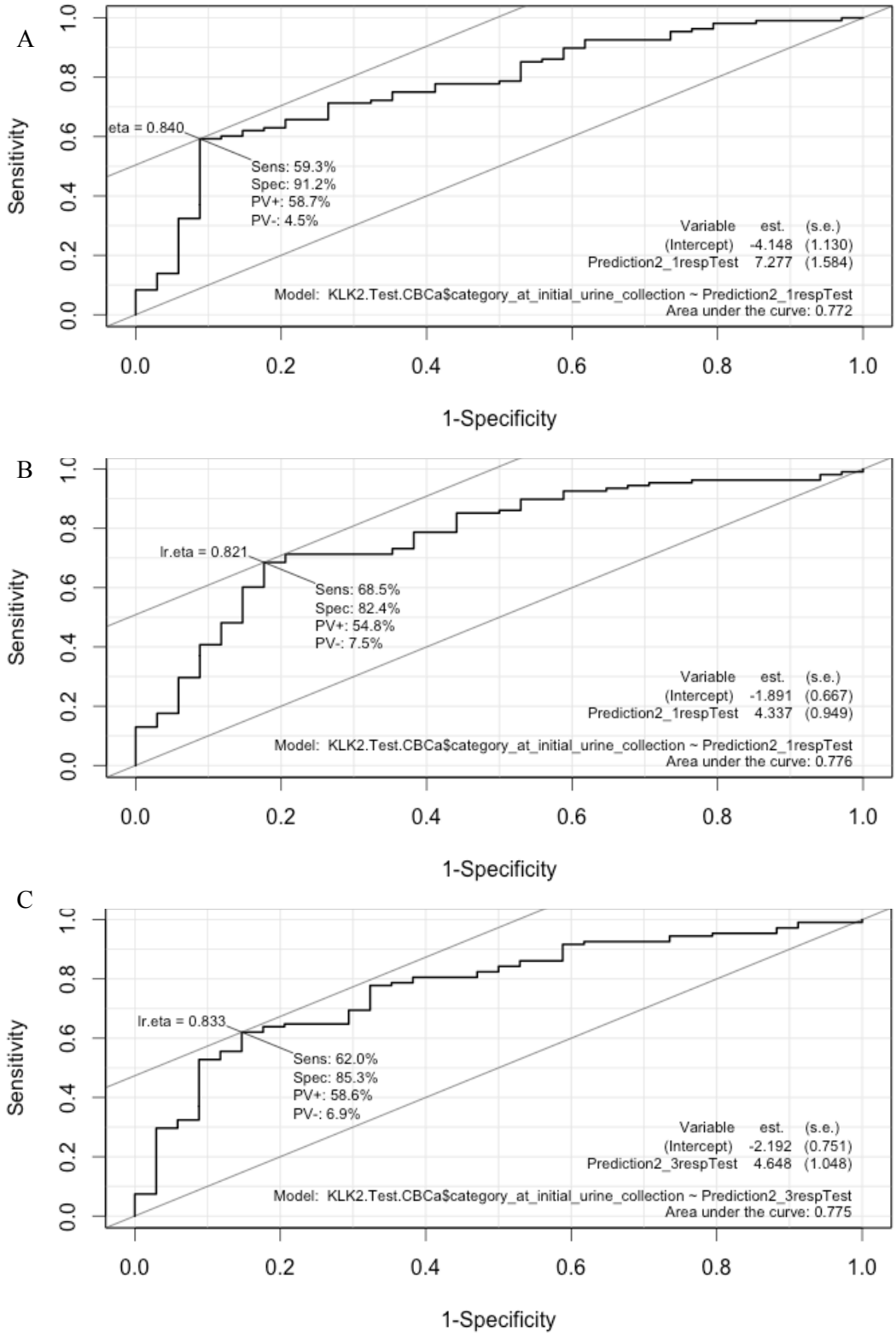


Supplementary Figure 1 KLK2 Ratio Data ROC curves for test data using models detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 5 Lasso output for models detecting between CB and Ca (L I H) using KLK2 ratio data.

All Transcripts	*Significant Transcripts*	*Multiple testing corrected*

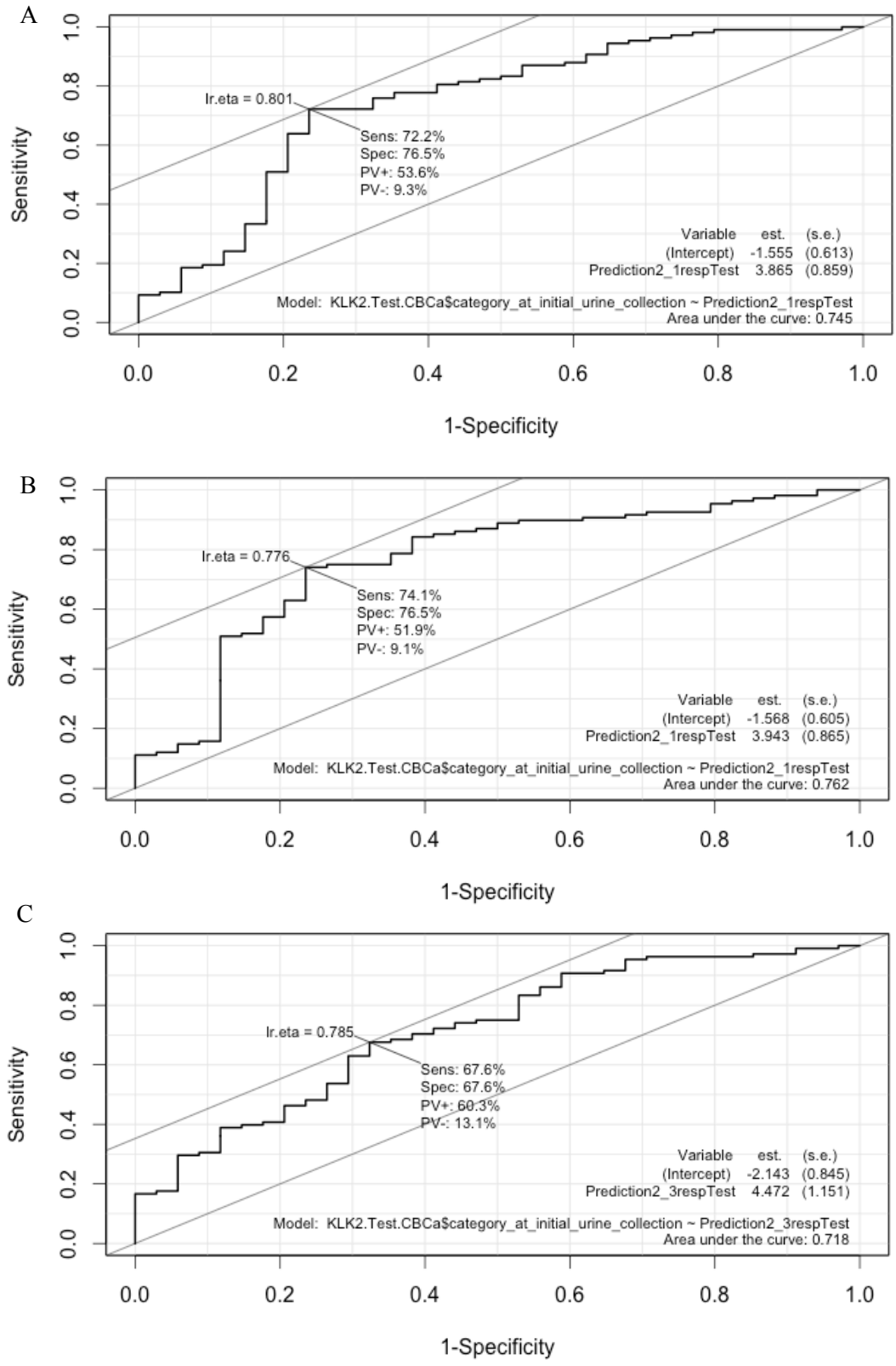
				Transcripts	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
ERG3' exons 4-5	***0.28***	***ERG3' exons 4-5***	***0.21***	***ERG3' exons 4-5***	***0.25***
TMPRSS2:ERG	***0.22***	***TMPRSS2:ERG***	***0.20***	***TMPRSS2:ERG***	***0.24***
PCA3	***0.20***	***PCA3***	***0.17***	***PCA3***	***0.20***
HOXC6	***0.08***	***HOXC6***	***0.07***	***HOXC6***	***0.08***
ISX	***0.08***			***ISX***	***0.03***
APOC1	***0.06***			***GJB1***	***0.02***
GJB1	***0.06***			***DLX1***	***0.01***
AMACR	***0.05***			***TDRD***	***0.01***
NEAT1	***0.03***				
DLX1	***0.02***				
TDRD	***0.02***				
TMEM47	***0.01***				
SULT1A1	***0.01***				
RNF157	***0.01***				
ST6GALNAC1	***0.00***				
IGFBP3	***-0.01***				
ARexon9	***-0.06***				
PPP1R12B	***-0.08***				
CP	***-0.11***				
MXI1	***-0.16***				
KLK4	***-0.24***				



Supplementary Figure 2 KLK2 Adjusted Data ROC curves for test data using models detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 6 Lasso output for models detecting between CB and Ca (L I H) using KLK2 adjusted data.

All Transcript		*Significant Transcripts*		*Multiple Testing correction Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***2.89***	***PCA3***	***2.85***	***PCA3***	***4.59***
AMACR	***1.73***	***AMACR***	***2.68***	***SIM2.short***	***2.88***
SIM2.short	***1.07***	***ERG3' exons 4-5***	***1.77***	***ERG3' exons 4-5***	***2.50***
SMIM1	***0.89***	***SMIM1***	***1.43***	***AMACR***	***2.44***
AMH	***0.89***	***RP11.97O12.7***	***1.38***	***HPN***	***2.02***
ERG3' exons 4-5	***0.70***	***SIM2.short***	***1.25***	***SMIM1***	***1.89***
CLU	***0.49***	***CAMKK2***	***1.03***		
HPN	***0.46***	***AMH***	***0.93***		
CAMKK2	***0.39***	***CLU***	***0.90***		
GAPDH	***0.29***	***RNF157***	***0.68***		
RP11.97O12.7	***0.24***	***DNAH5***	***0.63***		
DNAH5	***0.20***	***RIOK3***	***0.52***		
RNF157	***0.18***	***NEAT1***	***0.43***		
APOC1	***0.17***	***APOC1***	***0.42***		
ERG3' exons 6-7	***0.17***	***DLX1***	***0.38***		
RIOK3	***0.05***	***TBP***	***0.33***		
MMP25	***0.02***	***MMP11***	***0.23***		
CP	***-0.01***	***SYNM***	***0.23***		
CD10	***-0.03***	***CADPS***	***0.18***		
AR exon 9	***-0.08***	***SLC12A1***	***0.15***		
PTN	***-0.42***	***MIC1***	***0.11***		
IGFBP3	***-0.45***	***HPN***	***0.10***		
MYOF	***-0.68***	***STOM***	***0.08***		
GABARAPL2	***-0.75***	***MKi67***	***0.05***		
KLK4	***-1.09***	***RPS11***	***-0.13***		
MARCH5	***-1.11***	***CD10***	***-0.98***		
		IGFBP3	***-1.17***		
		PTN	***-1.25***		
		MYOF	***-1.66***		
		KLK4	***-1.87***		
		GABARAPL2	***-2.90***		

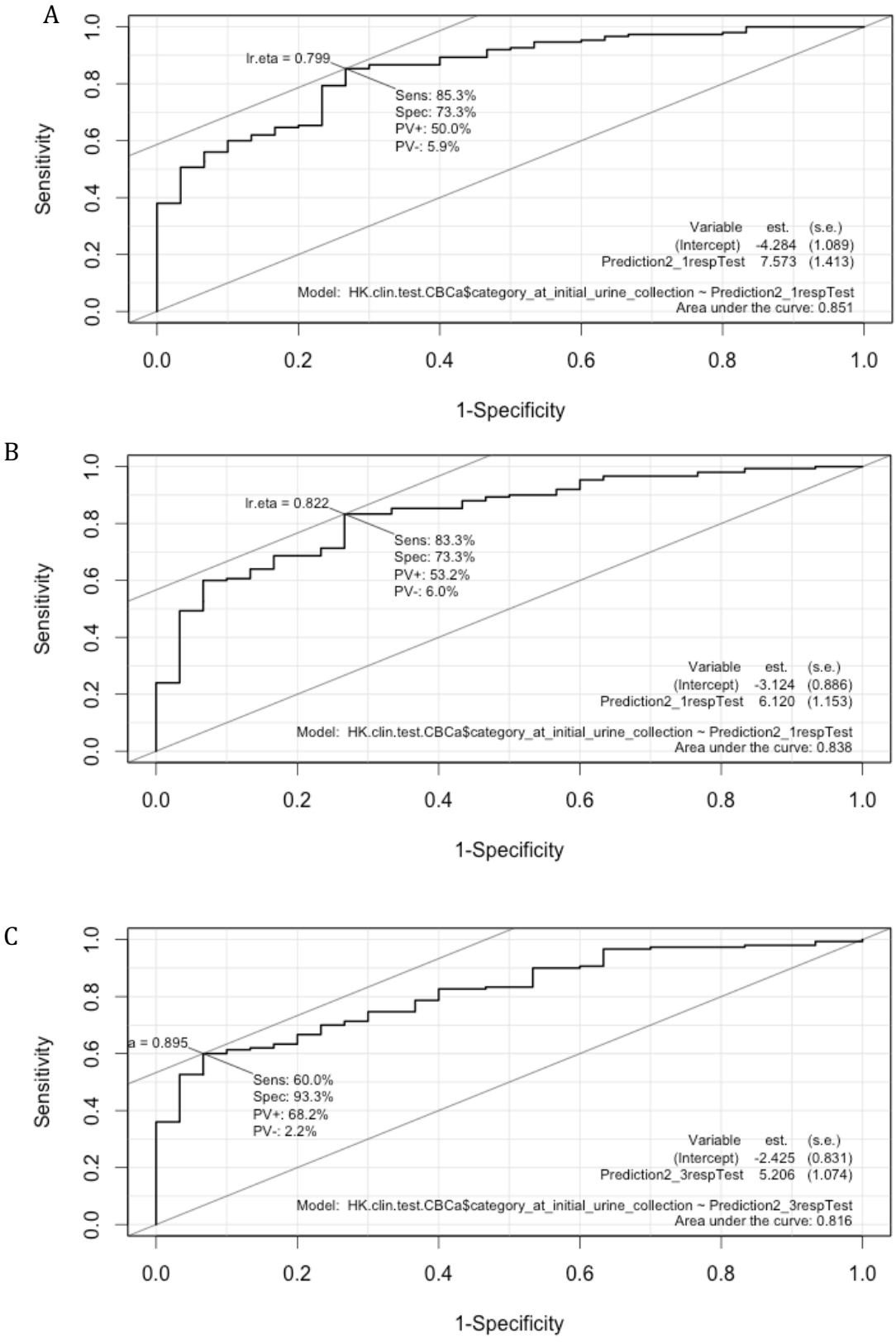


Supplementary Figure 3 KLK3 Adjusted Data ROC curves for test data using models detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 7 Lasso output for models detecting between CB and Ca (L I H) using KLK3 adjusted data

All Transcript		*Significant Transcripts*		*Multiple Testing correction Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***2.94***	***PCA3***	***2.88***	***PCA3***	***4.40***
ERG3' exons 4-5	***1.74***	***AMACR***	***2.30***	***SIM2.short***	***2.87***
AMACR	***1.57***	***ERG3' exons 4-5***	***2.01***	***HPN***	***2.15***
SIM2.short	***1.55***	***SMIM1***	***1.25***	***ERG3' exons 4-5***	***1.97***
SMIM1	***1.24***	***SIM2.short***	***1.19***		
AMH	***1.23***	***AMH***	***0.82***		
APOC1	***0.63***	***CLU***	***0.71***		
NEAT1	***0.60***	***RIOK3***	***0.70***		
MMP25	***0.59***	***CAMKK2***	***0.69***		
TBP	***0.55***	***APOC1***	***0.64***		
SERPINB5	***0.52***	***HPN***	***0.62***		
HPN	***0.47***	***RNF157***	***0.52***		
CLU	***0.34***	***DLX1***	***0.36***		
DLX1	***0.31***	***MMP11***	***0.30***		
RNF157	***0.30***	***SLC12A1***	***0.22***		
CAMKK2	***0.29***	***SULT1A1***	***0.22***		
PPAP2A	***0.26***	***ISX***	***0.09***		
MMP11	***0.25***	***DNAH5***	***0.09***		
SLC12A1	***0.19***	***EN2***	***0.07***		
STOM	***0.17***	***STOM***	***0.06***		
CADPS	***0.15***	***ANKRD34B***	***0.02***		
RIOK3	***0.15***	***CD10***	***-0.47***		
EN2	***0.13***	***RPS11***	***-0.53***		
ISX	***0.13***	***IGFBP3***	***-1.08***		
COL10A1	***0.12***	***KLK4***	***-1.17***		
ST6GALNAC1	***0.12***	***MYOF***	***-1.46***		
MNX1	***0.11***	***PTN***	***-1.57***		
DNAH5	***0.11***	***GABARAPL2***	***-2.02***		
SULT1A1	***0.08***	***MARCH5***	***-2.04***		
HOXC6	***0.07***				
GJB1	***0.04***				
ERG5	***0.03***				
RP11.244H18.1.P712P	***-0.14***				
SPON2	***-0.16***				
CLIC2	***-0.20***				
PPP1R12B	***-0.21***				
CD10	***-0.23***				
CP	***-0.27***				
AR exon 9	***-0.30***				
MXI1	***-0.32***				
CDC20	***-0.39***				
CKAP2L	***-0.43***				

Timp4	***-0.45***
RPS11	***-0.61***
IGFBP3	***-0.90***
MYOF	***-1.18***
PTN	***-1.29***
KLK4	***-1.36***
MARCH5	***-1.51***
GABARAPL2	***-1.54***

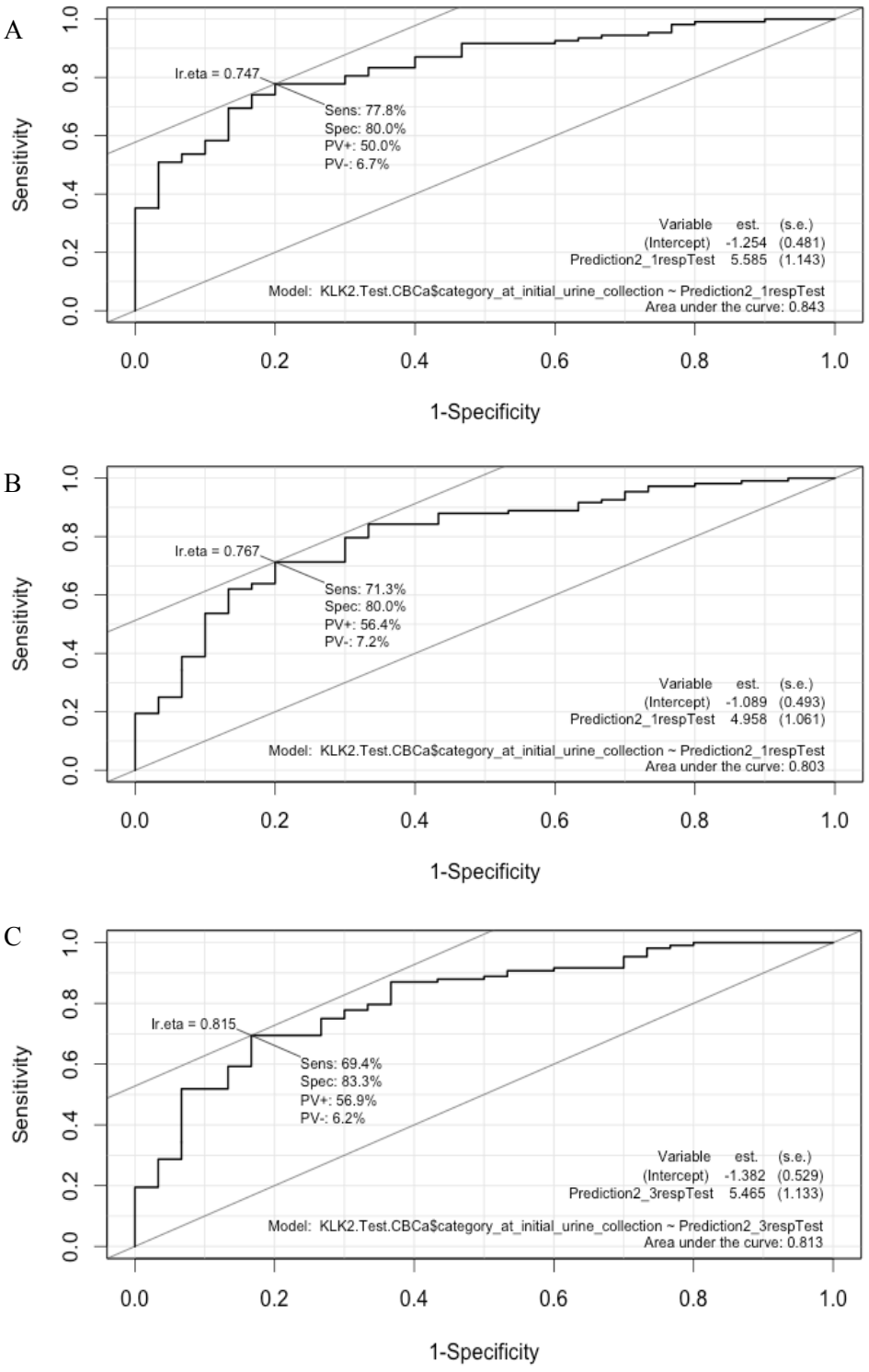


Supplementary Figure 4 HK normalised data ROC curves for test data using models detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 8 Lasso output for models detecting between CB and Ca (L I H) using HK normalised data.

All Transcript		*Significant Transcripts*		*Multiple Testing correction Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***0.29***	***PCA3***	***0.35***	***PCA3***	***0.29***
TMPRSS2:ERG	***0.18***	***ERG3' exons 4-5***	***0.19***	***TMPRSS2:ERG***	***0.23***
ERG3' exons 4-5	***0.17***	***TMPRSS2:ERG***	***0.18***	***ERG3' exons 4-5***	***0.19***
APOC1	***0.11***	***APOC1***	***0.13***	***HPN***	***0.04***
ISX	***0.04***	***SLC12A1***	***0.05***	***HOXC6***	***0.03***
SLC12A1	***0.04***	***ISX***	***0.04***		
HOXC6	***0.04***	***MCTP1***	***0.03***		
MCTP1	***0.03***	***HOXC6***	***0.02***		
TDRD	***0.00***	***SULT1A1***	***0.00***		
PDLIM5	***-0.01***	***KLK4***	***-0.41***		
CD10	***-0.02***				
GABARAPL2	***-0.02***				
PTN	***-0.02***				
AR exon 9	***-0.04***				
PPP1R12B	***-0.04***				
CP	***-0.08***				
MXI1	***-0.15***				
KLK4	***-0.20***				

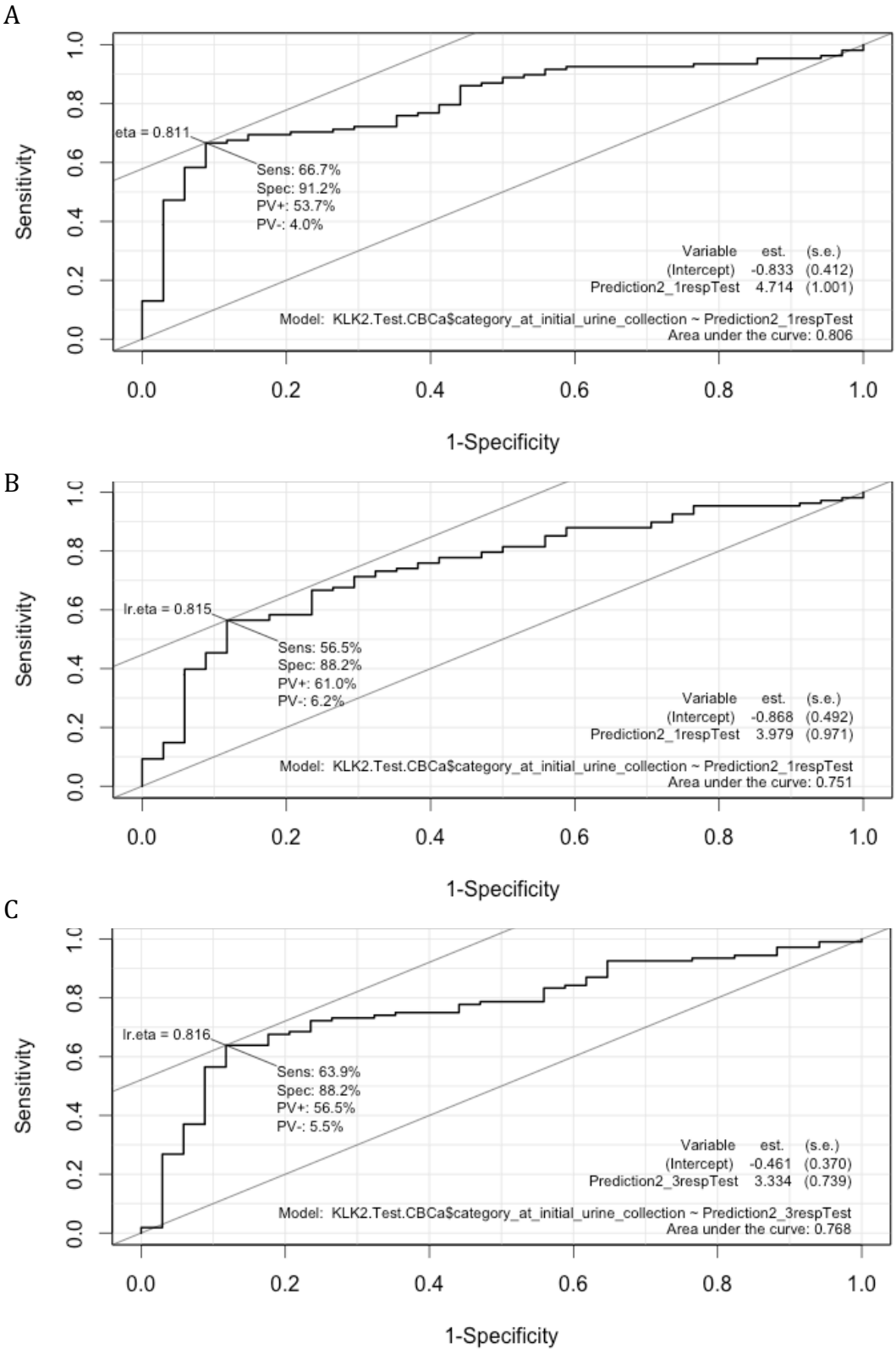
6.13 Binomial Testing between CB and Ca (Random Sampling)



Supplementary Figure 5 KLK2 ratio data ROC curves for test data using models (random sampling) detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 9 Lasso output for models (random sampling detecting between CB and Ca (L I H) using KLK2 ratio data.

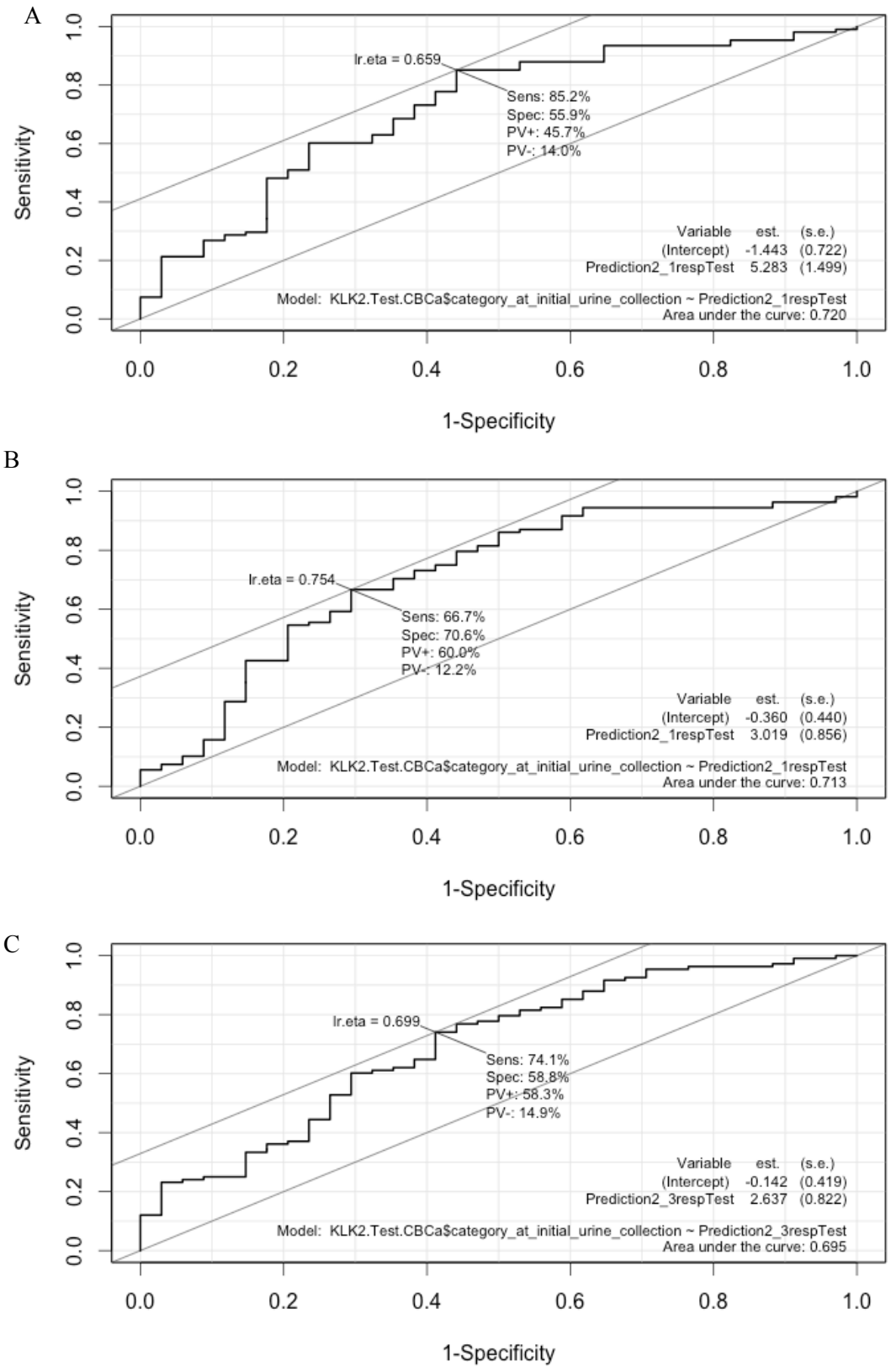
All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
ERG3' exons 4-5	***0.51***	***PCA3***	***0.21***	***ERG3' exons 4-5***	***0.25***
PCA3	***0.14***	***ERG3' exons 4-5***	***0.20***	***PCA3***	***0.24***
TMPRSS2:ERG	***0.14***	***TMPRSS2:ERG***	***0.15***	***TMPRSS2:ERG***	***0.17***
SLC12A1	***0.06***	***AMACR***	***0.08***	***HOXC6***	***0.02***
ERG5	***0.05***	***GJB1***	***0.06***	***GJB1***	***0.01***
GJB1	***0.04***	***NEAT1***	***0.03***		
HOXC6	***0.04***	***TDRD***	***0.03***		
TDRD	***0.01***	***DLX1***	***0.02***		
LASS1	***0.00***	***TRPM4***	***0.01***		
HIST1H2BF	***-0.01***				
CP	***-0.02***				
CKAP2L	***-0.03***				
DPP4	***-0.04***				
PTN	***-0.07***				
ZNF577	***-0.08***				
MYOF	***-0.10***				
GABARAPL2	***-0.31***				



Supplementary Figure 6 KLK2 Adjusted Data ROC curves for test data using models (random sampling) detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 10 Lasso output for models (random sampling) detecting between CB and Ca (L I H) using KLK2 adjusted data.

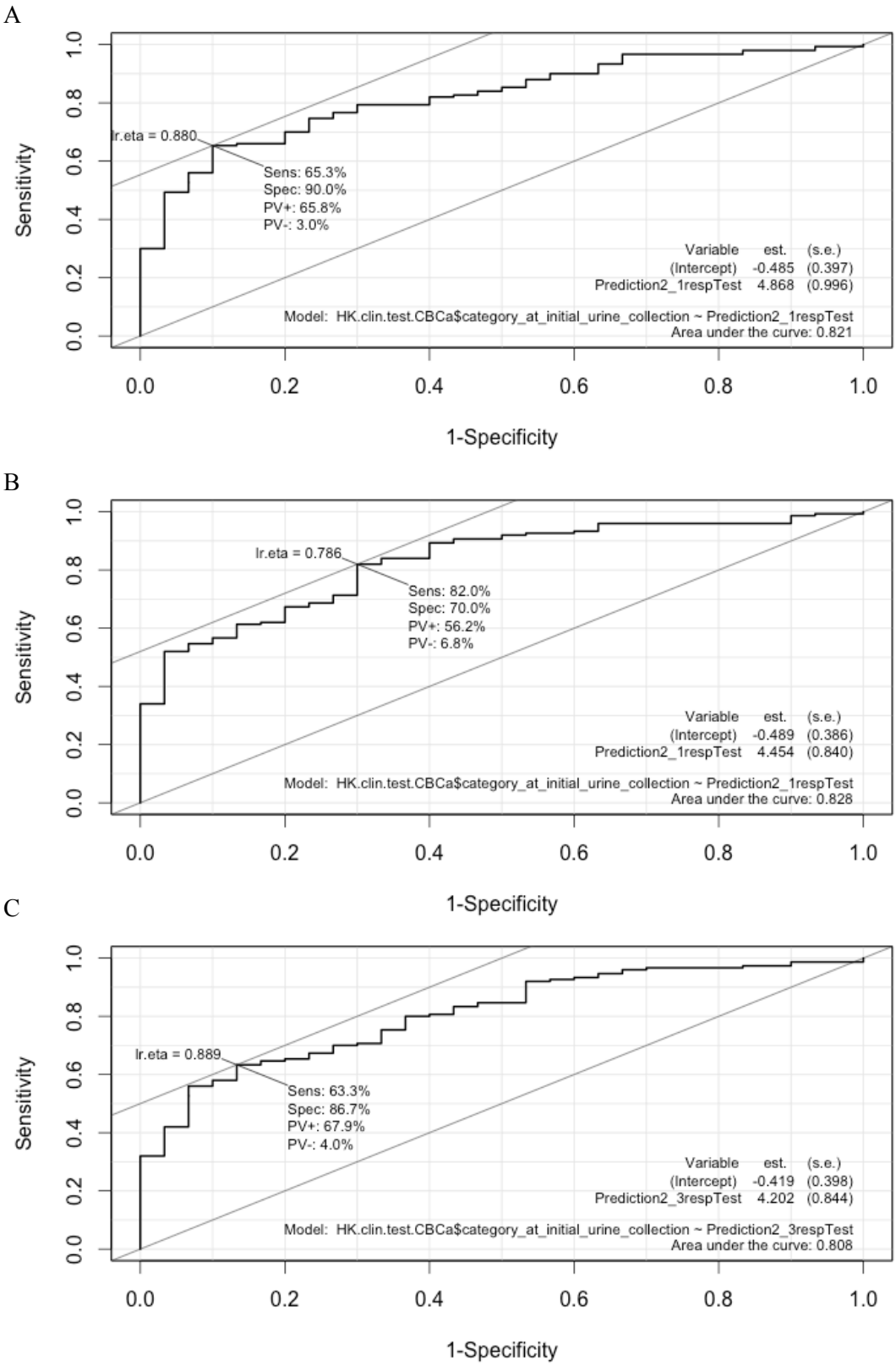
All Transcript		*Significant Transcripts*		*Multiple Testing correction Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
SIM2.short	***4.95***	***AMACR***	***2.32***	***SMIM1***	***6.89***
SMIM1	***2.78***	***SMIM1***	***2.20***	***PCA3***	***6.59***
ERG3' exons 6-7	***1.95***	***MMP11***	***1.69***	***SIM2.short***	***3.42***
AMH	***1.76***	***SIM2.short***	***1.54***	***AMACR***	***2.95***
HPN	***1.26***	***TMPRSS2:ERG***	***1.33***	***HPN***	***1.88***
PCA3	***1.20***	***HPN***	***1.31***	***ERG3' exons 4-5***	***1.18***
NEAT1	***1.06***	***ISX***	***0.93***		
PCSK6	***1.02***	***CLU***	***0.89***		
DNAH5	***0.68***	***DLX1***	***0.46***		
TMPRSS2:ERG	***0.66***	***APOC1***	***0.36***		
SEC61A1	***0.53***	***GJB1***	***0.30***		
HIST1H2BF	***0.47***	***CASKIN1***	***0.22***		
CADPS	***0.46***	***MIR146A.DQ658414***	***0.15***		
APOC1	***0.45***	***HOXC6***	***0.08***		
TBP	***0.37***	***PTN***	***-0.38***		
ERG5	***0.34***	***IGFBP3***	***-0.44***		
CAMKK2	***0.32***	***GABARAPL2***	***-0.69***		
CAMK2N2	***0.22***	***MYOF***	***-1.43***		
TMCC2	***0.19***	***KLK4***	***-1.86***		
SERPINB5	***0.12***				
EN2	***0.12***				
ERG3' exons 4-5	***0.03***				
SChLAP1	***0.00***				
PTN	***-0.01***				
PPP1R12B	***-0.15***				
SIRT1	***-0.30***				
PTPRC	***-0.36***				
IGFBP3	***-0.46***				
CD10	***-0.61***				
SNCA	***-0.68***				
MEMO1	***-0.75***				
RPLP2	***-1.46***				
MYOF	***-1.88***				
SACM1L	***-2.85***				
KLK4	***-4.13***				



Supplementary Figure 7 KLK3 Adjusted Data ROC curves for test data using models (random sampling) detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 11 Lasso output for models (random sampling) detecting between CB and Ca (L I H) using KLK3 adjusted data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***1.57***	***SMIM1***	***2.26***	***HPN***	***3.49***
SMIM1	***1.27***	***SIM2.short***	***1.84***	***SIM2.short***	***3.49***
SIM2.short	***1.21***	***SULT1A1***	***1.48***	***PCA3***	***3.33***
HPN	***0.76***	***ERG3' exons 4-5***	***1.42***	***ERG3' exons 4-5***	***2.55***
ERG3' exons 4-5	***0.57***	***GAPDH***	***1.37***		
GAPDH	***0.43***	***PCA3***	***1.34***		
AMH	***0.41***	***AMACR***	***0.85***		
HOXC6	***0.29***	***HPN***	***0.69***		
CLU	***0.16***	***MMP25***	***0.64***		
ISX	***0.15***	***ERG3' exons 6-7***	***0.52***		
MMP25	***0.09***	***CLU***	***0.48***		
APOC1	***0.05***	***GJB1***	***0.45***		
TMPRSS2:ERG	***0.04***	***ANKRD34B***	***0.27***		
MYOF	***-0.08***	***STOM***	***0.20***		
GABARAPL2	***-0.36***	***RAB17***	***0.06***		
KLK4	***-0.45***	***IGFBP3***	***-0.33***		
		RPS11	***-0.48***		
		PTN	***-0.78***		
		MYOF	***-1.08***		
		GABARAPL2	***-2.60***		



Supplementary Figure 8 GAPDH and RPLP2 Normalised Data ROC curves for test data using models (random sampling) detecting between CB and Ca (L I H) for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 12 Lasso output for models (random sampling) detecting between CB and Ca (L I H) using HK normalised data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***0.33***	***PCA3***	***0.63***	***PCA3***	***0.39***
ERG3' exons 4-5	***0.30***	***TMPRSS2:ERG***	***0.27***	***ERG3' exons 4-5***	***0.22***
TMPRSS2:ERG	***0.24***	***SMIM1***	***0.18***	***TMPRSS2:ERG***	***0.17***
TDRD	***0.04***	***TDRD***	***0.02***	***HOXC6***	***0.12***
CLU	***0.01***	***HOXC6***	***0.02***	***TDRD***	***0.06***
MMP25	***0.00***	***ERG5'***	***0.01***		
ALAS1	***-0.01***	***KLK4***	***-0.21***		
PDLIM5	***-0.09***				

6.14 Binomial Testing between CB and High-risk Ca

Supplementary Table 13 Glm test significant probes between CB and High-risk Ca

KLK2 ratio data				KLK2 adjusted data			
Transcript	***p*-value**	**Log_2(FC)**	**Adjusted *p*-value**	**Transcript**	***p*-value**	**Log_2(FC)**	**Adjusted *p*-value**
ERG3' exons 4-5	7.00E-07	1.868	0.0001	*HPN*	3.77x10-06	0.241	0.001
ERG3' exons 6-7	7.15E-07	2.966	0.0001	*PCA3*	4.78x10-06	0.222	0.001
PCA3	9.17E-07	0.376	0.0002	*GJB1*	0.0001	0.159	0.018
APOC1	7.71E-06	1.472	0.001	*AMACR*	0.0001	0.130	0.021
HPN	8.27E-06	0.352	0.001	*KLK4*	0.0003	-0.127	0.044
TMPRSS2:ERG	9.85E-06	NA	0.002	*ERG3' exons 4-5*	0.0004	0.109	0.063
HOXC6	2.14E-05	0.301	0.003	*ERG3' exons 6-7*	0.001	0.112	0.098
TDRD	2.54E-05	3.689	0.004	*TMPRSS2:ERG*	0.001	0.114	0.208
DLX1	4.09E-05	4.487	0.006	*HOXC6*	0.001	0.129	0.221
AMACR	7.47E-05	0.341	0.012	*RAB17*	0.002	0.227	0.272
GJB1	9.73E-05	0.320	0.015	*APOC1*	0.002	0.209	0.372
ANKRD34B	0.0002	5.892	0.025	*DLX1*	0.002	0.074	0.372
TRPM4	0.0002	0.730	0.029	*SPINK1*	0.003	0.163	0.445
MCTP1	0.0003	1.149	0.041	*MYOF*	0.003	-0.155	0.463
PPFIA2	0.0003	0.807	0.041	*SULT1A1*	0.003	0.126	0.509
ITGBL1	0.0003	0.799	0.042	*DPP4*	0.004	-0.105	0.552
HOXC4	0.0004	0.938	0.063	*ITGBL1*	0.004	0.087	0.611
SLC12A1	0.0004	1.022	0.064	*AR exons 4-8*	0.004	-0.121	0.637
ISX	0.001	2.371	0.077	*TRPM4*	0.004	0.080	0.640
RAB17	0.001	0.303	0.097	*CD10*	0.005	-0.092	0.771
VPS13A	0.001	0.110	0.118	*GABARAPL2*	0.006	-0.137	0.863
NEAT1	0.001	0.186	0.131	*RP11.244H18.1.P712P*	0.006	-0.100	0.890
STOM	0.001	NA	0.134	*TDRD*	0.007	0.064	0.983
PVT1	0.001	0.383	0.208	*UPK2*	0.007	0.108	0.996
SSTR1	0.001	0.369	0.209	*SLC12A1*	0.007	0.148	0.999

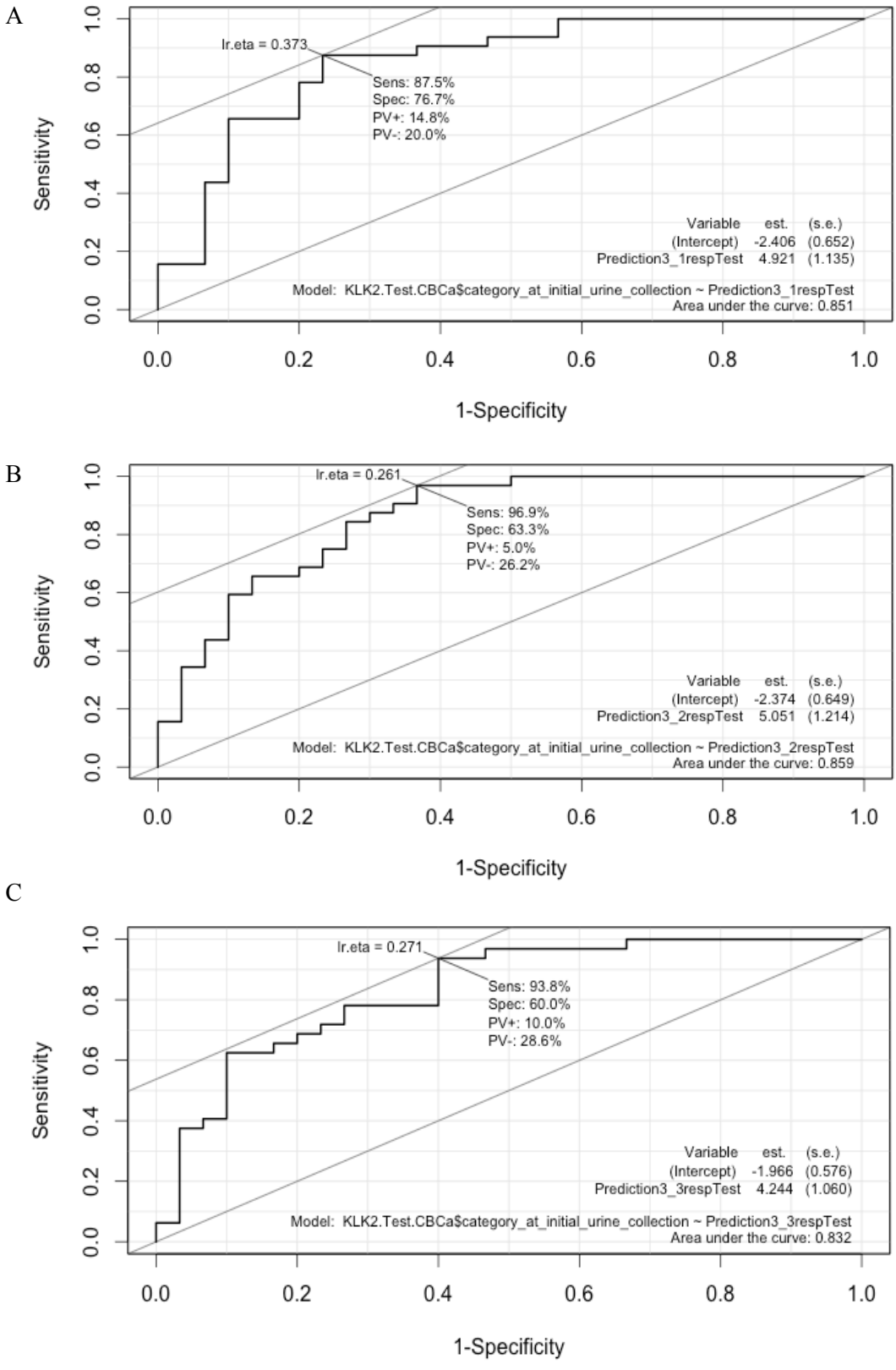
Met	0.002	2.490	0.266	*MIR4435.1HG.lOC541471*	0.008	0.087	0.999
SIM2.short	0.002	0.507	0.274	*GAPDH*	0.011	0.111	0.999
CDKN3	0.002	0.380	0.275	*RP11.97O12.7*	0.011	0.116	0.999
ERG5	0.002	NA	0.276	*STOM*	0.011	0.116	0.999
SPINK1	0.002	0.265	0.294	*SMIM1*	0.012	0.111	0.999
SULT1A1	0.002	0.225	0.323	*ANKRD34B*	0.012	0.134	0.999
TMEM45B	0.002	0.410	0.325	*NEAT1*	0.012	0.094	0.999
UPK2	0.002	0.676	0.334	*SIM2.short*	0.019	0.077	0.999
AMH	0.003	0.404	0.348	*MCTP1*	0.024	0.094	0.999
MIR146A.DQ658414	0.003	0.549	0.352	*MED4*	0.029	-0.054	0.999
SULF2	0.003	2.000	0.352	*DNAH5*	0.029	0.048	0.999
RP11.97O12.7	0.003	0.086	0.355	*ISX*	0.029	0.081	0.999
MMP11	0.003	0.451	0.382	*PPFIA2*	0.031	0.072	0.999
TMCC2	0.003	4.761	0.436	*Met*	0.035	0.098	0.999
PALM3	0.004	0.269	0.474	*SNCA*	0.038	-0.053	0.999
MIR4435.1HG.lOC541471	0.005	0.199	0.605	*VPS13A*	0.041	0.040	0.999
MIC1	0.005	0.445	0.642	*PTN*	0.044	-0.082	0.999
LASS1	0.005	0.473	0.665	*PVT1*	0.049	0.113	0.999
RIOK3	0.005	0.104	0.676				
MEX3A	0.006	0.801	0.694				
RPL23AP53	0.006	0.417	0.758				
CASKIN1	0.007	0.178	0.872				
TWIST1	0.008	0.192	0.893				
IMPDH2	0.008	0.099	0.894				
SIM2.long	0.009	0.133	0.966				
PECI	0.009	0.063	0.966				
GAPDH	0.009	0.067	0.966				
DNAH5	0.009	0.409	0.966				
EN2	0.009	0.358	0.966				
MKi67	0.010	-1.667	0.966				
NAALADL2	0.010	0.081	0.966				
SMIM1	0.010	0.137	0.966				
MMP26	0.011	0.411	0.966				
MNX1	0.011	0.241	0.966				

MMP25	0.012	0.932	0.966
HIST1H1C	0.012	0.055	0.966
SChLAP1	0.013	0.525	0.966
MGAT5B	0.013	0.258	0.966
PCSK6	0.014	0.219	0.966
CLIC2	0.014	NA	0.966
MCM7	0.015	0.273	0.966
MFSD2A	0.016	-2.398	0.966
TERT	0.017	0.153	0.966
HPRT	0.017	0.073	0.966
SSPO	0.017	0.221	0.966
HIST3H2A	0.020	0.091	0.966
ITPR1	0.022	0.069	0.966
B4GALNT4	0.022	NA	0.966
SLC4A1.S	0.022	NA	0.966
RPLP2	0.023	0.053	0.966
SACM1L	0.025	0.058	0.966
SYNM	0.025	0.214	0.966
VAX2	0.026	0.270	0.966
TMEM86A	0.026	0.795	0.966
RPS11	0.027	0.035	0.966
ABCB9	0.028	NA	0.966
CLU	0.030	0.248	0.966
CCDC88B	0.030	-5.601	0.966
HIST1H2BG	0.032	0.124	0.966
FOLH1	0.032	0.063	0.966
COL9A2	0.034	-2.208	0.966
BRAF	0.035	0.072	0.966
RPL18A	0.035	0.045	0.966
CAMKK2	0.036	0.085	0.966
AURKA	0.036	0.428	0.966
ARHGEF25	0.036	0.278	0.966
ALAS1	0.037	0.021	0.966
SFRP4	0.039	0.502	0.966

TERF2IP	0.041	0.031	0.966
PTPRC	0.046	NA	0.966
COL10A1	0.047	0.274	0.966
ACTR5	0.049	0.194	0.966
PSTPIP1	0.050	NA	0.966

KLK3 adjusted data				GAPDH and RPLP2 Normalised data			
Transcript	***p*-value**	**Log_2(FC)**	**Adjusted *p*-value**	**Transcript**	***p*-value**	**Log_2(FC)**	**Adjusted *p*-value**
HPN	1.32E-05	0.190	0.002	*ERG3' exons 4-5*	1.43E-06	0.793	0.000
PCA3	1.45E-05	0.184	0.002	*PCA3*	6.29E-06	0.196	0.001
KLK4	0.0002	-0.159	0.034	*TMPRSS2:ERG*	1.05E-05	0.953	0.002
ERG3' exons 4-5	0.0004	0.125	0.066	*ERG3' exons 6-7*	1.40E-05	1.265	0.002
GJB1	0.0005	0.142	0.075	*APOC1*	2.12E-05	0.505	0.003
AMACR	0.001	0.121	0.105	*HPN*	8.13E-05	0.149	0.013
ERG3' exons 6-7	0.001	0.088	0.106	*KLK4*	0.0005	-0.069	0.074
MYOF	0.001	-0.128	0.126	*HOXC6*	0.001	0.170	0.125
TMPRSS2:ERG	0.001	0.084	0.156	*TDRD*	0.002	0.800	0.245
ARexons4.8	0.001	-0.072	0.163	*SLC12A1*	0.002	0.371	0.273
HOXC6	0.002	0.120	0.273	*DLX1*	0.002	0.391	0.279
RP11.244H18.1.P712P	0.002	-0.096	0.300	*ITGBL1*	0.002	0.144	0.383
DPP4	0.002	-0.107	0.329	*MYOF*	0.005	-0.055	0.758
APOC1	0.002	0.162	0.347	*DPP4*	0.005	-0.039	0.762
DLX1	0.003	0.057	0.525	*SPINK1*	0.005	0.126	0.808
SULT1A1	0.004	0.110	0.532	*GABARAPL2*	0.005	-0.050	0.821
SPINK1	0.004	0.131	0.565	*RAB17*	0.005	0.125	0.826
ITGBL1	0.005	0.108	0.676	*CD10*	0.006	-0.074	0.967
RAB17	0.005	0.142	0.676	*HOXC4*	0.008	0.160	0.995
CD10	0.006	-0.095	0.856	*AR exons 4-8*	0.008	-0.059	0.995
KLK2	0.007	-0.080	0.952	*NEAT1*	0.010	0.126	0.995
GABARAPL2	0.007	-0.143	0.974	*UPK2*	0.010	0.306	0.995
SLC12A1	0.007	0.092	0.985	*PPFIA2*	0.011	0.325	0.995
UPK2	0.008	0.098	0.998	*GJB1*	0.012	0.127	0.995
STOM	0.012	0.089	0.998	*SRSF3*	0.014	-0.178	0.995

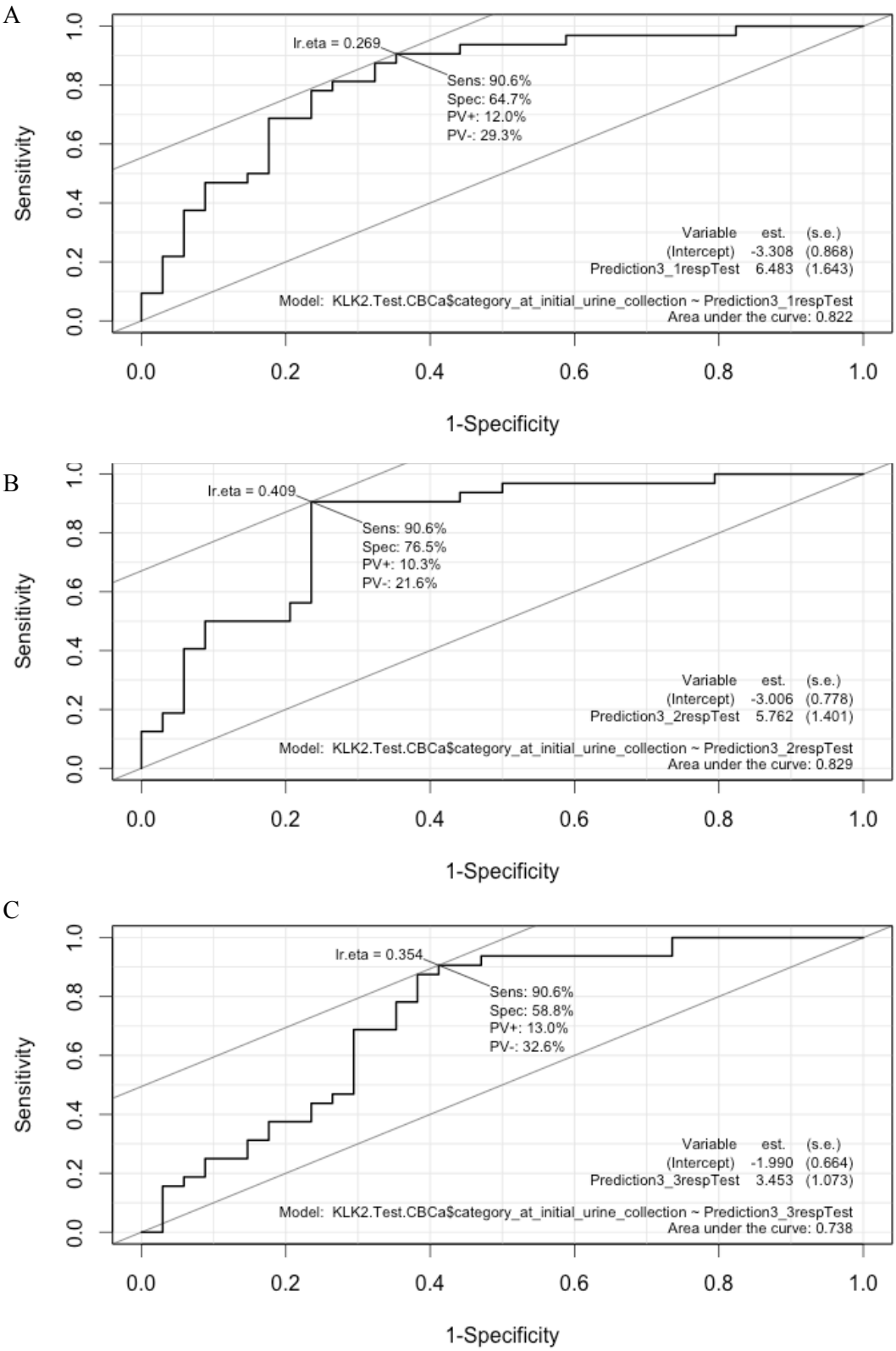
PTN	0.014	-0.104	0.998	*MCTP1*	0.014	0.370	0.995
MIR4435.1HG.lOC541471	0.014	0.122	0.998	*Met*	0.016	0.924	0.995
MED4	0.015	-0.061	0.998	*KLK2*	0.016	-0.043	0.995
TDRD	0.016	0.040	0.998	*AMACR*	0.016	0.140	0.995
SNCA	0.020	-0.058	0.998	*ANKRD34B*	0.018	0.083	0.995
TRPM4	0.022	0.053	0.998	*STOM*	0.023	0.180	0.995
NEAT1	0.027	0.059	0.998	*AR.ex9*	0.024	-0.448	0.995
MARCH5	0.030	-0.077	0.998	*MXI1*	0.025	-0.043	0.995
ANKRD34B	0.032	0.082	0.998	*P712P*	0.026	-0.054	0.995
MEMO1	0.032	-0.085	0.998	*STEAP2*	0.028	-0.032	0.995
SMIM1	0.035	0.103	0.998	*SULT1A1*	0.029	0.078	0.995
SIM2.short	0.039	0.028	0.998	*PDLIM5*	0.030	-0.042	0.995
RP11.97O12.7	0.040	0.063	0.998	*PTN*	0.031	-0.101	0.995
SRSF3	0.040	-0.094	0.998	*TRPM4*	0.032	0.219	0.995



Supplementary Figure 9 KLK2 Ratio Data ROC curves for test data using models detecting between CB and high risk Ca for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 14 Lasso output for models detecting between CB and high risk Ca using KLK2 ratio data.

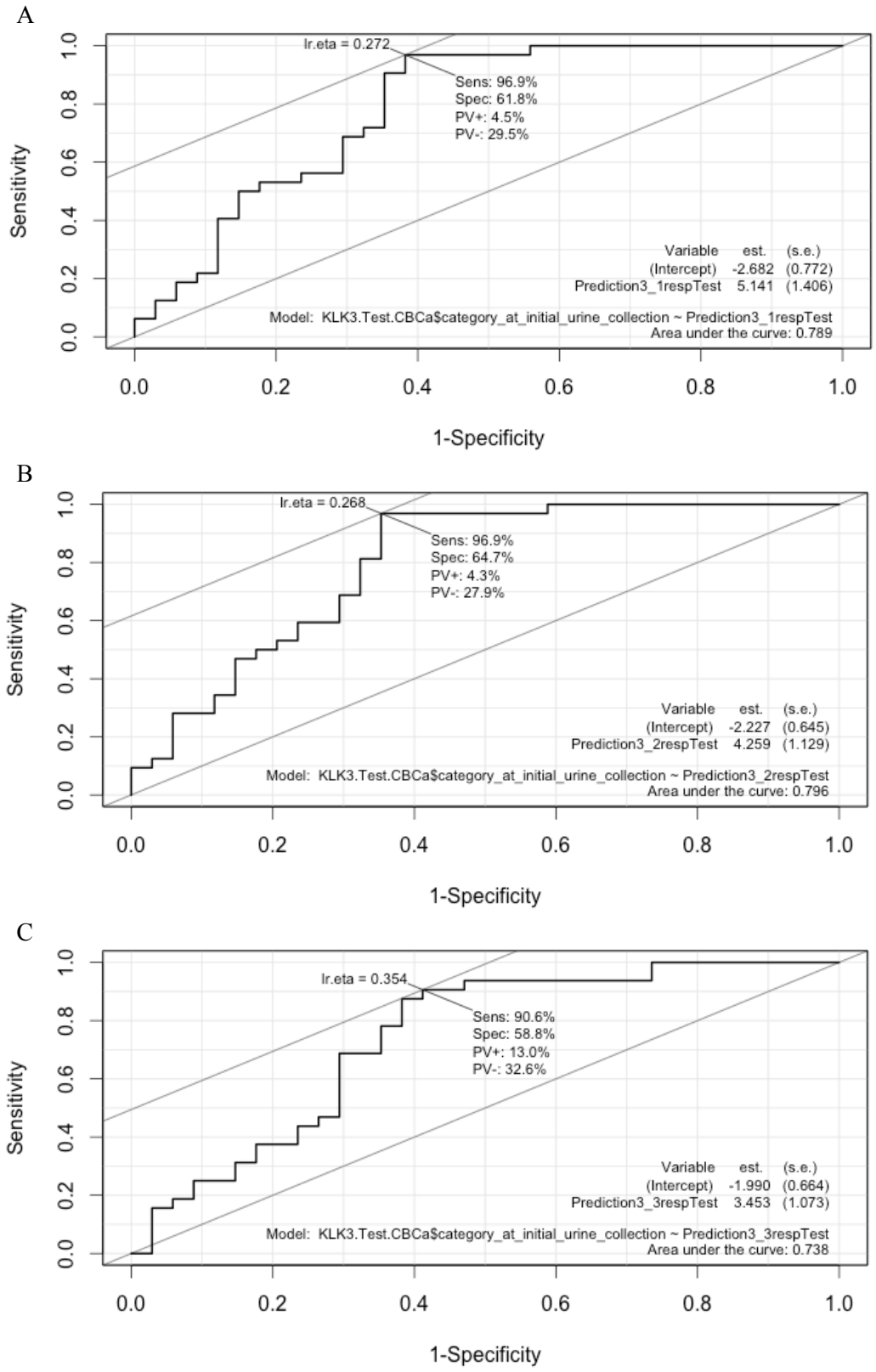
All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
ERG3' exons 4-5	***0.55***	***ERG3' exons 4-5***	***0.49***	***ERG3' exons 4-5***	***0.44***
PCA3	***0.22***	***APOC1***	***0.23***	***PCA3***	***0.20***
ANKRD34B	***0.17***	***PCA3***	***0.19***	***APOC1***	***0.19***
APOC1	***0.16***	***AMACR***	***0.15***	***AMACR***	***0.16***
AMACR	***0.14***	***HOXC6***	***0.10***	***TMPRSS2:ERG***	***0.14***
HOXC6	***0.09***	***TMPRSS2:ERG***	***0.10***	***HOXC6***	***0.12***
TMPRSS2:ERG	***0.08***	***ANKRD34B***	***0.07***	***ANKRD34B***	***0.05***
TMEM47	***0.07***	***HPN***	***0.06***	***DLX1***	***0.03***
MMP25	***0.05***	***NEAT1***	***0.04***	***PPFIA2***	***-0.01***
DLX1	***0.03***	***DLX1***	***0.03***		
NEAT1	***0.03***	***AURKA***	***-0.02***		
ISX	***0.01***	***PTPRC***	***-0.03***		
MAK	***0.00***	***ALAS1***	***-0.05***		
MED4	***-0.02***	***PSTPIP1***	***-0.06***		
CP	***-0.02***	***ACTR5***	***-0.14***		
CKAP2L	***-0.02***	***RPL18A***	***-0.22***		
IGFBP3	***-0.02***				
AR exon 9	***-0.03***				
SRSF3	***-0.04***				
PDLIM5	***-0.07***				
BTG2	***-0.07***				
STEAP4	***-0.08***				
CD10	***-0.14***				
AR exons 4-8	***-0.17***				
KLK4	***-0.27***				
DPP4	***-0.29***				



Supplementary Figure 10 KLK2 Adjusted Data ROC curves for test data using models detecting between CB and high risk Ca for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 15 Lasso output for models detecting between CB and high risk Ca using KLK2 adjusted data.

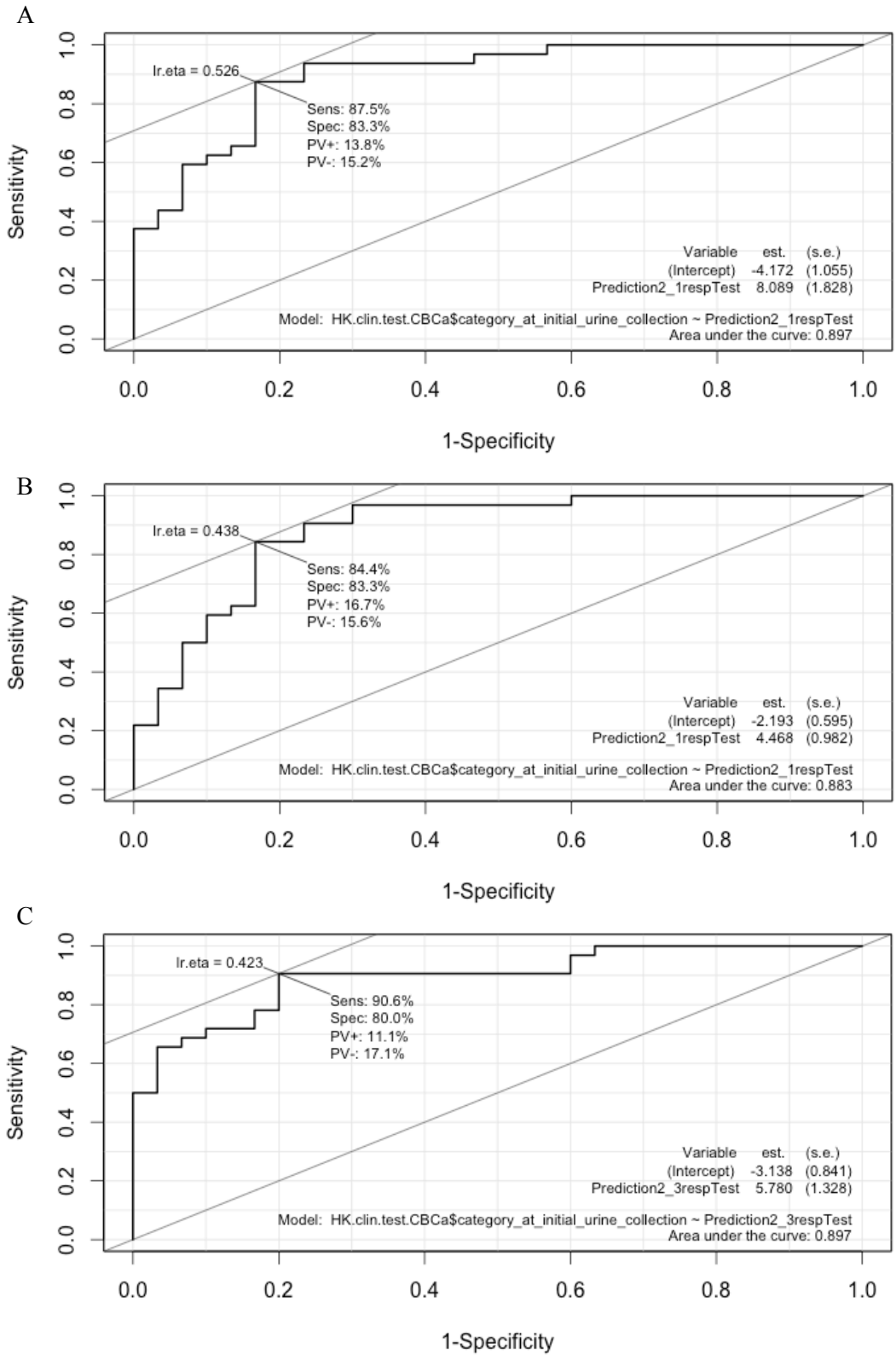
All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***3.08***	***PCA3***	***3.34***	***PCA3***	***5.34***
HPN	***2.79***	***HPN***	***2.40***	***HPN***	***5.02***
AMACR	***1.09***	***AMACR***	***1.93***	***GJB1***	***2.32***
ERG3' exons 6-7	***0.83***	***SIM2.short***	***1.45***	***AMACR***	***1.85***
SIM2.short	***0.72***	***DNAH5***	***1.03***	***KLK4***	***-2.59***
RAB17	***0.37***	***ERG3' exons 6-7***	***0.94***		
APOC1	***0.34***	***RAB17***	***0.50***		
MMP25	***0.34***	***ANKRD34B***	***0.45***		
ANKRD34B	***0.27***	***APOC1***	***0.44***		
DLX1	***0.25***	***DLX1***	***0.43***		
CLU	***0.23***	***SLC12A1***	***0.38***		
DNAH5	***0.16***	***STOM***	***0.16***		
SLC12A1	***0.14***	***ERG3' exons 4-5***	***0.09***		
ERG3' exons 4-5	***0.03***	***KLK4***	***-0.46***		
STOM	***0.03***	***MYOF***	***-0.71***		
MYOF	***-0.29***	***DPP4***	***-1.37***		
KLK4	***-0.40***	***CD10***	***-2.13***		
DPP4	***-0.98***				
CD10	***-1.43***				



Supplementary Figure 11 KLK3 Adjusted Data ROC curves for test data using models detecting between CB and high risk Ca for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 16 Lasso output for models detecting between CB and high risk Ca using KLK3 adjusted data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***2.56***	***PCA3***	***2.96***	***HPN***	***5.06***
HPN	***2.37***	***HPN***	***2.13***	***PCA3***	***4.95***
ERG3' exons 4-5	***0.77***	***SIM2.short***	***1.46***	***KLK4***	***-3.20***
SIM2.short	***0.64***	***ERG3' exons 4-5***	***1.17***		
APOC1	***0.44***	***AMACR***	***1.10***		
MMP25	***0.41***	***APOC1***	***0.64***		
AMACR	***0.30***	***ANKRD34B***	***0.57***		
ANKRD34B	***0.29***	***SLC12A1***	***0.43***		
SLC12A1	***0.18***	***SULT1A1***	***0.41***		
ERG3' exons 6-7	***0.17***	***DLX1***	***0.30***		
DLX1	***0.13***	***RAB17***	***0.27***		
RAB17	***0.11***	***STOM***	***0.17***		
SULT1A1	***0.07***	***ERG3' exons 6-7***	***0.16***		
STOM	***0.06***	***PTN***	***-0.14***		
PTN	***-0.01***	***RP11.244H18.1.P712P***	***-0.37***		
KLK4	***-0.18***	***MARCH5***	***-0.72***		
MARCH5	***-0.23***	***MYOF***	***-1.00***		
RP11.244H18.1.P712P	***-0.39***	***DPP4***	***-1.25***		
MYOF	***-0.65***	***CD10***	***-1.68***		
DPP4	***-0.82***				
CD10	***-1.17***				



Supplementary Figure 12 HK Normalised Data ROC curves for test data using models detecting between CB and high risk Ca for models using the following inputs A) all probes, B) significant probes, C) adjusted significant probes.

Supplementary Table 17 Lasso output for models detecting between CB and high risk Ca using HK normalised data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
ERG3' exons 4-5	***0.35***	***ERG3' exons 4-5***	***0.67***	***PCA3***	***0.50***
PCA3	***0.27***	***ANKRD34B***	***0.34***	***APOC1***	***0.39***
APOC1	***0.18***	***PCA3***	***0.33***	***HPN***	***0.37***
HPN	***0.13***	***APOC1***	***0.24***	***ERG3' exons 4-5***	***0.23***
SLC12A1	***0.03***	***AMACR***	***0.19***	***TMPRSS2:ERG***	***0.20***
TMPRSS2:ERG	***0.02***	***HPN***	***0.12***		
ANKRD34B	***0.02***	***SULT1A1***	***0.08***		
HOXC6	***0.01***	***NEAT1***	***0.08***		
AR exons 4-8	***-0.05***	***TMPRSS2:ERG***	***0.07***		
GABARAPL2	***-0.06***	***DLX1***	***0.04***		
CD10	***-0.14***	***HOXC6***	***0.03***		
KLK4	***-0.15***	***STOM***	***0.02***		
DPP4	***-0.18***	***SLC12A1***	***0.01***		
		AR exon 9	***-0.03***		
		MYOF	***-0.06***		
		SRSF3	***-0.07***		
		AR exons 4-8	***-0.20***		
		CD10	***-0.26***		
		GABARAPL2	***-0.28***		
		DPP4	***-0.36***		
		PDLIM5	***-0.56***		

6.15 Multinomial CBLIH Trend

Supplementary Table 18 Glm test significant probes for CB, L, I, H trend

KLK2 ratio data			*KLK2* adjusted data		
Transcript	***p*-value**	**Adjusted *p*-value**	**Transcript**	***p*-value**	**Adjusted *p*-value**
ERG3' exons 4-5	1.86x10-13	3.09x10-11	*PCA3*	1.45x10-08	2.41x10-06
PCA3	8.90x10-13	1.47x10-10	*ERG3' exons 4-5*	1.18x10-07	1.94x10-05
TMPRSS2:ERG	1.88x10-11	3.09x10-09	*ERG3' exons 6-7*	3.73x10-07	6.11x10-05
ERG3' exons 6-7	6.66x10-10	1.09x10-07	*SPINK1*	1.03x10-06	0.0002
HOXC6	7.88x10-09	1.28x10-06	*HOXC6*	3.85x10-06	0.0006
HPN	4.19x10-08	6.75x10-06	*HPN*	7.34x10-06	0.0012
APOC1	6.38x10-08	1.02x10-05	*TMPRSS2:ERG*	7.35x10-06	0.0012
TDRD	1.63x10-07	2.59x10-05	*KLK4*	3.65x10-05	0.0058
ANKRD34B	1.47x10-06	0.0002	*SLC12A1*	4.71x10-05	0.0074
ITGBL1	3.19x10-06	0.001	*UPK2*	8.47x10-05	0.0133
SLC12A1	5.82x10-06	0.001	*TDRD*	0.0002	0.0242
DLX1	7.26x10-06	0.001	*ITGBL1*	0.0002	0.0281
RAB17	9.26x10-06	0.001	*RP11.244H18.1.P712P*	0.0002	0.0320
HOXC4	1.07x10-05	0.002	*GABARAPL2*	0.0002	0.0363
GJB1	1.49x10-05	0.002	*GJB1*	0.0002	0.0366
PPFIA2	1.84x10-05	0.003	*AMACR*	0.0005	0.0695
SPINK1	2.62x10-05	0.004	*MYOF*	0.0005	0.0775
AMACR	3.58x10-05	0.005	*APOC1*	0.0005	0.0816
AMH	4.60x10-05	0.007	*MED4*	0.0010	0.1423
TRPM4	5.74x10-05	0.008	*SULT1A1*	0.0011	0.1624
NEAT1	6.14x10-05	0.009	*RAB17*	0.0020	0.2865
SIM2.short	6.84x10-05	0.010	*ANKRD34B*	0.0025	0.3583
SSTR1	7.31x10-05	0.011	*SNCA*	0.0035	0.4989
UPK2	7.83x10-05	0.011	*MMP26*	0.0047	0.6782
SULT1A1	8.22x10-05	0.012	*PTN*	0.0055	0.7804
MEX3A	9.10x10-05	0.013	*DLX1*	0.0055	0.7804

MIR146A.DQ658414	0.0001	0.015	*IFT57*	0.0058	0.8138
TMEM45B	0.0001	0.015	*SIM2.short*	0.0061	0.8487
ISX	0.0001	0.017	*DPP4*	0.0073	0.9916
MIC1	0.0001	0.019	*STOM*	0.0080	0.9916
TWIST1	0.0002	0.021	*GAPDH*	0.0105	0.9916
Met	0.0002	0.021	*VPS13A*	0.0135	0.9916
MMP11	0.0002	0.023	*MIR146A.DQ658414*	0.0168	0.9916
CDKN3	0.0002	0.023	*PPAP2A*	0.0182	0.9916
RP11.97O12.7	0.0002	0.023	*ZNF577*	0.0185	0.9916
STOM	0.0003	0.035	*SMIM1*	0.0233	0.9916
PALM3	0.0003	0.043	*PPFIA2*	0.0249	0.9916
LASS1	0.0003	0.043	*Met*	0.0251	0.9916
SSPO	0.0003	0.044	*MIC1*	0.0268	0.9916
MMP26	0.000	0.049	*EIF2D*	0.0316	0.9916
VPS13A	0.000	0.049	*CD10*	0.0336	0.9916
PECI	0.000	0.050	*STEAP2*	0.0432	0.9916
PCSK6	0.000	0.054	*MIR4435.1HG.lOC541471*	0.0437	0.9916
GAPDH	0.000	0.056	*ITPR1*	0.0445	0.9916
PVT1	0.000	0.056	*MXI1*	0.0487	0.9916
TERT	0.000	0.060			
CASKIN1	0.001	0.061			
TMCC2	0.001	0.064			
RPLP2	0.001	0.080			
MNX1	0.001	0.102			
SIM2.long	0.001	0.106			
RPS11	0.001	0.106			
SULF2	0.001	0.130			
HIST1H1C	0.001	0.133			
EN2	0.001	0.133			
DNAH5	0.001	0.166			
MMP25	0.002	0.198			
MFSD2A	0.002	0.212			
MIR4435.1HG.lOC541471	0.002	0.226			
SMIM1	0.002	0.239			

MGAT5B	0.003	0.266
RIOK3	0.003	0.267
MCTP1	0.003	0.315
RPS10	0.003	0.329
VAX2	0.003	0.337
TMEM86A	0.003	0.340
ERG5	0.004	0.358
IMPDH2	0.004	0.368
COL10A1	0.004	0.400
ABCB9	0.004	0.424
B4GALNT4	0.005	0.471
MKi67	0.005	0.472
CLIC2	0.006	0.526
SChLAP1	0.007	0.671
CCDC88B	0.009	0.807
PTPRC	0.009	0.809
CAMKK2	0.009	0.838
NAALADL2	0.009	0.844
HIST3H2A	0.010	0.873
HPRT	0.010	0.897
TERF2IP	0.011	0.949
ITPR1	0.014	0.994
SLC4A1.S	0.014	0.994
COL9A2	0.014	0.994
MCM7	0.015	0.994
CKAP2L	0.017	0.994
RPL18A	0.017	0.994
BRAF	0.017	0.994
MAPK8IP2	0.017	0.994
SFRP4	0.018	0.994
FDPS	0.018	0.994
SACM1L	0.019	0.994
MSMB	0.020	0.994
HMBS	0.020	0.994

SPON2	0.021	0.994
ANPEP	0.021	0.994
CACNA1D	0.022	0.994
SYNM	0.023	0.994
ALAS1	0.026	0.994
RNF157	0.027	0.994
HIST1H1E	0.027	0.994
ARHGEF25	0.028	0.994
RPL23AP53	0.028	0.994
AURKA	0.031	0.994
PSTPIP1	0.032	0.994
FOLH1	0.032	0.994
GOLM1	0.033	0.994
EIF2D	0.035	0.994
IFT57	0.039	0.994
SLC43A1	0.039	0.994
CDC20	0.039	0.994
CAMK2N2	0.047	0.994
GABARAPL2	0.049	0.994
CDC37L1	0.050	0.994

KLK3 adjusted data			*GAPDH* and *RPLP2* normalised data		
Transcript	***p*-value**	**Adjusted *p*-value**	**Transcript**	***p*-value**	**Adjusted *p*-value**
PCA3	1.52x10-07	2.52x10-05	*ERG3' exons 4-5*	1.44x10-08	2.41x10-06
SPINK1	5.80x10-06	0.001	*TMPRSS2:ERG*	1.18x10-07	1.96x10-05
ERG3' exons 4-5	6.32x10-06	0.001	*PCA3*	2.06x10-07	3.39x10-05
ERG3' exons 6-7	7.48x10-06	0.001	*ERG3' exons 6-7*	2.28x10-06	0.0004
KLK4	8.86x10-06	0.001	*APOC1*	9.64x10-06	0.002
SLC12A1	4.36x10-05	0.007	*HOXC6*	1.34x10-05	0.002
HOXC6	4.72x10-05	0.008	*HPN*	2.01x10-05	0.003
UPK2	5.48x10-05	0.009	*DPP4*	9.43x10-05	0.015
HPN	7.32x10-05	0.012	*GABARAPL2*	0.0001	0.017
TMPRSS2:ERG	0.0001	0.017	*ITGBL1*	0.0001	0.017
SULT1A1	0.0002	0.036	*MYOF*	0.0001	0.018

APOC1	0.0004	0.056	*KLK2*	0.0004	0.065
GJB1	0.0004	0.064	*SLC12A1*	0.0004	0.069
MYOF	0.0005	0.074	*TDRD*	0.0004	0.069
CD10	0.001	0.081	*SRSF3*	0.001	0.088
ITGBL1	0.001	0.138	*SPINK1*	0.001	0.097
RP11.244H18.1.P712P	0.001	0.138	*P712P*	0.001	0.098
DLX1	0.001	0.174	*KLK4*	0.001	0.110
RAB17	0.001	0.179	*RAB17*	0.001	0.156
GABARAPL2	0.001	0.189	*AR exons 4-8*	0.001	0.191
STOM	0.001	0.215	*IFT57*	0.001	0.191
TDRD	0.002	0.266	*CD10*	0.002	0.276
PTN	0.002	0.293	*PTN*	0.003	0.375
AMACR	0.002	0.327	*DLX1*	0.003	0.418
MED4	0.003	0.378	*ANKRD34B*	0.003	0.419
SNCA	0.004	0.622	*ZNF577*	0.003	0.434
NEAT1	0.005	0.743	*UPK2*	0.003	0.443
ANKRD34B	0.005	0.743	*MXI1*	0.004	0.606
MIR4435.1HG.lOC541471	0.006	0.783	*HOXC4*	0.006	0.767
KLK2	0.007	0.928	*SNCA*	0.006	0.769
Met	0.012	0.980	*STEAP2*	0.006	0.800
AURKA	0.014	0.980	*MEMO1*	0.006	0.834
SIM2.short	0.014	0.980	*CACNA1D*	0.006	0.834
MIC1	0.019	0.980	*STEAP4*	0.007	0.954
PPFIA2	0.020	0.980	*PPAP2A*	0.008	0.997
MEMO1	0.021	0.980	*Met*	0.011	0.997
ZNF577	0.023	0.980	*MED4*	0.011	0.997
CACNA1D	0.025	0.980	*MIATNB*	0.011	0.997
AR exon 9	0.026	0.980	*GJB1*	0.013	0.997
PDLIM5	0.027	0.980	*AR exon 9*	0.014	0.997
RP11.97O12.7	0.033	0.980	*SULT1A1*	0.018	0.997
IFT57	0.033	0.980	*PPFIA2*	0.018	0.997
MMP26	0.033	0.980	*FDPS*	0.019	0.997
MARCH5	0.034	0.980	*MARCH5*	0.019	0.997
RPS10	0.036	0.980	*MSMB*	0.020	0.997

AR exons 4-8	0.036	0.980	*KLK3 exons 2-3*	0.024	0.997
ITPR1	0.039	0.980	*SNORA20*	0.026	0.997
SMIM1	0.043	0.980	*NEAT1*	0.027	0.997
SNORA20	0.044	0.980	*RPS10*	0.028	0.997
VPS13A	0.046	0.980	*SERPINB5*	0.035	0.997
			TRPM4	0.038	0.997
			NLRP3	0.040	0.997
			HIST1H2BF	0.049	0.997

Supplementary Table 19 Lasso output for models detecting CB, L, I, H trend using KLK2 ratio data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***0.21***	***PCA3***	***0.18***	***PCA3***	***0.14***
ERG3' exons 4-5	***0.11***	***ERG3' exons 4-5***	***0.12***	***ERG3' exons 4-5***	***0.11***
APOC1	***0.11***	***APOC1***	***0.08***	***APOC1***	***0.05***
ANKRD34B	***0.07***	***TMPRSS2:ERG***	***0.04***	***TMPRSS2:ERG***	***0.04***
NEAT1	***0.05***	***SLC12A1***	***0.03***	***HOXC6***	***0.01***
HOXC6	***0.05***	***HOXC6***	***0.02***	***cp1***	***2.05***
HPN	***0.05***	***NEAT1***	***0.02***	***cp2***	***1.26***
TMPRSS2:ERG	***0.04***	***HPN***	***0.01***	***cp3***	***-0.36***
ITGBL1	***0.03***	***ANKRD34B***	***0.01***		
SLC12A1	***0.03***	***DLX1***	***0.00***		
SULT1A1	***0.03***	***PSTPIP1***	***0.00***		
ISX	***0.03***	***HIST1H1E***	***-0.05***		
DLX1	***0.02***	***GABARAPL2***	***-0.16***		
ERG3' exons 6-7	***0.01***	***cp1***	***2.19***		
TMEM47	***0.01***	***cp2***	***1.32***		
TDRD	***0.01***	***cp3***	***-0.38***		
AMACR	***0.01***				
HIST1H1E	***-0.01***				
IGFBP3	***-0.01***				
PSGR	***-0.01***				
BTG2	***-0.01***				
MED4	***-0.02***				
AR exons 4-8	***-0.02***				
PPP1R12B	***-0.02***				
AR exon 9	***-0.02***				
Timp4	***-0.03***				
DPP4	***-0.03***				
CP	***-0.04***				
MYOF	***-0.04***				
GCNT1	***-0.04***				
MEMO1	***-0.05***				
SRSF3	***-0.06***				
ZNF577	***-0.06***				
CD10	***-0.06***				
MXI1	***-0.10***				
KLK4	***-0.14***				
cp1	***2.45***				
cp2	***1.43***				
cp3	***-0.43***				

Supplementary Table 20 Lasso output for models detecting CB, L, I, H trend using KLK2 adjusted data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
AMACR	***0.14***	***AMACR***	***0.46***	***ERG3' exons 4-5***	***0.75***
ERG3' exons 4-5	***0.74***	***ANKRD34B***	***0.33***	***GABARAPL2***	***-1.05***
GJB1	***0.60***	***APOC1***	***0.53***	***GJB1***	***0.63***
HOXC6	***0.36***	***CD10***	***-1.04***	***HOXC6***	***0.38***
HPN	***0.74***	***DLX1***	***0.13***	***HPN***	***0.70***
ITGBL1	***0.12***	***DPP4***	***-0.14***	***ITGBL1***	***0.18***
KLK4	***-1.22***	***ERG3' exons 4-5***	***0.80***	***KLK4***	***-1.10***
PCA3	***2.38***	***GABARAPL2***	***-0.88***	***PCA3***	***2.27***
SLC12A1	***0.25***	***GAPDH***	***-0.07***	***RP11.244H18.1.P712P***	***-1.34***
SPINK1	***0.44***	***GJB1***	***0.08***	***SLC12A1***	***0.26***
TMPRSS2:ERG	***0.36***	***HOXC6***	***0.22***	***SPINK1***	***0.38***
UPK2	***0.24***	***IFT57***	***0.95***	***TDRD***	***0.14***
cp1	***2.06***	***ITPR1***	***0.13***	***TMPRSS2:ERG***	***0.38***
cp2	***1.35***	***KLK4***	***-0.60***	***UPK2***	***0.19***
cp3	***-0.42***	***MED4***	***-0.85***	***cp1***	***2.21***
		Met	***-0.11***	***cp2***	***1.41***
		MIC1	***0.28***	***cp3***	***-0.43***
		MIR146A.DQ658414	***0.24***		
		MMP26	***0.50***		
		MXI1	***-1.15***		
		MYOF	***-1.45***		
		PCA3	***2.69***		
		PPAP2A	***0.08***		
		PPFIA2	***-0.65***		
		PTN	***-0.79***		
		RP11.244H18.1.P712P	***-0.72***		
		SIM2.short	***1.12***		
		SLC12A1	***0.14***		
		SMIM1	***0.40***		

SNCA	***-0.77***
SPINK1	***0.47***
STEAP2	***0.82***
STOM	***0.11***
SULT1A1	***0.85***
TMPRSS2:ERG	***0.13***
UPK2	***0.25***
ZNF577	***-0.49***
cp1	***2.47***
cp2	***1.52***
cp3	***-0.48***

Supplementary Table 21 Lasso output for models detecting CB, L, I, H trend using KLK3 adjusted data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
ACTR5	***-0.40***	***MARCH5***	***-0.88***	***APOC1***	***0.78***
AMH	***1.60***	***AMACR***	***0.27***	***ERG3' exons 4-5***	***0.56***
ANKRD34B	***0.20***	***ANKRD34B***	***0.21***	***ERG3' exons 6-7***	***0.07***
APOC1	***0.96***	***APOC1***	***0.60***	***GJB1***	***0.51***
AR exon 9	***-0.05***	***AR exon 9***	***-0.49***	***HOXC6***	***0.36***
AURKA	***0.07***	***AURKA***	***0.70***	***HPN***	***0.52***
B2M	***-0.23***	***CACNA1D***	***0.42***	***KLK4***	***-1.25***
BRAF	***0.48***	***CD10***	***-0.75***	***PCA3***	***2.21***
BTG2	***-1.11***	***DLX1***	***0.27***	***SLC12A1***	***0.19***
CASKIN1	***-0.01***	***ERG3' exons 4-5***	***0.64***	***SULT1A1***	***0.65***
CCDC88B	***-0.27***	***GABARAPL2***	***-1.14***	***TMPRSS2:ERG***	***0.25***
CD10	***-0.95***	***GJB1***	***0.28***	***UPK2***	***0.50***
CDC20	***-0.44***	***HOXC6***	***0.45***	***cp1***	***1.14***
CKAP2L	***-0.39***	***ITGBL1***	***-0.31***	***cp2***	***1.35***
CLIC2	***-0.52***	***ITPR1***	***0.68***	***cp3***	***-0.36***
CLU	***0.11***	***KLK2***	***0.42***		
CP	***-0.46***	***KLK4***	***-0.53***		
CTA.211A9.5.MIATNB	***-0.41***	***MED4***	***-0.86***		
DLX1	***0.33***	***MEMO1***	***-0.90***		
DNAH5	***0.23***	***MIC1***	***0.40***		
ERG3' exons 4-5	***0.54***	***MIR4435.1HG.lOC541471***	***-0.35***		
ERG3' exons 6-7	***0.36***	***MMP26***	***0.24***		
GABARAPL2	***-1.42***	***MYOF***	***-1.84***		
GOLM1	***-0.02***	***NEAT1***	***0.80***		
HIST3H2A	***-0.34***	***PCA3***	***3.73***		
HOXC4	***-0.97***	***PPFIA2***	***-0.97***		
HOXC6	***0.76***	***PTN***	***-1.10***		
HPRT	***0.81***	***RAB17***	***-0.89***		
IGFBP3	***-0.87***	***RP11.244H18.1.P712P***	***-0.34***		
IMPDH2	***0.03***	***RP11.97O12.7***	***0.48***		
ITPR1	***0.17***	***SIM2.short***	***1.35***		
KLK4	***-1.18***	***SLC12A1***	***0.39***		
LASS1	***-0.16***	***SMIM1***	***0.47***		
LBH	***0.51***	***SNCA***	***-0.20***		
MCM7	***0.27***	***STOM***	***0.31***		
MDK	***0.01***	***SULT1A1***	***1.05***		
MED4	***-0.38***	***TMPRSS2:ERG***	***-0.02***		

MEMO1	***-0.97***	***UPK2***	***0.94***
MGAT5B	***-0.24***	***ZNF577***	***-0.14***
MIC1	***0.64***	***cp1***	***1.77***
MIR146A.DQ658414	***0.20***	***cp2***	***1.63***
MIR4435.1HG.lOC5414 71	***0.02***	***cp3***	***-0.46***
MMP11	***0.61***		
MMP25	***0.83***		
MNX1	***0.42***		
MXI1	***-0.18***		
MYOF	***-1.42***		
NAALADL2	***-0.32***		
NEAT1	***0.98***		
PSGR	***-0.15***		
PALM3	***0.21***		
PCA3	***3.16***		
PPAP2A	***0.89***		
PSTPIP1	***-0.62***		
PTN	***-0.71***		
PVT1	***0.17***		
RPL23AP53	***0.06***		
RPS11	***-0.02***		
SACM1L	***-0.41***		
SERPINB5	***0.28***		
SIM2.short	***0.90***		
SIRT1	***-0.68***		
SLC12A1	***0.23***		
SMIM1	***0.20***		
SNCA	***-0.22***		
SPINK1	***0.49***		
SPON2	***-0.43***		
SRSF3	***-0.01***		
ST6GALNAC1	***0.25***		
STOM	***0.36***		
SULT1A1	***0.61***		
SYNM	***0.19***		
TDRD	***0.03***		
Timp4	***-0.91***		
TWIST1	***0.65***		
UPK2	***0.75***		
VAX2	***-0.27***		
ZNF577	***-0.02***		
cp1	***2.04***		
cp2	***1.75***		
cp3	***-0.50***		

Supplementary Table 22 Lasso output for models detecting CB, L, I, H trend using HK normalised data.

All Transcripts	*Significant Transcripts*	*Multiple testing*

				corrected Transcripts	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
ACTR5	***0.00***	***ANKRD34B***	***0.11***	***APOC1***	***0.14***
AMH	***0.04***	***APOC1***	***0.12***	***DPP4***	***-0.25***
ANKRD34B	***0.07***	***AR exon 9***	***-0.04***	***ERG3' exons 4-5***	***0.11***
APOC1	***0.11***	***AR exons 4-8***	***-0.04***	***ERG3' exons 6-7***	***0.01***
AR exon 9	***-0.02***	***CD10***	***-0.08***	***GABARAPL2***	***-0.26***
AR exons 4-8	***-0.02***	***MIATNB***	***0.00***	***HOXC6***	***0.08***
CD10	***-0.05***	***DLX1***	***0.03***	***HPN***	***0.10***
CP	***-0.05***	***DPP4***	***-0.07***	***ITGBL1***	***0.10***
DLX1	***0.01***	***ERG3' exons 4-5***	***0.10***	***KLK4***	***-0.30***
DPP4	***-0.06***	***ERG3' exons 6-7***	***0.01***	***MYOF***	***-0.17***
ERG3' exons 4-5	***0.10***	***FDPS***	***-0.03***	***PCA3***	***0.23***
ERG3' exons 6-7	***0.01***	***GABARAPL2***	***-0.03***	***TDRD***	***0.04***
GABARAPL2	***-0.08***	***GJB1***	***0.02***	***TMPRSS2:ERG***	***0.04***
GCNT1	***-0.03***	***HOXC6***	***0.06***	***cp1***	***2.63***
HIST1H1E	***-0.03***	***HPN***	***0.08***	***cp2***	***1.52***
HIST1H2BF	***0.00***	***ITGBL1***	***0.07***	***cp3***	***-0.46***
HOXC6	***0.05***	***KLK4***	***-0.21***		
HPN	***0.05***	***MED4***	***-0.13***		
IGFBP3	***-0.01***	***MEMO1***	***-0.09***		
ISX	***0.02***	***MSMB***	***0.10***		
ITGBL1	***0.07***	***MXI1***	***-0.24***		
KLK4	***-0.16***	***MYOF***	***-0.06***		
MED4	***-0.02***	***NEAT1***	***0.06***		
MEMO1	***-0.05***	***PCA3***	***0.21***		
MXI1	***-0.13***	***RPS10***	***-0.02***		
MYOF	***-0.05***	***SLC12A1***	***0.04***		
NEAT1	***0.04***	***SPINK1***	***0.00***		
PCA3	***0.19***	***SRSF3***	***-0.09***		
PPP1R12B	***-0.02***	***SULT1A1***	***0.06***		
SLC12A1	***0.03***	***TDRD***	***0.03***		
SRSF3	***-0.05***	***TMPRSS2:ERG***	***0.03***		
SULT1A1	***0.03***	***TRPM4***	***-0.02***		

TDRD	***0.01***	***UPK2***	***0.02***
Timp4	***-0.02***	***ZNF577***	***-0.07***
TMPRSS2:ERG	***0.04***	***cp1***	***2.67***
UPK2	***0.00***	***cp2***	***1.55***
ZNF577	***-0.04***	***cp3***	***-0.47***
cp1	***2.42***		
cp2	***1.41***		
cp3	***-0.42***		

6.16 Multinomial CBCaM Trend

Supplementary Table 23 Glm test significant probes for CB, Ca, Mets trend

KLK2 ratio data			*KLK2* adjusted data		
Transcript	***p*-value**	**Adjusted *p*-value**	**Transcript**	***p*-value**	**Adjusted *p*-value**
HOXC6	5.20X10-10	8.62X10-08	*UPK2*	2.91X10-08	4.83X10-06
ERG3' exons 4-5	5.70X10-10	9.41X10-08	*SPINK1*	2.04X10-07	3.36X10-05
PCA3	4.58X10-09	7.51X10-07	*SLC12A1*	1.61X10-06	0.0003
TMPRSS2:ERG	1.27X10-08	2.07X10-06	*HOXC6*	5.86X10-05	0.0096
APOC1	1.42X10-08	2.30X10-06	*HPN*	7.26X10-05	0.0118
TDRD	2.07X10-08	3.34X10-06	*MFSD2A*	7.70X10-05	0.0124
SLC12A1	2.82X10-08	4.51X10-06	*GAPDH*	0.0001	0.0196
HPN	3.41X10-08	5.42X10-06	*RAB17*	0.0002	0.0285
HOXC4	2.04X10-07	3.23X10-05	*KLK4*	0.0002	0.0313
RAB17	2.07X10-07	3.26X10-05	*PCA3*	0.0005	0.0758
GJB1	2.14X10-07	3.34X10-05	*GJB1*	0.0005	0.0826
ERG3' exons 6-7	2.32X10-07	3.59X10-05	*MIR4435.1HG.lOC541471*	0.0007	0.1054
AMACR	4.06X10-07	6.25X10-05	*GABARAPL2*	0.0007	0.1085
SPINK1	4.92X10-07	7.53X10-05	*TMEM45B*	0.0009	0.1315
SSTR1	5.08X10-07	7.71X10-05	*APOC1*	0.0009	0.1376
UPK2	5.69X10-07	8.60X10-05	*AURKA*	0.0012	0.1751
TMCC2	6.63X10-07	9.94X10-05	*ANPEP*	0.0012	0.1760
TMEM45B	6.75X10-07	0.0001	*SULT1A1*	0.0017	0.2479
PPFIA2	7.28X10-07	0.0001	*RP11.244H18.1.P712P*	0.0020	0.2933
DLX1	8.97X10-07	0.0001	*ERG3' exons 4-5*	0.0023	0.3385
PALM3	9.33X10-07	0.0001	*PALM3*	0.0023	0.3421
SULT1A1	1.03X10-06	0.0001	*TMPRSS2:ERG*	0.0024	0.3487
RP11.97O12.7	1.60X10-06	0.0002	*TDRD*	0.0028	0.4042
SULF2	2.21X10-06	0.0003	*ERG3' exons 6-7*	0.0029	0.4165
AMH	2.21X10-06	0.0003	*TBP*	0.0031	0.4376
EN2	2.70X10-06	0.0004	*HMBS*	0.0035	0.4878

CASKIN1	2.79X10-06	0.0004	*ITGBL1*	0.0041	0.5687
MIR4435.1HG.lOC541471	2.91X10-06	0.0004	*AMACR*	0.0042	0.5824
HIST1H1C	3.23X10-06	0.0004	*TMCC2*	0.0044	0.6065
MEX3A	3.24X10-06	0.0004	*MYOF*	0.0058	0.7895
PECI	3.81X10-06	0.0005	*RNF157*	0.0073	0.9921
SIM2.short	3.92X10-06	0.0005	*PTPRC*	0.0079	0.9921
ISX	4.01X10-06	0.0005	*SIM2.short*	0.0081	0.9921
TMEM86A	4.87X10-06	0.0006	*SULF2*	0.0088	0.9921
ERG5	5.11X10-06	0.0007	*EN2*	0.0094	0.9921
TWIST1	5.24X10-06	0.0007	*PTN*	0.0095	0.9921
ITGBL1	6.47X10-06	0.0008	*ALAS1*	0.0104	0.9921
MGAT5B	6.56X10-06	0.0008	*TMEM86A*	0.0115	0.9921
MMP11	7.01X10-06	0.0009	*RP11.97O12.7*	0.0117	0.9921
HMBS	7.33X10-06	0.0009	*PPFIA2*	0.0122	0.9921
MCTP1	8.97X10-06	0.0011	*DPP4*	0.0125	0.9921
GAPDH	1.06X10-05	0.0013	*STOM*	0.0132	0.9921
STOM	1.09X10-05	0.0014	*Met*	0.0139	0.9921
HIST3H2A	1.26X10-05	0.0016	*ZNF577*	0.0153	0.9921
RPL23AP53	1.29X10-05	0.0016	*ERG5*	0.0251	0.9921
MFSD2A	1.49X10-05	0.0018	*ITPR1*	0.0280	0.9921
TERT	1.72X10-05	0.0021	*MARCH5*	0.0299	0.9921
Met	2.06X10-05	0.0025	*HIST1H1E*	0.0342	0.9921
B4GALNT4	2.08X10-05	0.0025	*SMIM1*	0.0380	0.9921
NLRP3	2.10X10-05	0.0025	*DLX1*	0.0384	0.9921
PVT1	2.14X10-05	0.0025	*RPL23AP53*	0.0433	0.9921
MIR146A.DQ658414	2.34X10-05	0.0027	*CASKIN1*	0.0457	0.9921
CCDC88B	2.70X10-05	0.0031	*SEC61A1*	0.0474	0.9921
PPAP2A	2.71X10-05	0.0031	*AMH*	0.0478	0.9921
ITPR1	3.03X10-05	0.0034	*IFT57*	0.0481	0.9921
ABCB9	3.22X10-05	0.0035	*CLU*	0.0497	0.9921
ANPEP	3.25X10-05	0.0035	*VPS13A*	0.0499	0.9921
VPS13A	3.25X10-05	0.0035			
MMP25	3.30X10-05	0.0036			
PSTPIP1	3.39X10-05	0.0036			

AURKA	3.44X10-05	0.0036
VAX2	3.72X10-05	0.0039
TRPM4	3.81X10-05	0.0039
PTPRC	3.82X10-05	0.0039
RIOK3	3.85X10-05	0.0039
OGT	4.06X10-05	0.0041
MNX1	4.10X10-05	0.0041
SLC4A1.S	4.52X10-05	0.0045
HPRT	4.84X10-05	0.0047
TBP	4.91X10-05	0.0048
HIST1H1E	5.41X10-05	0.0052
NAALADL2	5.44X10-05	0.0052
SIM2.long	5.69X10-05	0.0053
CLIC2	6.01X10-05	0.0056
DNAH5	6.67X10-05	0.0061
SMIM1	7.71X10-05	0.0070
PCSK6	8.13X10-05	0.0073
MKi67	8.24X10-05	0.0073
COL9A2	8.76X10-05	0.0077
BRAF	8.85X10-05	0.0077
COL10A1	9.42X10-05	0.0081
TERF2IP	9.55X10-05	0.0081
SSPO	0.0001	0.0091
RPLP2	0.0001	0.0097
SFRP4	0.0001	0.0097
MAPK8IP2	0.0001	0.0097
CDC37L1	0.0001	0.0097
RNF157	0.0001	0.0101
ACTR5	0.0001	0.0103
RPS11	0.0001	0.0110
RPS10	0.0001	0.0110
SYNM	0.0001	0.0111
CDKN3	0.0002	0.0120
AATF	0.0002	0.0127

EIF2D	0.0002	0.0132
ALAS1	0.0002	0.0139
IMPDH2	0.0002	0.0145
FDPS	0.0002	0.0157
SACM1L	0.0002	0.0165
TFDP1	0.0003	0.0170
MCM7	0.0003	0.0202
NEAT1	0.0003	0.0208
CAMKK2	0.0004	0.0232
IFT57	0.0004	0.0242
MEMO1	0.0005	0.0299
ANKRD34B	0.0005	0.0309
RPL18A	0.0005	0.0310
PPP1R12B	0.0006	0.0364
CACNA1D	0.0006	0.0370
MIC1	0.0008	0.0436
SPON2	0.0008	0.0441
CLU	0.0008	0.0441
BTG2	0.0010	0.0526
GABARAPL2	0.0014	0.0709
SMAP1 exons 7-8	0.0014	0.0709
STEAP4	0.0014	0.0709
CKAP2L	0.0015	0.0727
KLK3 exons 2-3	0.0015	0.0727
ARHGEF25	0.0017	0.0796
LASS1	0.0017	0.0796
STEAP2	0.0019	0.0887
B2M	0.0021	0.0964
MMP26	0.0023	0.0996
MXI1	0.0024	0.1046
SChLAP1	0.0025	0.1056
CDC20	0.0026	0.1085
CAMK2N2	0.0028	0.1104
SIRT1	0.0032	0.1231

GOLM1	0.0044	0.1676
LBH	0.0062	0.2206
SEC61A1	0.0062	0.2206
AR exons 4-8	0.0064	0.2206
MAK	0.0066	0.2206
PDLIM5	0.0067	0.2206
SRSF3	0.0074	0.2378
SNCA	0.0087	0.2628
FOLH1	0.0089	0.2628
CADPS	0.0091	0.2628
CD10	0.0103	0.2879
MDK	0.0133	0.3598
KLK3 exons 1-2	0.0145	0.3767
MED4	0.0152	0.3792
HIST1H2BG	0.0193	0.4640
PTN	0.0245	0.5637
IGFBP3	0.0268	0.5789
DPP4	0.0276	0.5789
SERPINB5	0.0302	0.6032
ST6GALNAC1	0.0343	0.6512
MARCH5	0.0372	0.6697
HIST1H2BF	0.0405	0.6880
MSMB	0.0471	0.7190
SLC43A1	0.0479	0.7190

KLK3 adjusted data			*GAPDH* and *RPLP2* normalised data		
Transcript	*p*-value	Adjusted *p*-value	Transcript	*p*-value	Adjusted *p*-value
UPK2	2.39x10-08	3.97x10-06	*HOXC6*	3.39X10-06	0.0006
SPINK1	1.87x10-06	0.0003	*SLC12A1*	3.93X10-06	0.0007
SLC12A1	2.58x10-06	0.0004	*APOC1*	7.43X10-06	0.0012
RAB17	4.04x10-06	0.0007	*ERG3' exons 4-5*	2.17X10-05	0.0036
MIR4435.1HG.lOC541471	3.58x10-05	0.0058	*SPINK1*	2.71X10-05	0.0044
HPN	5.22x10-05	0.0084	*KLK2*	3.80X10-05	0.0062
KLK4	7.61x10-05	0.0122	*TMPRSS2:ERG*	5.96X10-05	0.0096

HOXC6	0.0001	0.0168	*HPN*	6.42X10-05	0.0103
GABARAPL2	0.0003	0.0404	*UPK2*	6.50X10-05	0.0103
MFSD2A	0.0005	0.0832	*RAB17*	6.91X10-05	0.0109
SULT1A1	0.0007	0.1103	*TDRD*	0.0003	0.0531
GJB1	0.0007	0.1155	*KLK4*	0.0005	0.0831
APOC1	0.0010	0.1464	*ERG3' exons 6-7*	0.0005	0.0843
PCA3	0.0012	0.1782	*GABARAPL2*	0.0008	0.1191
TMEM45B	0.0012	0.1888	*HOXC4*	0.0010	0.1498
RP11.244H18.1.P712P	0.0017	0.2624	*TMEM45B*	0.0010	0.1563
MYOF	0.0018	0.2720	*P712P*	0.0012	0.1870
SULF2	0.0018	0.2720	*SULT1A1*	0.0016	0.2338
MARCH5	0.0019	0.2796	*GJB1*	0.0021	0.3090
GAPDH	0.0019	0.2826	*DLX1*	0.0023	0.3475
ERG3' exons 4-5	0.0023	0.3348	*PCA3*	0.0024	0.3598
TMPRSS2:ERG	0.0023	0.3376	*MSMB*	0.0027	0.3995
PTN	0.0025	0.3557	*MIR4435_1HG*	0.0028	0.4081
AURKA	0.0036	0.5179	*ITGBL1*	0.0033	0.4719
ERG3' exons 6-7	0.0041	0.5766	*DPP4*	0.0037	0.5275
TDRD	0.0054	0.7671	*SULF2*	0.0048	0.6756
PALM3	0.0056	0.7794	*ZNF577*	0.0051	0.7173
ITGBL1	0.0065	0.9059	*PPFIA2*	0.0052	0.7324
CD10	0.0071	0.9744	*PALM3*	0.0061	0.8496
TBP	0.0074	0.9897	*Met*	0.0064	0.8839
KLK2	0.0093	0.9897	*PTN*	0.0096	0.9961
ZNF577	0.0106	0.9897	*MCTP1*	0.0105	0.9961
RNF157	0.0107	0.9897	*CACNA1D*	0.0152	0.9961
PTPRC	0.0107	0.9897	*AMACR*	0.0174	0.9961
ANPEP	0.0136	0.9897	*CP*	0.0181	0.9961
NEAT1	0.0137	0.9897	*SSTR1*	0.0215	0.9961
Met	0.0137	0.9897	*GCNT1*	0.0221	0.9961
STOM	0.0148	0.9897	*PTPRC*	0.0310	0.9961
BTG2	0.0150	0.9897	*STOM*	0.0317	0.9961
AMACR	0.0159	0.9897	*IFT57*	0.0322	0.9961
MCTP1	0.0175	0.9897	*HIST1H2BF*	0.0333	0.9961

CACNA1D	0.0178	0.9897	*RP11_97O12.7*	0.0360	0.9961
ALAS1	0.0201	0.9897	*STEAP2*	0.0398	0.9961
ERG5	0.0223	0.9897	*TMCC2*	0.0409	0.9961
EN2	0.0225	0.9897	*MARCH5*	0.0431	0.9961
PPFIA2	0.0270	0.9897			
SIM2.short	0.0274	0.9897			
DLX1	0.0287	0.9897			
ITPR1	0.0289	0.9897			
TMEM86A	0.0296	0.9897			
TMCC2	0.0325	0.9897			
RP11.97O12.7	0.0336	0.9897			
HMBS	0.0343	0.9897			
MSMB	0.0355	0.9897			
IFT57	0.0375	0.9897			
HOXC4	0.0390	0.9897			

Supplementary Table 24 Lasso output for models detecting CB,Ca, Mets trend using KLK2 ratio data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
PCA3	***0.12***	***ERG3' exons 4-5***	***0.11***	***ERG3' exons 4-5***	***0.11***
ERG3' exons 4-5	***0.11***	***PCA3***	***0.09***	***PCA3***	***0.10***
APOC1	***0.09***	***HOXC6***	***0.08***	***HOXC6***	***0.08***
HOXC6	***0.08***	***APOC1***	***0.06***	***APOC1***	***0.06***
SLC12A1	***0.05***	***DLX1***	***0.03***	***DLX1***	***0.04***
DLX1	***0.04***	***SLC12A1***	***0.02***	***SLC12A1***	***0.03***
TDRD	***0.03***	***TDRD***	***0.01***	***TDRD***	***0.01***
TMPRSS2:ERG	***0.00***	***cp1***	***5.00***	***ERG3' exons 6-7***	***0.01***
ERG3' exons 6-7	***0.00***	***cp2***	***4.85x10^{-14}***	***cp1***	***5.09***
ZNF577	***-0.03***			***cp2***	***4.69x10^{-14}***
GCNT1	***-0.04***				
CP	***-0.09***				
cp1	***5.09***				
cp2	***3.99x10^{-14}***				

Supplementary Table 25 Lasso output for models detecting CB,Ca, Mets trend using KLK2 adjusted data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
MARCH5	***-2.11***	***MARCH5***	***-4.03***	***GABARAPL2***	***-3.18***
AMACR	***0.09***	***AMACR***	***1.17***	***GAPDH***	***0.82***
APOC1	***0.16***	***ANPEP***	***0.06***	***GJB1***	***1.07***
CASKIN1	***0.46***	***APOC1***	***0.28***	***HOXC6***	***1.52***
CDC20	***-0.15***	***AURKA***	***0.05***	***HPN***	***1.12***
EN2	***0.15***	***DLX1***	***0.02***	***KLK4***	***-1.13***
ERG5	***0.08***	***EN2***	***0.33***	***MFSD2A***	***0.97***
GABARAPL2	***-2.25***	***ERG3' exons 6-7***	***0.08***	***MIR4435.1HG .lOC541471***	***0.79***
GJB1	***0.56***	***ERG5***	***0.23***	***RAB17***	***0.23***
HIST1H1C	***1.00***	***GABARAPL2***	***-2.92***	***SPINK1***	***-0.50***
HMBS	***0.94***	***GJB1***	***0.67***	***UPK2***	***0.75***
HOXC6	***0.92***	***HMBS***	***1.97***	***cp1***	***5.36***
HPN	***0.32***	***HOXC6***	***1.02***	***cp2***	***-4.06 x10^{-}***

				14
IGFBP3	***-0.12***	***IFT57***	***1.08***	
KLK4	***-0.68***	***ITGBL1***	***-0.46***	
MFSD2A	***0.69***	***KLK4***	***-0.93***	
MIR4435.1HG.lOC541471	***0.28***	***Met***	***-0.30***	
MYOF	***-1.04***	***MFSD2A***	***0.96***	
NLRP3	***0.05***	***MIR4435.1HG.lOC541471***	***0.35***	
PALM3	***0.11***	***MYOF***	***-1.57***	
PCA3	***0.81***	***PALM3***	***0.19***	
PPAP2A	***0.21***	***PCA3***	***1.49***	
PTN	***-0.25***	***PPFIA2***	***-0.65***	
PTPRC	***0.12***	***PTN***	***-0.72***	
RNF157	***0.38***	***PTPRC***	***0.21***	
RP11.244H18.1.P712P	***-0.23***	***RNF157***	***0.83***	
RPL23AP53	***0.49***	***RP11.244H18.1.P712P***	***-0.88***	
SFRP4	***0.13***	***RP11.97O12.7***	***0.24***	
SIM2.short	***1.04***	***SIM2.short***	***1.77***	
SLC12A1	***0.07***	***SLC12A1***	***0.35***	
TBP	***0.22***	***STOM***	***0.11***	
TDRD	***0.10***	***TBP***	***1.04***	
Timp4	***-0.66***	***TDRD***	***0.17***	
TMCC2	***0.41***	***TMCC2***	***0.56***	
TMEM45B	***0.20***	***TMEM45B***	***0.36***	
TMEM86A	***0.13***	***TMEM86A***	***0.45***	
TMPRSS2:ERG	***0.08***	***TMPRSS2:ERG***	***0.15***	
UPK2	***0.56***	***UPK2***	***0.73***	
ZNF577	***-0.32***	***ZNF577***	***-1.06***	
cp1	***5.23***	***cp1***	***5.95***	
cp2	***-3.58x10-14***	***cp2***	***-3.55x10-14***	

Supplementary Table 26 Lasso output for models detecting CB,Ca, Mets trend using KLK3 adjusted data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***
MARCH5	***-2.05***	***MARCH5***	***-3.75***	***GABARAPL2***	***-1.96***
AMH	***0.42***	***AMACR***	***0.30***	***GJB1***	***0.66***
APOC1	***0.24***	***APOC1***	***0.45***	***HOXC6***	***1.57***
CP	***-0.20***	***DLX1***	***0.05***	***HPN***	***0.90***
EN2	***0.09***	***EN2***	***0.32***	***KLK4***	***-1.24***
ERG3' exons 4-5	***0.09***	***ERG3' exons 4-5***	***0.37***	***MIR4435.1HG.lOC541471***	***0.18***

ERG3' exons 6-7	***0.00***	***ERG5***	***0.22***	***RAB17***	***0.66***
ERG5	***0.05***	***GABARAPL2***	***-1.42***	***SPINK1***	***-0.68***
GABARAPL2	***-1.46***	***GJB1***	***0.48***	***UPK2***	***1.15***
GJB1	***0.18***	***HMBS***	***1.05***	***cp1***	***4.65***
HIST1H1C	***0.40***	***HOXC4***	***-0.40***	***cp2***	***6.23x10^{-14}***
HMBS	***0.41***	***HOXC6***	***1.00***		
HOXC6	***0.79***	***IFT57***	***0.30***		
HPN	***0.10***	***ITGBL1***	***-0.59***		
IGFBP3	***-0.35***	***ITPR1***	***0.29***		
KLK4	***-0.46***	***KLK2***	***0.06***		
MFSD2A	***0.33***	***KLK4***	***-0.42***		
MIR4435.1HG.lOC541471	***0.02***	***Met***	***-0.12***		
MMP25	***0.15***	***MFSD2A***	***0.51***		
MXI1	***-0.05***	***MIR4435.1HG.lOC541471***	***0.21***		
MYOF	***-0.41***	***MYOF***	***-0.63***		
PALM3	***0.06***	***NEAT1***	***0.22***		
PCA3	***1.17***	***PALM3***	***0.09***		
PSTPIP1	***0.05***	***PCA3***	***2.10***		
PTN	***-0.26***	***PPFIA2***	***-0.46***		
RNF157	***0.27***	***PTN***	***-0.75***		
RP11.244H18.1.P712P	***-0.04***	***PTPRC***	***0.03***		
RPL23AP53	***0.71***	***RNF157***	***0.59***		
SIM2.short	***0.87***	***RP11.244H18.1.P712P***	***-0.45***		
SLC12A1	***0.23***	***RP11.97O12.7***	***0.37***		
TBP	***0.65***	***SIM2.short***	***1.56***		
TDRD	***0.05***	***SLC12A1***	***0.49***		
Timp4	***-0.29***	***STOM***	***0.02***		
TMPRSS2:ERG	***0.02***	***TBP***	***1.70***		
UPK2	***0.84***	***TMCC2***	***0.23***		
cp1	***4.52***	***TMEM45B***	***0.24***		
cp2	***7.49x10^{-14}***	***TMEM86A***	***0.32***		
		UPK2	***0.91***		
		ZNF577	***-0.55***		
		cp1	***5.18***		
		cp2	***9.67x10^{-14}***		

Supplementary Table 27 Lasso output for models detecting CB,Ca, Mets trend using HK normalised data.

All Transcripts		*Significant Transcripts*		*Multiple testing corrected Transcripts*	
Transcript	***Beta***	***Transcript***	***Beta***	***Transcript***	***Beta***

MARCH5	***-0.02***	***MARCH5***	***-0.15***	***APOC1***	***0.12***
ACTR5	***-0.02***	***AMACR***	***0.07 0.07***	***ERG3' exons 4-5***	***0.17***
AMACR	***0.11***	***APOC1***	***0.20***	***HOXC6***	***0.12***
AMH	***0.15***	***CACNA1D***	***-0.34***	***HPN***	***0.05***
ANKRD34B	***-0.05***	***CP***	***-0.33***	***KLK2***	***-0.29***
APOC1	***0.22***	***DLX1***	***0.09***	***SLC12A1***	***0.08***
AR exon 9	***-0.07***	***ERG3' exons 4-5***	***0.06***	***SPINK1***	***0.04***
AURKA	***0.02***	***ERG3' exons 6-7***	***0.04***	***TMPRSS2:ERG***	***0.03***
CACNA1D	***-0.30***	***GABARAPL2***	***-0.41***	***UPK2***	***0.00***
CASKIN1	***0.18***	***GCNT1***	***-0.13***	***cp1***	***5.27***
CD10	***-0.01***	***GJB1***	***0.07***	***cp2***	***3.65 x10^{-14}***
CDC20	***-0.04***	***HOXC6***	***0.13***		
CP	***-0.30***	***IFT57***	***0.10***		
DLX1	***0.09***	***ITGBL1***	***-0.10***		
EN2	***0.03***	***KLK2***	***-0.11***		
ERG3' exons 4-5	***0.02***	***KLK4***	***-0.24***		
ERG3' exons 6-7	***0.06***	***MCTP1***	***0.05***		
GABARAPL2	***-0.87***	***Met***	***-0.01***		
GCNT1	***-0.13***	***MIR4435_1HG***	***0.09***		
GJB1	***0.08***	***MSMB***	***0.07***		
GOLM1	***-0.02***	***PALM3***	***0.13***		
HIST1H1C	***0.41***	***PCA3***	***0.21***		
HMBS	***0.24***	***PTN***	***-0.01***		
HOXC6	***0.17***	***SLC12A1***	***0.05***		
IGFBP3	***-0.01***	***SSTR1***	***0.11***		
ISX	***0.09***	***STOM***	***0.03***		
ITGBL1	***-0.05***	***SULF2***	***0.02***		
KLK2	***-0.01***	***TDRD***	***0.03***		
KLK4	***-0.23***	***TMCC2***	***0.28***		
LASS1	***-0.11***	***TMEM45B***	***0.30***		
MCTP1	***0.01***	***ZNF577***	***-0.17***		
MED4	***-0.03***	***cp1***	***6.15***		
MEMO1	***-0.08***	***cp2***	***3.32x10^{-14}***		
MEX3A	***0.03***				
MGAT5B	***0.15***				
MIC1	***-0.03***				
MIR146A	***-0.10***				
MIR4435_1HG	***0.09***				
MMP25	***0.11***				
MMP26	***-0.06***				
MXI1	***-0.07***				
NEAT1	***0.04***				
NLRP3	***0.12***				
OR52A2	***-0.08***				
PALM3	***0.09***				

PCA3	***0.15***
PDLIM5	***-0.17***
PPAP2A	***0.42***
PPFIA2	***0.03***
PPP1R12B	***-0.19***
RAB17	***0.00***
RNF157	***0.01***
RPL23AP53	***0.12***
SChLAP1	***0.00***
SEC61A1	***-0.02***
SFRP4	***0.07***
SIRT1	***-0.02***
SLC12A1	***0.04***
SLC43A1	***-0.12***
SSPO	***-0.30***
SSTR1	***0.07***
SULT1A1	***0.01***
TDRD	***0.03***
Timp4	***-0.13***
TMCC2	***0.20***
TMEM45B	***0.29***
TMEM86A	***0.07***
TMPRSS2:ERG	***0.03***
ZNF577	***-0.11***
cp1	***6.38***
cp2	***2.17x10^{-14}***

6.17 Looking for Housekeepers

Supplementary Table 28 Top twenty transcripts with the lowest variance in cell sediment urine fraction data

Transcript	*Variance*
MNX1	***0.94***
TWIST1	***0.95***
SSPO	***1.03***
SLC4A1 S	***1.04***
COL9A2	***1.09***
TERT	***1.12***
SSTR1	***1.14***
ABCB9	***1.15***
CASKIN1	***1.27***
MMP11	***1.28***
TMCC2	***1.33***
HIST1H1C	***1.42***
AMH	***1.42***
ISX	***1.50***
RPS11	***1.50***
AATF	***1.51***
HIST1H1E	***1.53***
VAX2	***1.55***
ARHGEF25	***1.66***
FDPS	***1.66***

Supplementary Table 29 Top twenty transcripts with the lowest IQR in cell sediment urine fraction data

Transcript	*IQR*
SSPO	***0.95***
RPS11	***1.01***
SLC4A1 S	***1.03***
TWIST1	***1.04***
ABCB9	***1.05***
HIST1H1E	***1.05***
B2M	***1.08***
VAX2	***1.13***
CASKIN1	***1.13***
RIOK3	***1.15***
CADPS	***1.16***
RP11_97O12.7	***1.18***
COL9A2	***1.18***
MMP26	***1.20***
MGAT5B	***1.20***
ISX	***1.21***
TFDP1	***1.21***

MNX1	***1.21***
SSTR1	***1.22***
RPL18A	***1.23***

Supplementary Table 30 Comparing the expression between all clinical categories using Tukey test, looking for potential house keeping transcripts.

Transcript	*Significant*	*Transcript*	*Significant*	*Transcript*	*Significant*
MARCH5	***0***	***PTN***	***0***	***OR52A2***	***2***
ABCB9	***0***	***PVT1***	***0***	***PCA3***	***2***
ACTR5	***0***	***RAB17***	***0***	***PDLIM5***	***2***
AMACR	***0***	***RNF157***	***0***	***PSTPIP1***	***2***
AMH	***0***	***RP11_97O12.7***	***0***	***SIM2 long***	***2***
AR exon 9	***0***	***RPL18A***	***0***	***SIM2 short***	***2***
AR exons 4-8	***0***	***RPL23AP53***	***0***	***SLC12A1***	***2***
ARHGEF25	***0***	***RPLP2***	***0***	***SNORA20***	***2***
BRAF	***0***	***RPS10***	***0***	***ST6GALNAC1***	***2***
CAMK2N2	***0***	***SACM1L***	***0***	***TMCC2***	***2***
CASKIN1	***0***	***SChLAP1***	***0***	***TMEM86A***	***2***
CDC20	***0***	***SLC4A1 S***	***0***	***AGR2***	***3***
CDC37L1	***0***	***SMAP1 exons7-8***	***0***	***BTG2***	***3***
CDKN3	***0***	***SMIM1***	***0***	***FOLH1***	***3***
CLU	***0***	***SPINK1***	***0***	***GABARAPL2***	***3***
COL10A1	***0***	***SPON2***	***0***	***MAPK8IP2***	***3***
CP	***0***	***STEAP2***	***0***	***SLC43A1***	***3***
MIATNB	***0***	***STEAP4***	***0***	***SNCA***	***3***
DLX1	***0***	***SYNM***	***0***	***TDRD***	***3***
ERG3' exons 4-5	***0***	***TERT***	***0***	***ANPEP***	***4***
ERG5'	***0***	***TFDP1***	***0***	***B2M***	***4***
FDPS	***0***	***Timp4***	***0***	***CLIC2***	***4***
GOLM1	***0***	***TMEM47***	***0***	***EIF2D***	***4***
HIST1H1C	***0***	***TRPM4***	***0***	***GAPDH***	***4***
HIST1H1E	***0***	***TWIST1***	***0***	***LASS1***	***4***
HIST1H2BF	***0***	***VAX2***	***0***	***MIR146A***	***4***
HIST3H2A	***0***	***VPS13A***	***0***	***MIR4435_1HG***	***4***
HMBS	***0***	***ZNF577***	***0***	***NLRP3***	***4***
HOXC4	***0***	***ALAS1***	***1***	***SRSF3***	***4***
IFT57	***0***	***ANKRD34B***	***1***	***SSPO***	***4***
IGFBP3	***0***	***AURKA***	***1***	***TERF2IP***	***4***
IMPDH2	***0***	***CKAP2L***	***1***	***TMPRSS2:ERG fusion***	***4***
ITGBL1	***0***	***COL9A2***	***1***	***AATF***	***5***
KLK2	***0***	***DNAH5***	***1***	***B4GALNT4***	***5***
KLK3 exons 2-3	***0***	***GJB1***	***1***	***CADPS***	***5***

KLK4	***0***	***KLK3 exons 1-2***	***1***	***CAMKK2***	***5***
LBH	***0***	***MED4***	***1***	***CCDC88B***	***5***
MAK	***0***	***MEMO1***	***1***	***HPN***	***5***
MCM7	***0***	***MMP26***	***1***	***ISX***	***5***
MDK	***0***	***MNX1***	***1***	***ITPR1***	***5***
Met	***0***	***NAALADL2***	***1***	***MFSD2A***	***5***
MEX3A	***0***	***P712P***	***1***	***MMP25***	***5***
MGAT5B	***0***	***RPS11***	***1***	***SEC61A1***	***5***
MIC1	***0***	***SFRP4***	***1***	***HPRT***	***6***
MKi67	***0***	***SSTR1***	***1***	***RIOK3***	***6***
MMP11	***0***	***SULT1A1***	***1***	***SERPINB5***	***6***
MSMB	***0***	***TBP***	***1***	***SIRT1***	***6***
MYOF	***0***	***TMEM45B***	***1***	***ERG3' exons 6-7***	***7***
NKAIN1	***0***	***UPK2***	***1***	***HOXC6***	***7***
OGT	***0***	***CACNA1D***	***2***	***STOM***	***9***
PALM3	***0***	***CD10***	***2***	***APOC1***	***11***
PCSK6	***0***	***DPP4***	***2***	***MCTP1***	***11***
PECI	***0***	***EN2***	***2***	***NEAT1***	***11***
PPAP2A	***0***	***GCNT1***	***2***	***PTPRC***	***12***
PPFIA2	***0***	***HIST1H2BG***	***2***	***SULF2***	***12***
PPP1R12B	***0***	***MXI1***	***2***		

6.18 Cancer Vs CB

Supplementary Table 31 Transcripts that have significant differential expression between CB and cancer samples (L, I, H) in the baseline normalised NanoString data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***0.0002***	***0.024***	***0.0014***	***0.2049***	***1.64***
ERG3' exons 6-7	***$2.84x10^{-07}$***	***$4.74x10^{-05}$***	***0.0008***	***0.128***	***1.38***
TMPRSS2:ERG	***$4.52x10^{-05}$***	***0.0069***	***0.0013***	***0.1979***	***1.31***
SLC43A1	***0.0003***	***0.0406***	***0.0019***	***0.2745***	***1.17***
CLIC2	***$2.66x10^{-05}$***	***0.0042***	***0.001***	***0.1645***	***1.05***
B4GALNT4	***$3.38x10^{-05}$***	***0.0053***	***0.0012***	***0.1807***	***1.04***
CADPS	***$1.37x10^{-05}$***	***0.0022***	***0.0004***	***0.0682***	***1.04***
CKAP2L	***0.0116***	***1***	***0.0033***	***0.4318***	***1.01***
HPN	***$7.04x10^{-05}$***	***0.0103***	***0.0006***	***0.1041***	***0.97***
LASS1	***0.0002***	***0.022***	***0.0011***	***0.1703***	***0.97***
TDRD	***0.0002***	***0.022***	***0.0047***	***0.5935***	***0.97***
SFRP4	***0.0004***	***0.0478***	***0.0031***	***0.4076***	***0.87***
OR52A2	***0.0343***	***1***	***0.0284***	***0.9777***	***0.85***
ANKRD34B	***0.0013***	***0.1706***	***0.0093***	***0.9777***	***0.83***
MAPK8IP2	***0.0063***	***0.6758***	***0.0067***	***0.7955***	***0.8***
PCA3	***0.0004***	***0.0565***	***0.0019***	***0.2745***	***0.8***
CDKN3	***0.024***	***1***	***0.011***	***0.9777***	***0.76***

ERG5'	***0.0016***	***0.2005***	***0.009***	***0.9777***	***0.68***
MFSD2A	***1.32x10^{-05}***	***0.0021***	***0.001***	***0.1645***	***0.66***
MMP25	***7.80x10^{-05}***	***0.0113***	***0.0008***	***0.1278***	***0.66***
APOC1	***1.85x10^{-06}***	***0.0003***	***0.0004***	***0.0586***	***0.65***
TMCC2	***0.0126***	***1***	***0.0075***	***0.8811***	***0.65***
NKAIN1	***0.0379***	***1***	***0.0429***	***0.9777***	***0.62***
SIM2 long	***3.72x10^{-05}***	***0.0057***	***0.0031***	***0.4076***	***0.62***
MCTP1	***3.97x10^{-07}***	***6.59x10^{-05}***	***0.0002***	***0.0406***	***0.61***
ISX	***0.0007***	***0.086***	***0.0024***	***0.3365***	***0.6***
X?	***0.0032***	***0.3694***			***0.59***
MMP26	***0.0448***	***1***			***0.56***
AMH	***0.0032***	***0.3694***	***0.0335***	***0.9777***	***0.55***
SLC12A1	***0.0014***	***0.1798***	***0.0064***	***0.7769***	***0.55***
SULF2	***9.18x10^{-06}***	***0.0015***	***0.0011***	***0.1754***	***0.55***
CCDC88B	***6.34x10^{-05}***	***0.0094***	***0.0012***	***0.1785***	***0.54***
NLRP3	***0.0024***	***0.29***	***0.002***	***0.2817***	***0.54***
UPK2	***0.0071***	***0.7532***	***0.0147***	***0.9777***	***-0.54***
TMEM86A	***0.0001***	***0.0202***	***0.0019***	***0.2742***	***0.53***
CAMKK2	***2.13x10^{-06}***	***0.0003***	***0.0005***	***0.0859***	***0.51***
FOLH1	***0.013***	***1***	***0.0121***	***0.9777***	***0.49***
ANPEP	***0.0011***	***0.1472***	***0.003***	***0.4076***	***0.46***
SRSF3	***0.0002***	***0.0311***	***0.0038***	***0.4879***	***0.45***
MIR146A	***0.0019***	***0.235***	***0.0023***	***0.3187***	***0.44***
GCNT1	***0.003***	***0.353***	***0.0035***	***0.4645***	***0.43***
SIRT1	***5.71x10^{-05}***	***0.0085***	***0.0012***	***0.1785***	***0.41***
SERPINB5	***0.031***	***1***			***-0.4***
NAALADL2	***0.0059***	***0.6513***	***0.0308***	***0.9777***	***-0.38***
SNORA20	***0.0081***	***0.8455***	***0.0206***	***0.9777***	***0.37***
CDC20	***0.0063***	***0.6758***	***0.0117***	***0.9777***	***0.35***
TMEM45B	***0.0253***	***1***			***-0.35***
AATF	***5.14x10^{-05}***	***0.0077***	***0.0014***	***0.2012***	***0.34***
IGFBP3	***0.0067***	***0.7136***			***-0.34***
AURKA	***0.0016***	***0.1909***	***0.0056***	***0.6828***	***0.33***
CD10	***0.0014***	***0.1798***	***0.0053***	***0.6649***	***0.33***
PTPRC	***4.62x10^{-05}***	***0.007***	***0.0014***	***0.2012***	***0.33***
SSTR1	***0.0164***	***1***	***0.0177***	***0.9777***	***0.33***
SEC61A1	***0.0002***	***0.0285***	***0.0055***	***0.682***	***0.32***
SIM2 short	***0.0193***	***1***	***0.0164***	***0.9777***	***0.32***
SNCA	***0.0016***	***0.1909***	***0.007***	***0.8231***	***0.32***
MMP11	***0.013***	***1***	***0.0222***	***0.9777***	***0.31***
SPINK1	***0.0032***	***0.3694***	***0.0133***	***0.9777***	***-0.31***
HPRT	***0.0005***	***0.0666***	***0.0067***	***0.7955***	***0.29***
PSTPIP1	***0.046***	***1***	***0.0226***	***0.9777***	***0.28***
HOXC4	***0.0356***	***1***			***0.26***
RIOK3	***1.03x10^{-06}***	***0.0002***	***0.0013***	***0.1979***	***0.26***
	0.0469	***1***			***0.25***
	0.0266	***1***			***0.24***
EN2	***0.0192***	***1***	***0.0282***	***0.9777***	***0.24***
MEX3A	***0.0224***	***1***	***0.0252***	***0.9777***	***0.24***
CACNA1D	***0.0014***	***0.1798***	***0.0054***	***0.681***	***0.23***
MIR4435_1HG	***3.72x10^{-05}***	***0.0057***	***0.0026***	***0.3533***	***0.23***

MXI1	***0.0001***	***0.0168***	***0.0026***	***0.3537***	***0.23***
DPP4	***0.0048***	***0.54***	***0.0108***	***0.9777***	***0.22***
CASKIN1	***0.0343***	***1***	***0.0315***	***0.9777***	***0.21***
HIST3H2A	***0.0193***	***1***			***0.21***
ITPR1	***$4.15x10^{-05}$***	***0.0063***	***0.0019***	***0.2745***	***0.21***
NEAT1	***$1.89x10^{-05}$***	***0.003***	***0.0009***	***0.1377***	***0.21***
STOM	***0.0116***	***1***	***0.0219***	***0.9777***	***0.21***
PDLIM5	***0.0007***	***0.0916***	***0.0046***	***0.5856***	***0.2***
BTG2	***0.0037***	***0.4198***	***0.0173***	***0.9777***	***0.19***
GABARAPL2	***0.0003***	***0.044***	***0.0025***	***0.3502***	***0.19***
HIST1H2BG	***0.0048***	***0.54***	***0.0064***	***0.7753***	***0.19***
MAK	***0.0183***	***1***			***0.19***
EIF2D	***0.0026***	***0.309***	***0.0125***	***0.9777***	***0.18***
MGAT5B			***0.0446***	***0.9777***	***0.18***
TBP	***0.0063***	***0.6758***	***0.0358***	***0.9777***	***0.18***
TWIST1	***0.0138***	***1***			***0.18***
MED4	***0.0146***	***1***	***0.0361***	***0.9777***	***0.17***
TERF2IP	***$7.80x10^{-05}$***	***0.0113***	***0.0019***	***0.2745***	***0.17***
GAPDH	***$2.98x10^{-05}$***	***0.0047***	***0.0007***	***0.109***	***0.16***
ACTR5	***0.0164***	***1***	***0.0123***	***0.9777***	***0.15***
B2M	***0.0002***	***0.0285***	***0.0044***	***0.5721***	***0.14***
	0.0438	***1***			***0.11***
SACM1L	***0.0266***	***1***	***0.0347***	***0.9777***	***0.11***
RPL18A	***0.0123***	***1***			***0.1***
	0.0138	***1***			***0.09***
STEAP4			***0.0474***	***0.9777***	***0.09***
HIST1H2BF			***0.0424***	***0.9777***	***0.08***
MEMO1	***0.0361***	***1***	***0.0357***	***0.9777***	***0.08***
RPS11	***0.0173***	***1***			***0.08***
HMBS			***0.0391***	***0.9777***	***0.06***
SLC4A1 S			***0.0223***	***0.9777***	***0.03***
SMAP1 exons 7-8			***0.0458***	***0.9777***	***0.03***

Supplementary Table 32 Transcripts that have significant differential expression between CB and cancer samples (L, I, H) in the *KLK2* ratio NanoString data.

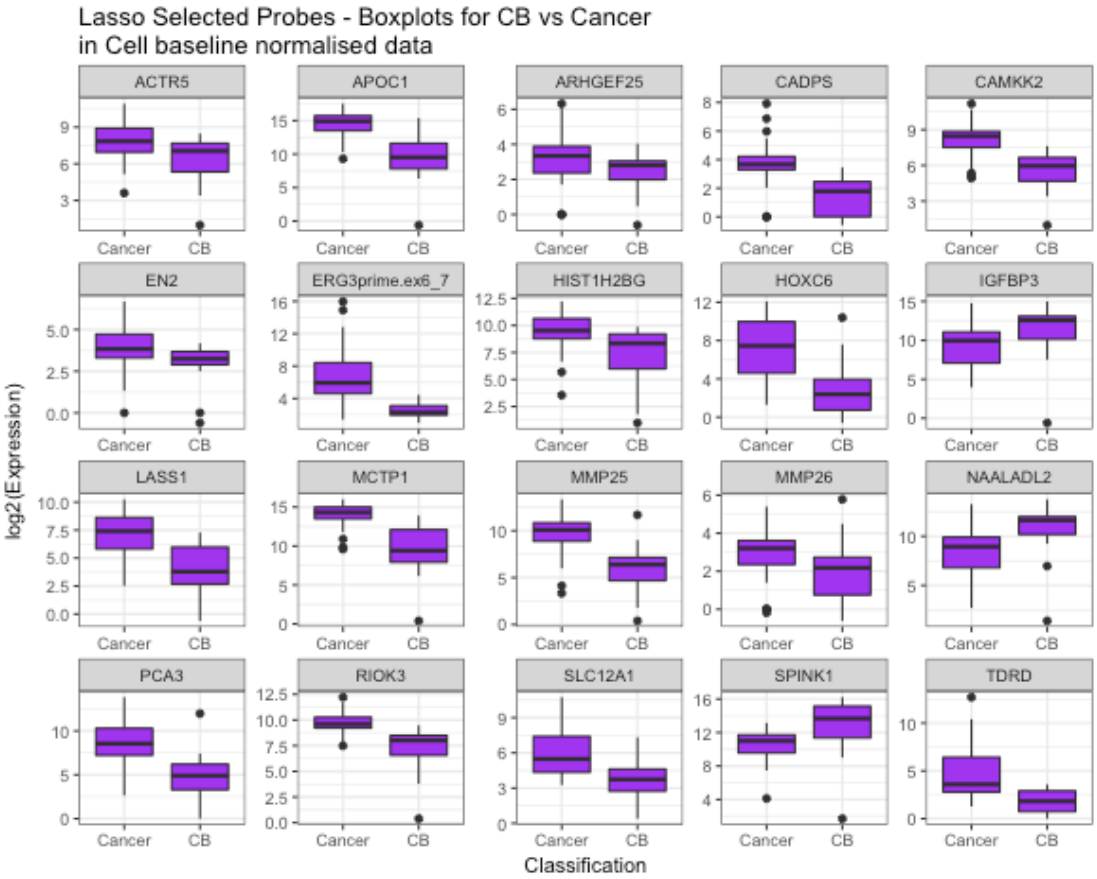
	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***$6.80x10^{-05}$***	***0.01***	***0.004***	***0.63***	***0.21***
ERG3' exons 6-7	***$7.80x10^{-05}$***	***0.01***	***0.001***	***0.24***	***0.18***
TDRD	***0.0004***	***0.06***	***0.004***	***0.72***	***0.18***
SLC43A1	***0.002***	***0.32***			***0.17***
CADPS	***0.004***	***0.67***	***0.01***	***1***	***0.16***
ERG5'	***0.01***	***0.99***			***0.15***
B4GALNT4	***0.01***	***0.87***			***0.14***
SLC12A1	***0.003***	***0.54***	***0.03***	***1***	***0.13***
TMCC2	***0.05***	***0.99***	***0.05***	***1***	***0.13***

TMPRSS2:ERG	***0.001***	***0.17***	***0.01***	***1***	***0.13***
CKAP2L	***0.02***	***0.99***	***0.02***	***1***	***0.12***
MFSD2A	***0.01***	***0.99***	***0.02***	***1***	***0.12***
CLIC2	***0.004***	***0.58***	***0.01***	***1***	***0.11***
LASS1	***0.01***	***0.89***	***0.01***	***1***	***0.11***
MMP25	***0.01***	***0.99***	***0.02***	***1***	***0.11***
PCA3	***$3.72x10^{-05}$***	***0.01***	***0.003***	***0.41***	***0.1***
ANKRD34B	***0.04***	***0.99***			***0.09***
HPN	***0.001***	***0.23***	***0.01***	***1***	***0.09***
TMEM86A	***0.01***	***0.99***	***0.03***	***1***	***0.09***
NAALADL2	***0.03***	***0.99***	***0.04***	***1***	***-0.09***
UPK2	***0.03***	***0.99***			***-0.09***
APOC1	***0.004***	***0.65***			***0.08***
CCDC88B	***0.01***	***0.99***	***0.05***	***1***	***0.08***
ST6GALNAC1	***0.03***	***0.99***			***-0.08***
ISX	***0.01***	***0.99***			***0.07***
MCTP1	***0.01***	***0.99***	***0.03***	***1***	***0.07***
MIR146A	***0.02***	***0.99***			***0.07***
NLRP3			***0.03***	***1***	***0.07***
SULF2	***0.01***	***0.99***			***0.07***
SERPINB5	***0.03***	***0.99***			***-0.06***
SFRP4	***0.04***	***0.99***			***0.06***
FOLH1	***0.03***	***0.99***			***0.05***
OR52A2			***0.03***	***1***	***0.05***
SIM2.long	***0.003***	***0.51***	***0.01***	***1***	***0.05***
CAMKK2	***0.004***	***0.62***	***0.04***	***1***	***0.04***
GCNT1			***0.02***	***1***	***0.03***
HIST1H2BG			***0.04***	***1***	***0.03***

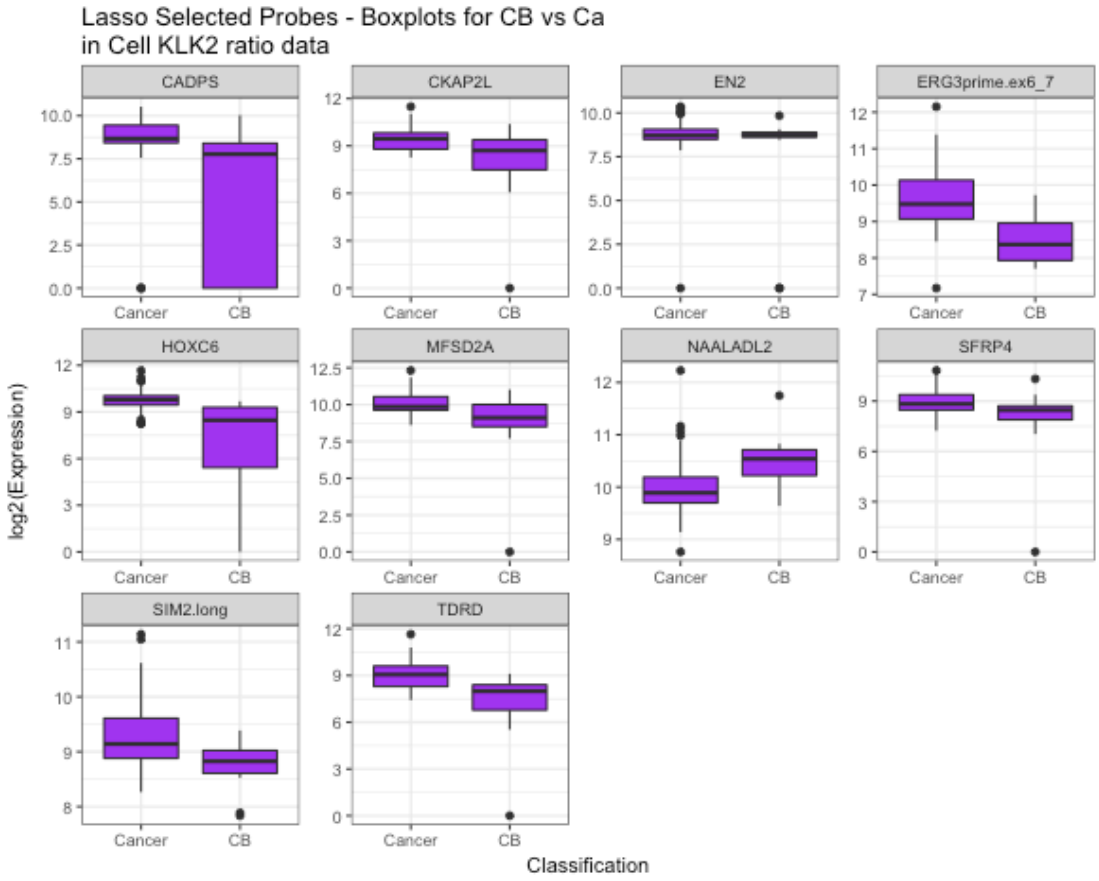
Supplementary Table 33 Transcripts that have significant differential expression between CB and cancer samples (L, I, H) in the HK normalised NanoString data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***0.0002***	***0.0374***	***0.0019***	***0.3087***	***1.5***
ERG3' exons 6-7	***0.0006***	***0.1045***	***0.0228***	***0.9861***	***1.1***
TMPRSS2:ERG	***0.0036***	***0.5527***	***0.0069***	***0.9861***	***1.1***
CP	***0.0146***	***0.9924***	***0.0109***	***0.9861***	***-1***
TDRD	***0.001***	***0.153***	***0.0105***	***0.9861***	***0.9***
NAALADL2	***$3.33x10^{-05}$***	***0.0056***	***0.0012***	***0.2012***	***-0.8***
SLC43A1	***0.0005***	***0.0895***	***0.0168***	***0.9861***	***0.8***
ST6GALNAC1	***0.0008***	***0.1311***	***0.0238***	***0.9861***	***-0.8***
SPINK1	***$7.80x10^{-05}$***	***0.0129***			***-0.7***
UPK2	***0.0007***	***0.1128***	***0.0026***	***0.4313***	***-0.7***
CADPS	***0.0083***	***0.9924***	***0.0076***	***0.9861***	***0.7***
HPN	***0.0022***	***0.3485***	***0.0072***	***0.9861***	***0.7***
MFSD2A	***0.0123***	***0.9924***	***0.0082***	***0.9861***	***0.7***
DNAH5	***0.0116***	***0.9924***			***-0.7***
IGFBP3	***0.0086***	***0.9924***			***-0.7***
SERPINB5	***0.0003***	***0.0489***	***0.0205***	***0.9861***	***-0.6***
B4GALNT4	***0.0273***	***0.9924***			***0.6***
CLIC2	***0.0055***	***0.8138***	***0.0097***	***0.9861***	***0.6***
LASS1	***0.0055***	***0.8138***	***0.0182***	***0.9861***	***0.6***
PCA3	***0.0006***	***0.1045***	***0.005***	***0.8153***	***0.6***
ITGBL1	***0.0227***	***0.9924***			***-0.6***
CCDC88B	***0.024***	***0.9924***	***0.0349***	***0.9861***	***0.5***
ERG5'	***0.0164***	***0.9924***			***0.5***
ISX	***0.0169***	***0.9924***	***0.0124***	***0.9861***	***0.5***
MMP25	***0.0071***	***0.9924***			***0.5***
AGR2	***0.0227***	***0.9924***	***0.0348***	***0.9861***	***-0.5***
GJB1	***0.0024***	***0.3722***	***0.0136***	***0.9861***	***-0.5***
MCTP1	***0.0164***	***0.9924***	***0.0349***	***0.9861***	***0.4***
PPAP2A	***0.0008***	***0.1311***			***-0.4***
PPP1R12B	***0.0138***	***0.9924***			***-0.4***
TMEM86A	***0.0037***	***0.5561***	***0.0227***	***0.9861***	***-0.4***
APOC1	***0.003***	***0.4577***	***0.0172***	***0.9861***	***0.3***
CAMKK2	***0.024***	***0.9924***			***0.3***
GCNT1	***0.0482***	***0.9924***	***0.0425***	***0.9861***	***0.3***
SIM2 long	***0.0109***	***0.9924***	***0.0125***	***0.9861***	***0.3***
SLC12A1	***0.0398***	***0.9924***			***0.3***
SULF2			***0.0391***	***0.9861***	***0.3***
MDK	***0.0022***	***0.3485***			***-0.3***
MNX1	***0.0193***	***0.9924***			***-0.3***
OGT	***0.0253***	***0.9924***			***-0.3***
PALM3	***0.0343***	***0.9924***			***-0.3***
RAB17	***0.0052***	***0.7725***			***-0.3***
RPS10	***0.0155***	***0.9924***			***-0.3***
STEAP2	***0.0076***	***0.9924***			***-0.3***
MIR146A			***0.0347***	***0.9861***	***0.2***

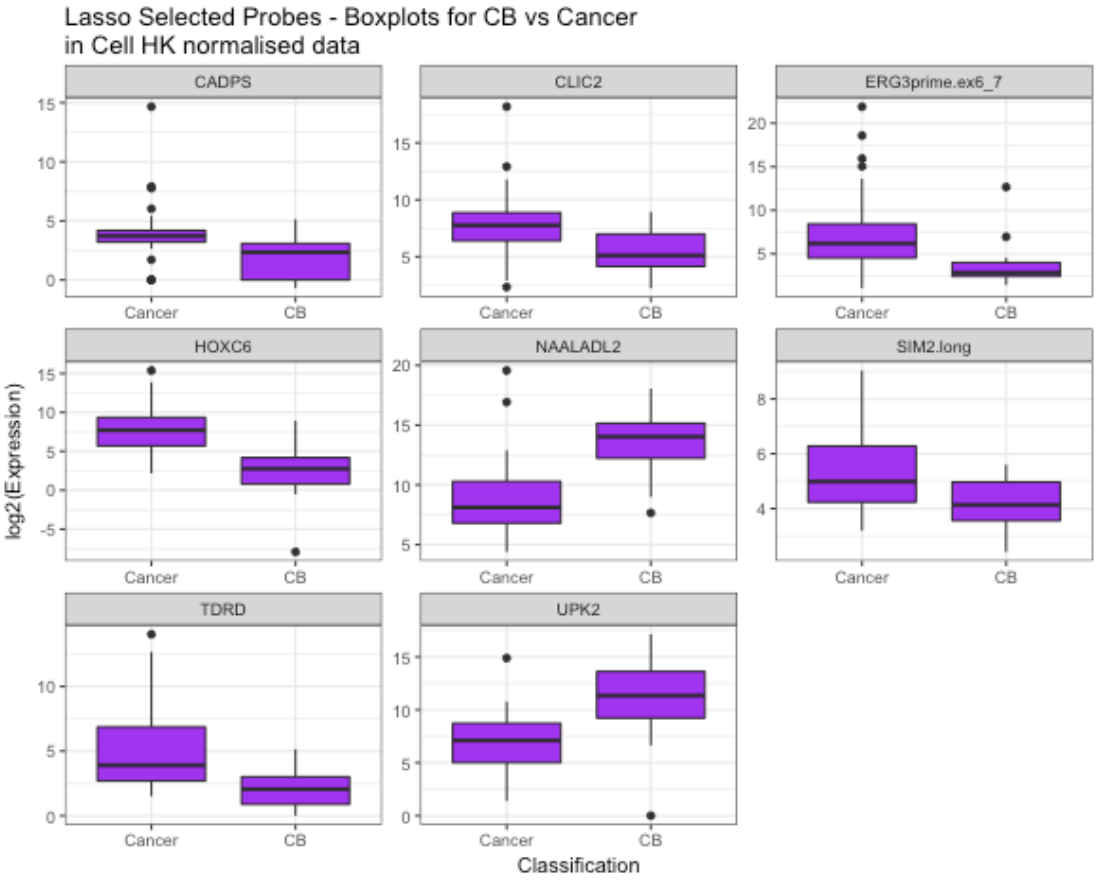
RIOK3	***0.0379***	***0.9924***	***0.0486***	***0.9861***	***0.2***
TMEM86A	***0.0155***	***0.9924***			***0.2***
HIST1H1C	***0.0379***	***0.9924***			***-0.2***
IFT57	***0.0081***	***0.9924***			***-0.2***
IMPDH2	***0.0361***	***0.9924***			***-0.2***
MSMB	***0.0091***	***0.9924***			***-0.2***
MYOF	***0.0081***	***0.9924***			***-0.2***
PTN	***0.0266***	***0.9924***			***-0.2***
RPL18A	***0.0253***	***0.9924***			***-0.2***
RPLP2	***0.0138***	***0.9924***			***-0.2***
RPS11	***0.0253***	***0.9924***			***-0.2***
ZNF577	***0.0193***	***0.9924***			***-0.2***
HIST1H1E	***0.0438***	***0.9924***			***-0.1***



Supplementary Figure 13 Boxplots of all of the Lasso selected probes involved in CB vs. cancer (L, I, and H) models from the baseline normalised data.



Supplementary Figure 14 Boxplots of all of the Lasso selected probes involved in CB vs. cancer (L, I, and H) models from the *KLK2* ratio data.



Supplementary Figure 15 Boxplots of all of the Lasso selected probes involved in CB vs. cancer (L, I, and H) models from the HK normalised data.

Supplementary Table 34 Random Forest results for Ca vs CBN baseline normalisation

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
TMPRSS2:ERG	***0.54***	***167***	***APOC1***	***0.57***	***85***	***TMPRSS2:ERG***	***0.78***	***94***
APOC1	***0.53***	***166***	***SPINK1***	***0.47***	***84***	***APOC1***	***0.59***	***93***
ERG3' exons 6-7	***0.53***	***165***	***TMPRSS2:ERG***	***0.47***	***83***	***NEAT1***	***0.49***	***92***
NEAT1	***0.45***	***164***	***RIOK3***	***0.46***	***82***	***RIOK3***	***0.47***	***91***
RIOK3	***0.41***	***163***	***ERG3' exons 6-7***	***0.44***	***81***	***ERG3' exons 6-7***	***0.39***	***90***
SIM2 long	***0.38***	***162***	***NEAT1***	***0.44***	***80***	***MCTP1***	***0.36***	***89***
MCTP1	***0.35***	***161***	***SIM2 long***	***0.39***	***79***	***MFSD2A***	***0.35***	***88***
SPINK1	***0.32***	***160***	***MCTP1***	***0.38***	***78***	***CADPS***	***0.34***	***87***
CCDC88B	***0.28***	***159***	***CADPS***	***0.36***	***77***	***SIM2 long***	***0.28***	***86***
CADPS	***0.27***	***158***	***MFSD2A***	***0.32***	***76***	***SPINK1***	***0.25***	***85***
MXI1	***0.22***	***157***	***CCDC88B***	***0.32***	***75***	***CCDC88B***	***0.24***	***84***
MFSD2A	***0.21***	***156***	***CKAP2L***	***0.24***	***74***	***CKAP2L***	***0.23***	***83***
MMP25	***0.21***	***155***	***CAMKK2***	***0.21***	***73***	***GAPDH***	***0.22***	***82***
CKAP2L	***0.19***	***154***	***SLC43A1***	***0.20***	***72***	***MXI1***	***0.21***	***81***
HOXC6	***0.16***	***153***	***MIR4435_1HG***	***0.19***	***71***	***CAMKK2***	***0.20***	***80***
SULF2	***0.16***	***152***	***MMP25***	***0.19***	***70***	***SLC43A1***	***0.17***	***79***
MIR4435_1HG	***0.15***	***151***	***SULF2***	***0.17***	***69***	***MMP25***	***0.16***	***78***
SIRT1	***0.15***	***150***	***HOXC6***	***0.17***	***68***	***AURKA***	***0.14***	***77***
CAMKK2	***0.14***	***149***	***GAPDH***	***0.16***	***67***	***HOXC6***	***0.14***	***76***
AURKA	***0.13***	***148***	***UPK2***	***0.15***	***66***	***HPRT***	***0.13***	***75***
GAPDH	***0.13***	***147***	***ISX***	***0.14***	***65***	***SIRT1***	***0.13***	***74***
B4GALNT4	***0.11***	***146***	***AATF***	***0.13***	***64***	***MIR4435_1HG***	***0.12***	***73***
TDRD	***0.11***	***145***	***TDRD***	***0.13***	***63***	***HPN***	***0.12***	***72***
SLC43A1	***0.10***	***144***	***MXI1***	***0.12***	***62***	***TDRD***	***0.11***	***71***
UPK2	***0.10***	***143***	***AURKA***	***0.11***	***61***	***IGFBP3***	***0.10***	***70***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HPN	***0.09***	***142***	***HPRT***	***0.11***	***60***	***SULF2***	***0.10***	***69***
IGFBP3	***0.09***	***141***	***SLC4A1 S***	***0.10***	***59***	***UPK2***	***0.09***	***68***
SNCA	***0.09***	***140***	***HPN***	***0.09***	***58***	***MMP26***	***0.09***	***67***
TMEM45B	***0.07***	***139***	***PCA3***	***0.09***	***57***	***AATF***	***0.09***	***66***
AATF	***0.07***	***138***	***MAPK8IP2***	***0.09***	***56***	***MAPK8IP2***	***0.09***	***65***
LASS1	***0.07***	***137***	***GCNT1***	***0.09***	***55***	***SNCA***	***0.09***	***64***
GCNT1	***0.06***	***136***	***STOM***	***0.09***	***54***	***LASS1***	***0.09***	***63***
HPRT	***0.06***	***135***	***B4GALNT4***	***0.08***	***53***	***CD10***	***0.08***	***62***
NAALADL2	***0.06***	***134***	***SLC12A1***	***0.08***	***52***	***GCNT1***	***0.08***	***61***
ISX	***0.06***	***133***	***PTPRC***	***0.07***	***51***	***TMCC2***	***0.07***	***60***
AMH	***0.05***	***132***	***DPP4***	***0.07***	***50***	***SFRP4***	***0.07***	***59***
SLC12A1	***0.05***	***131***	***CD10***	***0.06***	***49***	***ITPR1***	***0.06***	***58***
SLC4A1 S	***0.05***	***130***	***EN2***	***0.06***	***48***	***EN2***	***0.06***	***57***
CACNA1D	***0.05***	***129***	***SNCA***	***0.06***	***47***	***B4GALNT4***	***0.06***	***56***
RPL23AP53	***0.05***	***128***	***PDLIM5***	***0.05***	***46***	***ERG5'***	***0.06***	***55***
CDC37L1	***0.05***	***127***	***TMCC2***	***0.05***	***45***	***SLC12A1***	***0.06***	***54***
PCA3	***0.05***	***126***	***SIRT1***	***0.05***	***44***	***ISX***	***0.06***	***53***
ACTR5	***0.05***	***125***	***MGAT5B***	***0.05***	***43***	***PCA3***	***0.05***	***52***
PTPRC	***0.04***	***124***	***SNORA20***	***0.05***	***42***	***PDLIM5***	***0.05***	***51***
MMP26	***0.04***	***123***	***TMEM86A***	***0.05***	***41***	***STOM***	***0.05***	***50***
RNF157	***0.04***	***122***	***LASS1***	***0.04***	***40***	***ACTR5***	***0.05***	***49***
MAPK8IP2	***0.04***	***121***	***HIST1H2BG***	***0.04***	***39***	***DPP4***	***0.05***	***48***
STOM	***0.04***	***120***	***SRSF3***	***0.04***	***38***	***ERG3' exons 4-5***	***0.05***	***47***
CDC20	***0.04***	***119***	***NAALADL2***	***0.04***	***37***	***TMEM45B***	***0.04***	***46***
EN2	***0.04***	***118***	***AMH***	***0.04***	***36***	***AMH***	***0.04***	***45***
SRSF3	***0.04***	***117***	***STEAP4***	***0.04***	***35***	***NAALADL2***	***0.04***	***44***
ERG5'	***0.04***	***116***	***CACNA1D***	***0.04***	***34***	***HIST3H2A***	***0.04***	***43***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SFRP4	***0.04***	***115***	***ACTR5***	***0.04***	***33***	***TMEM86A***	***0.04***	***42***
MYOF	***0.04***	***114***	***ANKRD34B***	***0.03***	***32***	***FOLH1***	***0.04***	***41***
CLIC2	***0.04***	***113***	***SFRP4***	***0.03***	***31***	***HIST1H2BG***	***0.04***	***40***
HIST1H2BG	***0.04***	***112***	***CDC20***	***0.02***	***30***	***FDPS***	***0.03***	***39***
MAK	***0.04***	***111***	***SEC61A1***	***0.02***	***29***	***CLIC2***	***0.03***	***38***
Timp4	***0.03***	***110***	***CLIC2***	***0.02***	***28***	***MIR146A***	***0.03***	***37***
TMEM86A	***0.03***	***109***	***HIST1H2BF***	***0.02***	***27***	***SRSF3***	***0.03***	***36***
PPP1R12B	***0.03***	***108***	***FOLH1***	***0.02***	***26***	***PTPRC***	***0.03***	***35***
STEAP4	***0.03***	***107***	***ANPEP***	***0.02***	***25***	***MAK***	***0.03***	***34***
DPP4	***0.03***	***106***	***ERG5'***	***0.02***	***24***	***SEC61A1***	***0.03***	***33***
CD10	***0.03***	***105***	***MIR146A***	***0.02***	***23***	***TWIST1***	***0.03***	***32***
SULT1A1	***0.03***	***104***	***TERF2IP***	***0.02***	***22***	***SERPINB5***	***0.02***	***31***
PDLIM5	***0.03***	***103***	***MED4***	***0.02***	***21***	***NLRP3***	***0.02***	***30***
P712P	***0.03***	***102***	***ITPR1***	***0.01***	***20***	***CDC20***	***0.02***	***29***
MSMB	***0.03***	***101***	***BTG2***	***0.01***	***19***	***RPS11***	***0.02***	***28***
ERG3' exons 4-5	***0.03***	***100***	***NKAIN1***	***0.01***	***18***	***CACNA1D***	***0.02***	***27***
AGR2	***0.02***	***99***	***MEMO1***	***0.01***	***17***	***SACM1L***	***0.02***	***26***
PECI	***0.02***	***98***	***CASKIN1***	***0.01***	***16***	***RPL18A***	***0.02***	***25***
MNX1	***0.02***	***97***	***SMAP1 exons 7-8***	***0.01***	***15***	***ANKRD34B***	***0.01***	***24***
PPAP2A	***0.02***	***96***	***TBP***	***0.01***	***14***	***TERF2IP***	***0.01***	***23***
PPFIA2	***0.02***	***95***	***SIM2 short***	***0.01***	***13***	***GABARAPL2***	***0.01***	***22***
PALM3	***0.02***	***94***	***MEX3A***	***0.01***	***12***	***SNORA20***	***0.01***	***21***
ITPR1	***0.02***	***93***	***CDKN3***	***0.01***	***11***	***MEX3A***	***0.01***	***20***
RPS11	***0.02***	***92***	***SACM1L***	***0.01***	***10***	***HOXC4***	***0.01***	***19***
VAX2	***0.02***	***91***	***MMP11***	***0.01***	***9***	***ALAS1***	***0.01***	***18***
EIF2D	***0.02***	***90***	***OR52A2***	***0.00***	***8***	***CAMK2N2***	***0.01***	***17***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
FOLH1	***0.02***	***89***	***GABARAPL2***	***0.00***	***7***	***MED4***	***0.01***	***16***
PVT1	***0.02***	***88***	***EIF2D***	***0.00***	***6***	***NKAIN1***	***0.01***	***15***
AR.ex9	***0.02***	***87***	***PSTPIP1***	***0.00***	***5***	***MMP11***	***0.01***	***14***
ANKRD34B	***0.02***	***86***	***SSTR1***	***0.00***	***4***	***ANPEP***	***0.01***	***13***
MKi67	***0.02***	***85***	***NLRP3***	***0.00***	***3***	***ARHGEF25***	***0.01***	***12***
MGAT5B	***0.02***	***84***	***HMBS***	***0.00***	***2***	***CASKIN1***	***0.01***	***11***
SNORA20	***0.02***	***83***	***B2M***	***0.00***	***1***	***B2M***	***0.01***	***10***
IMPDH2	***0.01***	***82***				***OR52A2***	***0.00***	***9***
MED4	***0.01***	***81***				***BTG2***	***0.00***	***8***
GJB1	***0.01***	***80***				***SSTR1***	***0.00***	***7***
HIST3H2A	***0.01***	***79***				***SIM2 short***	***0.00***	***6***
CAMK2N2	***0.01***	***78***				***EIF2D***	***0.00***	***5***
OGT	***0.01***	***77***				***MEMO1***	***0.00***	***4***
HIST1H2BF	***0.01***	***76***				***CDKN3***	***0.00***	***3***
DLX1	***0.01***	***75***				***TBP***	***0.00***	***2***
MCM7	***0.01***	***74***				***PSTPIP1***	***0.00***	***1***
SEC61A1	***0.01***	***73***						
PSTPIP1	***0.01***	***72***						
ARHGEF25	***0.01***	***71***						
IFT57	***0.01***	***70***						
GOLM1	***0.01***	***69***						
TMCC2	***0.01***	***68***						
SERPINB5	***0.01***	***67***						
TERF2IP	***0.01***	***66***						
SPON2	***0.01***	***65***						
SSPO	***0.01***	***64***						
TMEM47	***0.01***	***63***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
GABARAPL2	***0.01***	***62***						
COL9A2	***0.01***	***61***						
RPS10	***0.01***	***60***						
SIM2 short	***0.01***	***59***						
MIR146A	***0.01***	***58***						
MEX3A	***0.01***	***57***						
ALAS1	***0.01***	***56***						
AMACR	***0.01***	***55***						
ITGBL1	***0.01***	***54***						
FDPS	***0.01***	***53***						
TWIST1	***0.01***	***52***						
HMBS	***0.01***	***51***						
KLK3 exons 1-2	***0.01***	***50***						
KLK4	***0.01***	***49***						
TFDP1	***0.01***	***48***						
VPS13A	***0.01***	***47***						
MEMO1	***0.01***	***46***						
ANPEP	***0.01***	***45***						
RAB17	***0.01***	***44***						
TRPM4	***0.01***	***43***						
HIST1H1C	***0.01***	***42***						
TBP	***0.01***	***41***						
RPL18A	***0.01***	***40***						
KLK2	***0.01***	***39***						
NKAIN1	***0.01***	***38***						
ZNF577	***0.01***	***37***						
BTG2	***0.01***	***36***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SChLAP1	***0.01***	***35***						
PCSK6	***0.00***	***34***						
CLU	***0.00***	***33***						
RPLP2	***0.00***	***32***						
ST6GALNAC1	***0.00***	***31***						
OR52A2	***0.00***	***30***						
SMIM1	***0.00***	***29***						
CDKN3	***0.00***	***28***						
MIC1	***0.00***	***27***						
ABCB9	***0.00***	***26***						
AR.ex4_8	***0.00***	***25***						
HIST1H1E	***0.00***	***24***						
DNAH5	***0.00***	***23***						
SMAP1 exons 7-8	***0.00***	***22***						
SYNM	***0.00***	***21***						
TERT	***0.00***	***20***						
PTN	***0.00***	***19***						
NLRP3	***0.00***	***18***						
CASKIN1	***0.00***	***17***						
BRAF	***0.00***	***16***						
Met	***0.00***	***15***						
MIATNB	***0.00***	***14***						
COL10A1	***0.00***	***13***						
HOXC4	***0.00***	***12***						
MDK	***0.00***	***11***						
SSTR1	***0.00***	***10***						
LBH	***0.00***	***9***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 85)*			*Transcripts identified by Mann Whitney U (n = 94)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
RP11_97O12.7	***0.00***	***8***						
STEAP2	***0.00***	***7***						
KLK3 exons 2-3	***0.00***	***5.5***						
SACM1L	***0.00***	***5.5***						
MARCH5	***0.00***	***4***						
CP	***0.00***	***3***						
B2M	***0.00***	***2***						
MMP11	***0.00***	***1***						

Supplementary Table 35 Random Forest results for comparing cancer samples with clinically benign samples in *KLK2* factorised cell data.

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ERG3' exons 6-7	***0.85***	***166***	***SLC12A1***	***0.98***	***24***	***HOXC6***	***1.20***	***33***
SLC12A1	***0.80***	***165***	***ERG3' exons 6-7***	***0.92***	***23***	***SLC12A1***	***0.76***	***32***
HOXC6	***0.69***	***164***	***HOXC6***	***0.92***	***22***	***ERG3' exons 6-7***	***0.74***	***31***
APOC1	***0.41***	***163***	***PCA3***	***0.63***	***21***	***PCA3***	***0.51***	***30***
CKAP2L	***0.38***	***162***	***HIST1H2BG***	***0.59***	***20***	***APOC1***	***0.50***	***29***
HIST1H2BG	***0.36***	***161***	***CADPS***	***0.46***	***19***	***CKAP2L***	***0.42***	***28***
CADPS	***0.27***	***160***	***TMPRSS2:ERG fusion***	***0.44***	***18***	***TMPRSS2:ERG fusion***	***0.39***	***27***
LASS1	***0.25***	***159***	***CKAP2L***	***0.43***	***17***	***CADPS***	***0.34***	***26***
SLC43A1	***0.24***	***158***	***NAALADL2***	***0.37***	***16***	***HPN***	***0.34***	***25***
NAALADL2	***0.23***	***157***	***SIM2 long***	***0.36***	***15***	***NAALADL2***	***0.31***	***24***
PCA3	***0.23***	***156***	***TDRD***	***0.35***	***14***	***TMEM86A***	***0.30***	***23***

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HPN	***0.23***	***155***	***HPN***	***0.35***	***13***	***UPK2***	***0.28***	***22***
SIM2 long	***0.21***	***154***	***GCNT1***	***0.34***	***12***	***TDRD***	***0.26***	***21***
TMPRSS2:ERG fusion	***0.19***	***153***	***TMEM86A***	***0.24***	***11***	***SIM2 long***	***0.25***	***20***
TMEM86A	***0.19***	***152***	***LASS1***	***0.24***	***10***	***SLC43A1***	***0.24***	***19***
ANKRD34B	***0.17***	***151***	***TMCC2***	***0.23***	***9***	***ST6GALNAC1***	***0.20***	***18***
AMACR	***0.17***	***150***	***CLIC2***	***0.21***	***8***	***LASS1***	***0.20***	***17***
TDRD	***0.17***	***149***	***MMP25***	***0.20***	***7***	***TMCC2***	***0.18***	***16***
GCNT1	***0.14***	***148***	***MFSD2A***	***0.16***	***6***	***ERG5'***	***0.18***	***15***
MFSD2A	***0.12***	***147***	***MCTP1***	***0.14***	***5***	***SERPINB5***	***0.18***	***14***
MCTP1	***0.10***	***146***	***OR52A2***	***0.14***	***4***	***CLIC2***	***0.17***	***13***
CAMKK2	***0.10***	***145***	***CAMKK2***	***0.11***	***3***	***SFRP4***	***0.17***	***12***
CLIC2	***0.09***	***144***	***CCDC88B***	***0.09***	***2***	***B4GALNT4***	***0.17***	***11***
TMCC2	***0.09***	***143***	***NLRP3***	***0.05***	***1***	***ANKRD34B***	***0.13***	***10***
B4GALNT4	***0.09***	***142***				***CAMKK2***	***0.10***	***9***
Timp4	***0.09***	***141***				***MCTP1***	***0.09***	***8***
UPK2	***0.09***	***140***				***MMP25***	***0.09***	***7***
ERG5'	***0.08***	***139***				***ISX***	***0.08***	***6***
DLX1	***0.08***	***138***				***FOLH1***	***0.08***	***5***
MMP25	***0.08***	***137***				***MFSD2A***	***0.07***	***4***
RNF157	***0.08***	***136***				***CCDC88B***	***0.05***	***3***
AURKA	***0.08***	***135***				***SULF2***	***0.03***	***2***
TERT	***0.08***	***134***				***MIR146A***	***0.03***	***1***
SFRP4	***0.07***	***133***						
CP	***0.06***	***132***						
NKAIN1	***0.06***	***131***						
CCDC88B	***0.05***	***130***						
OR52A2	***0.05***	***129***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
AR exons 4-8	***0.05***	***128***						
STOM	***0.04***	***127***						
ABCB9	***0.04***	***126***						
ERG3' exons 4-5	***0.04***	***125***						
SERPINB5	***0.04***	***124***						
SULF2	***0.04***	***123***						
MAPK8IP2	***0.03***	***122***						
AGR2	***0.03***	***121***						
ISX	***0.03***	***120***						
STEAP2	***0.03***	***119***						
CDKN3	***0.03***	***118***						
FOLH1	***0.03***	***117***						
MMP11	***0.03***	***116***						
TMEM45B	***0.03***	***115***						
SPINK1	***0.03***	***114***						
ITGBL1	***0.03***	***113***						
PPAP2A	***0.02***	***112***						
MEX3A	***0.02***	***111***						
IGFBP3	***0.02***	***110***						
PVT1	***0.02***	***109***						
P712P	***0.02***	***108***						
PPFIA2	***0.02***	***107***						
TRPM4	***0.02***	***106***						
MSMB	***0.02***	***105***						
SLC4A1.S	***0.02***	***104***						
PPP1R12B	***0.02***	***103***						
AMH	***0.02***	***102***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ST6GALNAC1	***0.02***	***101***						
DPP4	***0.02***	***100***						
SNORA20	***0.02***	***99***						
TMEM47	***0.02***	***98***						
VAX2	***0.02***	***97***						
HMBS	***0.02***	***96***						
VPS13A	***0.01***	***95***						
RPL23AP53	***0.01***	***94***						
EN2	***0.01***	***93***						
MKi67	***0.01***	***92***						
KLK4	***0.01***	***91***						
PALM3	***0.01***	***90***						
ALAS1	***0.01***	***89***						
RPL18A	***0.01***	***88***						
SEC61A1	***0.01***	***87***						
PTN	***0.01***	***86***						
MNX1	***0.01***	***85***						
TWIST1	***0.01***	***84***						
MGAT5B	***0.01***	***83***						
RPS11	***0.01***	***82***						
ZNF577	***0.01***	***81***						
PSTPIP1	***0.01***	***80***						
RIOK3	***0.01***	***79***						
KLK3 exons 2-3	***0.01***	***78***						
COL10A1	***0.01***	***77***						
OGT	***0.01***	***76***						
CASKIN1	***0.01***	***75***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
RPS10	***0.01***	***74***						
NLRP3	***0.01***	***73***						
CLU	***0.01***	***72***						
HIST1H1C	***0.01***	***71***						
SMIM1	***0.01***	***70***						
GJB1	***0.01***	***69***						
MIATNB	***0.01***	***68***						
CD10	***0.01***	***67***						
PDLIM5	***0.01***	***66***						
TBP	***0.009***	***65***						
MMP26	***0.009***	***64***						
CACNA1D	***0.009***	***63***						
SPON2	***0.009***	***62***						
MCM7	***0.009***	***61***						
MEMO1	***0.009***	***60***						
ACTR5	***0.008***	***59***						
RP11_97O12.7	***0.008***	***58***						
ITPR1	***0.008***	***57***						
TERF2IP	***0.008***	***56***						
STEAP4	***0.008***	***55***						
MAK	***0.008***	***54***						
SULT1A1	***0.007***	***53***						
NEAT1	***0.007***	***52***						
MYOF	***0.006***	***51***						
MIC1	***0.006***	***50***						
KLK3 exons 1-2	***0.006***	***49***						
HOXC4	***0.005***	***48***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SRSF3	***0.005***	***47***						
GAPDH	***0.005***	***46***						
MDK	***0.005***	***45***						
SACM1L	***0.005***	***44***						
HIST1H1E	***0.005***	***43***						
GABARAPL2	***0.005***	***42***						
MIR4435_1HG	***0.005***	***41***						
FDPS	***0.005***	***40***						
COL9A2	***0.004***	***39***						
DNAH5	***0.004***	***38***						
LBH	***0.004***	***37***						
RAB17	***0.003***	***36***						
SChLAP1	***0.003***	***35***						
BRAF	***0.003***	***34***						
TFDP1	***0.003***	***33***						
IFT57	***0.003***	***32***						
RPLP2	***0.003***	***31***						
HIST3H2A	***0.003***	***30***						
SIM2 short	***0.002***	***29***						
ANPEP	***0.002***	***28***						
AATF	***0.002***	***27***						
BTG2	***0.002***	***26***						
MXI1	***0.002***	***25***						
MED4	***0.002***	***24***						
IMPDH2	***0.002***	***23***						
SSTR1	***0.002***	***22***						
MIR146A	***0.002***	***21***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 24)*			*Transcripts identified by Mann Whitney U (n = 33)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
Mar-05	***0.002***	***20***						
SIRT1	***0.002***	***19***						
AR exon 9	***0.002***	***18***						
PECI	***0.002***	***17***						
SYNM	***$1.15x10^{-17}$***	***16***						
PTPRC	***$1.07x10^{-17}$***	***15***						
GOLM1	***$8.99x10^{-18}$***	***14***						
ARHGEF25	***$7.55x10^{-18}$***	***13***						
CDC37L1	***$7.11x10^{-18}$***	***12***						
CDC20	***$5.77x10^{-18}$***	***11***						
SSPO	***$4.00x10^{-18}$***	***10***						
SMAP1 exons 7-8	***$3.55x10^{-18}$***	***9***						
EIF2D	***$2.66x10^{-18}$***	***8***						
SNCA	***$1.78x10^{-18}$***	*7*						
B2M	***$4.44x10^{-19}$***	***6***						
CAMK2N2	***0***	***3***						
HIST1H2BF	***0***	***3***						
HPRT	***0***	***3***						
Met	***0***	***3***						
PCSK6	***0***	***3***						

Supplementary Table 36 Random Forest results for CB vs Cancer in the *RPLP2* and *TWIST1* normalised data.

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HOXC6	***0.76***	***167***	***ERG3' exons 6-7***	***1.05***	***87***	***ERG3' exons 6-7***	***0.76***	***65***
SPINK1	***0.73***	***166***	***APOC1***	***0.81***	***86***	***CCDC88B***	***0.53***	***64***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
NAALADL2	***0.67***	***165***	***SPINK1***	***0.65***	***85***	***CADPS***	***0.51***	***63***
UPK2	***0.61***	***164***	***CCDC88B***	***0.57***	***84***	***B4GALNT4***	***0.29***	***62***
CADPS	***0.43***	***163***	***CADPS***	***0.48***	***83***	***HOXC6***	***0.27***	***61***
ERG3' exons 6-7	***0.34***	***162***	***CKAP2L***	***0.43***	***82***	***TMPRSS2:ERG fusion***	***0.24***	***60***
HPN	***0.32***	***161***	***GAPDH***	***0.37***	***81***	***RIOK3***	***0.19***	***59***
ISX	***0.29***	***160***	***CAMKK2***	***0.36***	***80***	***SIM2 long***	***0.19***	***58***
CP	***0.25***	***159***	***AURKA***	***0.22***	***79***	***MIR4435_1HG***	***0.18***	***57***
TMPRSS2:ERG fusion	***0.24***	***158***	***HPN***	***0.22***	***78***	***NEAT1***	***0.18***	***56***
PCA3	***0.20***	***157***	***UPK2***	***0.21***	***77***	***AATF***	***0.15***	***55***
TDRD	***0.19***	***156***	***AATF***	***0.21***	***76***	***SIRT1***	***0.13***	***54***
B4GALNT4	***0.18***	***155***	***B4GALNT4***	***0.20***	***75***	***APOC1***	***0.12***	***53***
CKAP2L	***0.16***	***154***	***IGFBP3***	***0.18***	***74***	***HPRT***	***0.11***	***52***
ST6GALNAC1	***0.15***	***153***	***ISX***	***0.18***	***73***	***MMP25***	***0.11***	***51***
SFRP4	***0.14***	***152***	***TDRD***	***0.17***	***72***	***TDRD***	***0.10***	***50***
GCNT1	***0.14***	***151***	***PCA3***	***0.17***	***71***	***MCTP1***	***0.09***	***49***
Timp4	***0.13***	***150***	***CD10***	***0.16***	***70***	***TMEM86A***	***0.09***	***48***
APOC1	***0.12***	***149***	***CLIC2***	***0.13***	***69***	***CLIC2***	***0.09***	***47***
SLC43A1	***0.12***	***148***	***NAALADL2***	***0.11***	***68***	***SFRP4***	***0.09***	***46***
CLIC2	***0.11***	***147***	***ERG3' exons 4-5***	***0.10***	***67***	***ERG5'***	***0.09***	***45***
TMCC2	***0.11***	***146***	***SLC4A1 S***	***0.10***	***66***	***MEX3A***	***0.09***	***44***
EN2	***0.09***	***145***	***CDC37L1***	***0.09***	***65***	***SLC43A1***	***0.08***	***43***
AR exon 9	***0.08***	***144***	***CACNA1D***	***0.09***	***64***	***MFSD2A***	***0.08***	***42***
RNF157	***0.08***	***143***	***HIST1H2BG***	***0.09***	***63***	***SEC61A1***	***0.07***	***41***
ANKRD34B	***0.08***	***142***	***SNORA20***	***0.08***	***62***	***MAK***	***0.07***	***40***
CLU	***0.08***	***141***	***ACTR5***	***0.08***	***61***	***HPN***	***0.07***	***39***
MMP25	***0.07***	***140***	***TERF2IP***	***0.07***	***60***	***SULF2***	***0.07***	***38***
SIM2 long	***0.07***	***139***	***TMCC2***	***0.07***	***59***	***GCNT1***	***0.06***	***37***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
TMEM86A	***0.07***	***138***	***CDC20***	***0.06***	***58***	***EN2***	***0.06***	***36***
HMBS	***0.06***	***137***	***DPP4***	***0.06***	***57***	***SPINK1***	***0.06***	***35***
ERG5'	***0.06***	***136***	***TFDP1***	***0.06***	***56***	***PTPRC***	***0.06***	***34***
DNAH5	***0.06***	***135***	***AR exon 9***	***0.06***	***55***	***ANKRD34B***	***0.05***	***33***
MSMB	***0.05***	***134***	***MYOF***	***0.05***	***54***	***IGFBP3***	***0.05***	***32***
MFSD2A	***0.05***	***133***	***RNF157***	***0.05***	***53***	***UPK2***	***0.05***	***31***
SERPINB5	***0.05***	***132***	***AMH***	***0.05***	***52***	***AURKA***	***0.05***	***30***
P712P	***0.05***	***131***	***GABARAPL2***	***0.05***	***51***	***SNCA***	***0.05***	***29***
CAMKK2	***0.05***	***130***	***FOLH1***	***0.05***	***50***	***CACNA1D***	***0.05***	***28***
TMEM47	***0.04***	***129***	***ANPEP***	***0.05***	***49***	***LASS1***	***0.05***	***27***
PPFIA2	***0.04***	***128***	***EN2***	***0.05***	***48***	***GAPDH***	***0.05***	***26***
ITGBL1	***0.04***	***127***	***ST6GALNAC1***	***0.04***	***47***	***CAMKK2***	***0.04***	***25***
MNX1	***0.04***	***126***	***ANKRD34B***	***0.04***	***46***	***B2M***	***0.04***	***24***
RIOK3	***0.04***	***125***	***AMACR***	***0.04***	***45***	***ERG3' exons 4-5***	***0.04***	***23***
GJB1	***0.04***	***124***	***ERG5'***	***0.03***	***44***	***SLC12A1***	***0.04***	***22***
TWIST1	***0.04***	***123***	***CP***	***0.03***	***43***	***ITPR1***	***0.04***	***21***
SRSF3	***0.03***	***122***	***EIF2D***	***0.03***	***42***	***MAPK8IP2***	***0.03***	***20***
AGR2	***0.03***	***121***	***MCM7***	***0.03***	***41***	***SRSF3***	***0.03***	***19***
PPAP2A	***0.03***	***120***	***Met***	***0.03***	***40***	***ISX***	***0.03***	***18***
PPP1R12B	***0.03***	***119***	***DNAH5***	***0.02***	***39***	***FOLH1***	***0.03***	***17***
STEAP4	***0.03***	***118***	***SIM2 short***	***0.02***	***38***	***EIF2D***	***0.03***	***16***
MYOF	***0.03***	***117***	***SMAP1 exons 7-8***	***0.02***	***37***	***CDC20***	***0.03***	***15***
STEAP2	***0.03***	***116***	***AGR2***	***0.02***	***36***	***GABARAPL2***	***0.02***	***14***
IGFBP3	***0.03***	***115***	***NLRP3***	***0.02***	***35***	***MXI1***	***0.02***	***13***
CCDC88B	***0.03***	***114***	***KLK2***	***0.02***	***34***	***AMH***	***0.02***	***12***
SLC4A1.S	***0.03***	***113***	***AR exons 4-8***	***0.02***	***33***	***TBP***	***0.01***	***11***
SULF2	***0.03***	***112***	***MAK***	***0.02***	***32***	***PDLIM5***	***0.01***	***10***
DLX1	***0.03***	***111***	***TMEM47***	***0.02***	***31***	***ARHGEF25***	***0.01***	***9***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SEC61A1	***0.03***	***110***	***CDKN3***	***0.02***	***30***	***ACTR5***	***0.01***	***8***
PECI	***0.03***	***109***	***RPS11***	***0.02***	***29***	***NLRP3***	***0.01***	***7***
HIST1H2BG	***0.03***	***108***	***PPAP2A***	***0.02***	***28***	***CD10***	***0.01***	***6***
LASS1	***0.03***	***107***	***PALM3***	***0.02***	***27***	***TERF2IP***	***0.005***	***5***
NLRP3	***0.02***	***106***	***RP11_97O12.7***	***0.02***	***26***	***ANPEP***	***0.004***	***4***
SULT1A1	***0.02***	***105***	***CAMK2N2***	***0.02***	***25***	***MIC1***	***0.004***	***3***
ACTR5	***0.02***	***104***	***PECI***	***0.02***	***24***	***CASKIN1***	***0.003***	***2***
MDK	***0.02***	***103***	***FDPS***	***0.02***	***23***	***SACM1L***	***$5.46x10^{-17}$***	***1***
SLC12A1	***0.02***	***102***	***ARHGEF25***	***0.02***	***22***			
TMEM45B	***0.02***	***101***	***HOXC4***	***0.02***	***21***			
MAK	***0.02***	***100***	***MARCH5***	***0.01***	***20***			
SIRT1	***0.02***	***99***	***TBP***	***0.01***	***19***			
MAPK8IP2	***0.02***	***98***	***ABCB9***	***0.01***	***18***			
MCTP1	***0.02***	***97***	***B2M***	***0.01***	***17***			
AATF	***0.02***	***96***	***ALAS1***	***0.01***	***16***			
RAB17	***0.02***	***95***	***DLX1***	***0.01***	***15***			
MEMO1	***0.02***	***94***	***BTG2***	***0.01***	***14***			
PALM3	***0.02***	***93***	***PCSK6***	***0.01***	***13***			
TRPM4	***0.02***	***92***	***SSTR1***	***0.01***	***12***			
SMIM1	***0.02***	***91***	***STEAP2***	***0.01***	***11***			
ABCB9	***0.02***	***90***	***CLU***	***0.01***	***10***			
MIR146A	***0.02***	***89***	***LBH***	***0.01***	***9***			
IMPDH2	***0.02***	***88***	***MIATNB***	***0.01***	***8***			
MGAT5B	***0.02***	***87***	***COL10A1***	***0.01***	***7***			
DPP4	***0.02***	***86***	***COL9A2***	***0.01***	***6***			
MIR4435_1HG	***0.02***	***85***	***OGT***	***0.01***	***5***			
CACNA1D	***0.01***	***84***	***MEX3A***	***0.01***	***4***			
CDC20	***0.01***	***83***	***GOLM1***	***0.01***	***3***			

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
RPS10	***0.01***	***82***	***CASKIN1***	***0.004***	***2***			
CASKIN1	***0.01***	***81***	***BRAF***	***0.002***	***1***			
Met	***0.01***	***80***						
SPON2	***0.01***	***79***						
TERF2IP	***0.01***	***78***						
HIST1H1E	***0.01***	***77***						
GAPDH	***0.01***	***76***						
AURKA	***0.01***	***75***						
NKAIN1	***0.01***	***74***						
PVT1	***0.01***	***73***						
STOM	***0.01***	***72***						
VPS13A	***0.01***	***71***						
AMH	***0.01***	***70***						
COL9A2	***0.01***	***69***						
AMACR	***0.01***	***68***						
SIM2 short	***0.01***	***67***						
CD10	***0.01***	***66***						
FDPS	***0.01***	***65***						
MMP26	***0.01***	***64***						
MXI1	***0.01***	***63***						
ARHGEF25	***0.01***	***62***						
IFT57	***0.01***	***61***						
KLK2	***0.01***	***60***						
HOXC4	***0.01***	***59***						
KLK4	***0.01***	***58***						
MED4	***0.01***	***57***						
RPLP2	***0.01***	***56***						
CDKN3	***0.01***	***55***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
CDC37L1	***0.01***	***54***						
MMP11	***0.01***	***53***						
AR exons 4-8	***0.01***	***52***						
RPS11	***0.01***	***51***						
SMAP1 exons 7-8	***0.01***	***50***						
FOLH1	***0.01***	***49***						
GOLM1	***0.01***	***48***						
PTN	***0.01***	***47***						
HIST3H2A	***0.01***	***46***						
ERG3' exons 4-5	***0.01***	***45***						
TERT	***0.01***	***44***						
MEX3A	***0.01***	***43***						
SYNM	***0.01***	***42***						
B2M	***0.01***	***41***						
SChLAP1	***0.01***	***40***						
RP11_97O12.7	***0.01***	***39***						
RPL18A	***0.01***	***38***						
GABARAPL2	***0.01***	***37***						
HIST1H1C	***0.01***	***36***						
BRAF	***0.01***	***35***						
SNORA20	***0.01***	***34***						
OR52A2	***0.01***	***33***						
ANPEP	***0.01***	***32***						
PSTPIP1	***0.01***	***31***						
RPL23AP53	***0.01***	***30***						
COL10A1	***0.01***	***29***						
SSTR1	***0.01***	***28***						
LBH	***0.005***	***27***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ITPR1	***0.005***	***26***						
TFDP1	***0.005***	***25***						
CAMK2N2	***0.004***	***24***						
TBP	***0.004***	***23***						
PTPRC	***0.004***	***22***						
ZNF577	***0.004***	***21***						
MARCH5	***0.003***	***20***						
ALAS1	***0.003***	***19***						
HPRT	***0.003***	***18***						
OGT	***0.003***	***17***						
KLK3 exons 1-2	***0.002***	***16***						
MCM7	***0.002***	***15***						
VAX2	***0.002***	***14***						
SSPO	***0.002***	***13***						
BTG2	***0.002***	***12***						
MIC1	***0.002***	***11***						
NEAT1	***0.002***	***10***						
MKi67	***0.002***	***9***						
MIATNB	***0.002***	***8***						
EIF2D	***0.002***	***7***						
SACM1L	***$1.60x10^{-17}$***	***6***						
SNCA	***1.08x10-^{17}***	***5***						
PCSK6	***$7.99x10^{-18}$***	***4***						
HIST1H2BF	***$1.78x10^{-18}$***	***3***						
KLK3 exons 2-3	***$4.44x10^{-19}$***	***2***						
PDLIM5	***0***	***1***						

6.19 High Risk Vs CB

Supplementary Table 37 Transcripts that have significant differential expression (using glm and MWU tests) between clinically benign and high-risk cancer samples in the baseline normalized NanoString data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***0.0002***	***0.0299***	***0.004***	***0.6711***	***2***
ERG3' exons 6-7	***$6.21x10^{-06}$***	***0.001***	***0.0371***	***0.9942***	***1.6***
TDRD	***0.0011***	***0.1558***	***0.0333***	***0.9942***	***1.5***
TMPRSS2:ERG	***0.0004***	***0.0668***	***0.0386***	***0.9942***	***1.3***
B4GALNT4	***$2.88x10^{-05}$***	***0.0048***	***0.0409***	***0.9942***	***1.2***
SLC43A1	***0.002***	***0.2897***	***0.0117***	***0.9942***	***1.2***
CADPS	***$6.70x10^{-05}$***	***0.011***	***0.02***	***0.9942***	***1.1***
CLIC2	***0.0002***	***0.0386***	***0.0087***	***0.9942***	***1***
HPN	***0.0008***	***0.1258***	***0.0092***	***0.9942***	***0.9***
LASS1	***0.0011***	***0.1558***	***0.0103***	***0.9942***	***0.9***
MAPK8IP2	***0.0148***	***1***	***0.0336***	***0.9942***	***0.9***
SFRP4	***0.0013***	***0.1919***	***0.0155***	***0.9942***	***0.9***
CKAP2L			***0.0392***	***0.9942***	***0.9***
CDKN3			***0.0326***	***0.9942***	***0.9***
ANKRD34B	***0.0054***	***0.7002***	***0.0368***	***0.9942***	***0.8***
ERG3' exons 4-5	***0.0037***	***0.5042***	***0.0434***	***0.9942***	***0.8***
APOC1	***0.0002***	***0.0386***	***0.0055***	***0.9103***	***0.7***
ERG5'	***0.0077***	***0.9678***			***0.7***
MMP25	***0.0045***	***0.5959***	***0.0162***	***0.9942***	***0.7***
AMH	***0.0108***	***1***			***0.6***
CCDC88B	***0.0007***	***0.1014***	***0.0104***	***0.9942***	***0.6***
FOLH1	***0.0108***	***1***	***0.027***	***0.9942***	***0.6***
ISX	***0.0026***	***0.3638***	***0.0234***	***0.9942***	***0.6***
MCTP1	***0.0001***	***0.0228***	***0.0098***	***0.9942***	***0.6***
SIM2 long	***0.0002***	***0.0386***	***0.0124***	***0.9942***	***0.6***
SRSF3	***0.0234***	***1***	***0.0225***	***0.9942***	***0.6***
ANPEP	***0.0234***	***1***	***0.0324***	***0.9942***	***0.5***
GCNT1	***0.0054***	***0.7002***	***0.0209***	***0.9942***	***0.5***
MFSD2A	***0.0017***	***0.2364***	***0.0216***	***0.9942***	***0.5***
NLRP3	***0.0202***	***1***	***0.0277***	***0.9942***	***0.5***
SLC12A1	***0.0064***	***0.8317***	***0.0267***	***0.9942***	***0.5***
SULF2	***0.0007***	***0.1014***	***0.0141***	***0.9942***	***0.5***
TMEM86A	***0.0005***	***0.0823***	***0.022***	***0.9942***	***0.5***
AATF	***0.0007***	***0.1014***	***0.0108***	***0.9942***	***0.4***
CAMKK2	***0.0007***	***0.1014***	***0.0128***	***0.9942***	***0.4***
CDC20	***0.0202***	***1***	***0.0343***	***0.9942***	***0.4***
EN2	***0.0091***	***1***			***0.4***

ARHGEF25	***0.0464***	***1***			***0.3***
AURKA	***0.0234***	***1***			***0.3***
CD10	***0.0127***	***1***	***0.0481***	***0.9942***	***0.3***
HPRT	***0.0025***	***0.3503***	***0.0255***	***0.9942***	***0.3***
MEX3A	***0.0045***	***0.5959***	***0.0496***	***0.9942***	***0.3***
MIC1	***0.031***	***1***			***0.3***
PTPRC	***0.0013***	***0.1919***	***0.0116***	***0.9942***	***0.3***
RIOK3	***$5.63x10^{-05}$***	***0.0093***	***0.0237***	***0.9942***	***0.3***
SEC61A1	***0.0031***	***0.4184***	***0.0341***	***0.9942***	***0.3***
SIRT1	***0.0025***	***0.3503***	***0.0206***	***0.9942***	***0.3***
SNCA	***0.027***	***1***			***0.3***
ACTR5	***0.0127***	***1***			***0.2***
CACNA1D	***0.0202***	***1***	***0.0356***	***0.9942***	***0.2***
CASKIN1	***0.0464***	***1***			***0.2***
EIF2D	***0.0077***	***0.9678***	***0.0297***	***0.9942***	***0.2***
GABARAPL2	***0.0127***	***1***	***0.036***	***0.9942***	***0.2***
ITPR1	***0.0008***	***0.1258***	***0.0091***	***0.9942***	***0.2***
MAK	***0.0148***	***1***			***0.2***
MIR4435_1HG	***0.0007***	***0.1014***	***0.0169***	***0.9942***	***0.2***
MXI1	***0.0054***	***0.7002***	***0.0206***	***0.9942***	***0.2***
NEAT1	***0.0002***	***0.0299***	***0.0039***	***0.6541***	***0.2***
PDLIM5	***0.0464***	***1***			***0.2***
TBP	***0.0202***	***1***			***0.2***
B2M	***0.0008***	***0.1258***	***0.0203***	***0.9942***	***0.1***
GAPDH	***0.0031***	***0.4184***	***0.0111***	***0.9942***	***0.1***
SACM1L	***0.0234***	***1***			***0.1***
TERF2IP	***0.0045***	***0.5959***	***0.0197***	***0.9942***	***0.1***
IGFBP3	***0.0464***	***1***			***-0.3***
SPINK1	***0.0077***	***0.9678***			***-0.4***
UPK2	***0.0202***	***1***			***-0.8***

Supplementary Table 38 Transcripts that have significant differential expression (using glm and MWU tests) between clinically benign and high-risk cancer samples in the *KLK2* ratio NanoString data.

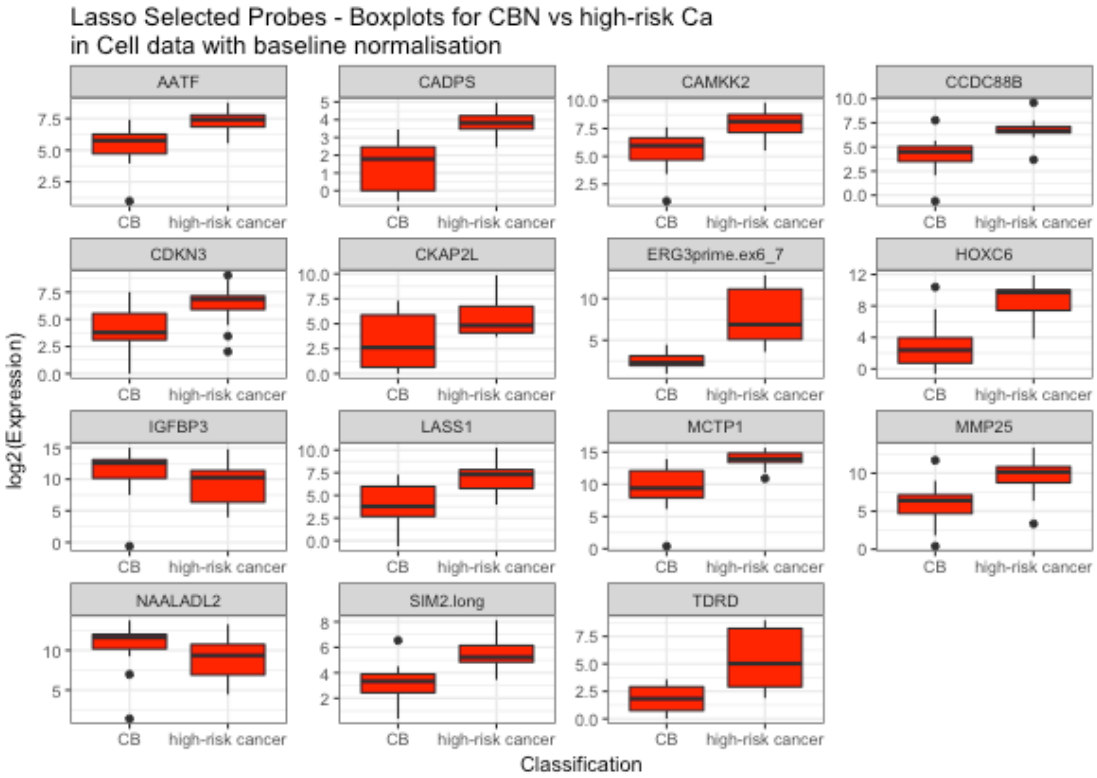
	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
TMPRSS2:ERG	***0.004***	***0.68***	***0.028***	***1.000***	***0.25***
ERG 3' exons 6-7	***0.000***	***0.07***	***0.008***	***1.000***	***0.25***
HOXC6	***4.28E-05***	***0.01***			***0.25***
TDRD	***0.001***	***0.09***	***0.017***	***1.000***	***0.24***
SLC43A1	***0.002***	***0.27***	***0.022***	***1.000***	***0.21***
CADPS	***0.007***	***1***			***0.18***
B4GALNT4	***0.002***	***0.33***	***0.035***	***1.000***	***0.17***
ERG 5'	***0.027***	***1***			***0.16***
SLC12A1	***0.013***	***1***			***0.15***
ERG 3' exons 4-5	***0.046***	***1***	***0.050***	***1.000***	***0.14***

LASS1	***0.015***	***1***	***0.029***	***1.000***	***0.13***
CLIC2	***0.004***	***0.59***	***0.027***	***1.000***	***0.13***
HPN	***0.003***	***0.49***	***0.032***	***1.000***	***0.11***
ISX	***0.017***	***1***			***0.11***
APOC1	***0.009***	***1***			***0.09***
TMEM86A	***0.017***	***1***			***0.09***
PCA3	***0.003***	***0.49***	***0.015***	***1.000***	***0.08***
CCDC88B	***0.027***	***1***			***0.08***
SFRP4	***0.031***	***1***			***0.08***
MCTP1	***0.020***	***1***			***0.08***
SIM2 long	***0.001***	***0.14***	***0.028***	***1.000***	***0.08***
FOLH1	***0.008***	***1***			***0.07***
CAMKK2	***0.036***	***1***			***0.05***
SEC61A1	***0.046***	***1***			***0.05***
GCNT1	***0.027***	***1***	***0.043***	***1.000***	***0.04***

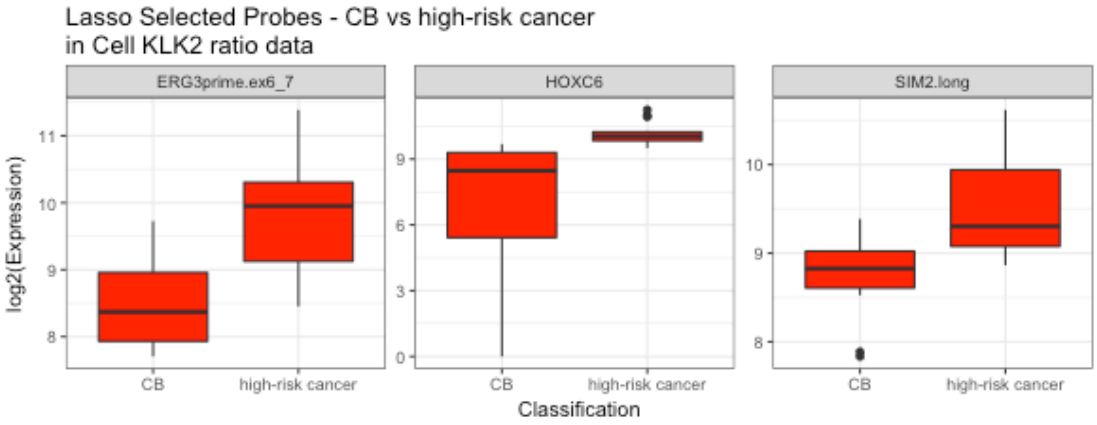
Supplementary Table 39 Transcripts that have significant differential expression (using glm and MWU tests) between clinically benign and high-risk cancer samples in the HK normalised NanoString data.

	MWU		*glm*		
Transcript	***p-value***	***Adjusted p-value***	***p-value***	***Adjusted p-value***	***Log_2(FC)***
HOXC6	***0.0005***	***0.0882***	***0.0059***	***0.9765***	***1.6***
ERG3' exons 6-7	***0.0013***	***0.2186***	***0.0266***	***0.9765***	***1.4***
TDRD	***0.0031***	***0.4948***	***0.0272***	***0.9765***	***1.1***
TMPRSS2:ERG fusion	***0.0094***	***1***	***0.033***	***0.9765***	***1.1***
ST6GALNAC1	***0.0037***	***0.5969***	***0.0168***	***0.9765***	***-1***
SLC43A1	***0.0013***	***0.2186***	***0.0197***	***0.9765***	***0.9***
B4GALNT4	***0.0202***	***1***			***0.8***
HPN	***0.0077***	***1***	***0.0314***	***0.9765***	***0.8***
CADPS	***0.0145***	***1***	***0.0326***	***0.9765***	***0.7***
CCDC88B	***0.031***	***1***	***0.0482***	***0.9765***	***0.7***
SPINK1	***0.0007***	***0.1115***	***0.0092***	***0.9765***	***-0.7***
UPK2	***0.0054***	***0.8564***	***0.0237***	***0.9765***	***-0.7***
CLIC2	***0.0108***	***1***	***0.0278***	***0.9765***	***0.6***
LASS1	***0.0202***	***1***	***0.0451***	***0.9765***	***0.6***
GJB1	***0.0108***	***1***	***0.0197***	***0.9765***	***-0.6***
IGFBP3	***0.0464***	***1***			***-0.6***
NAALADL2	***0.0031***	***0.4948***	***0.0133***	***0.9765***	***-0.6***
SERPINB5	***0.0054***	***0.8564***	***0.0199***	***0.9765***	***-0.6***
ISX	***0.0288***	***1***			***0.5***
MMP25	***0.0407***	***1***			***0.5***
GCNT1	***0.0356***	***1***	***0.0446***	***0.9765***	***0.4***
MCTP1	***0.0464***	***1***			***0.4***
SIM2 long	***0.0234***	***1***	***0.0317***	***0.9765***	***0.4***
PALM3	***0.0108***	***1***	***0.0394***	***0.9765***	***-0.4***

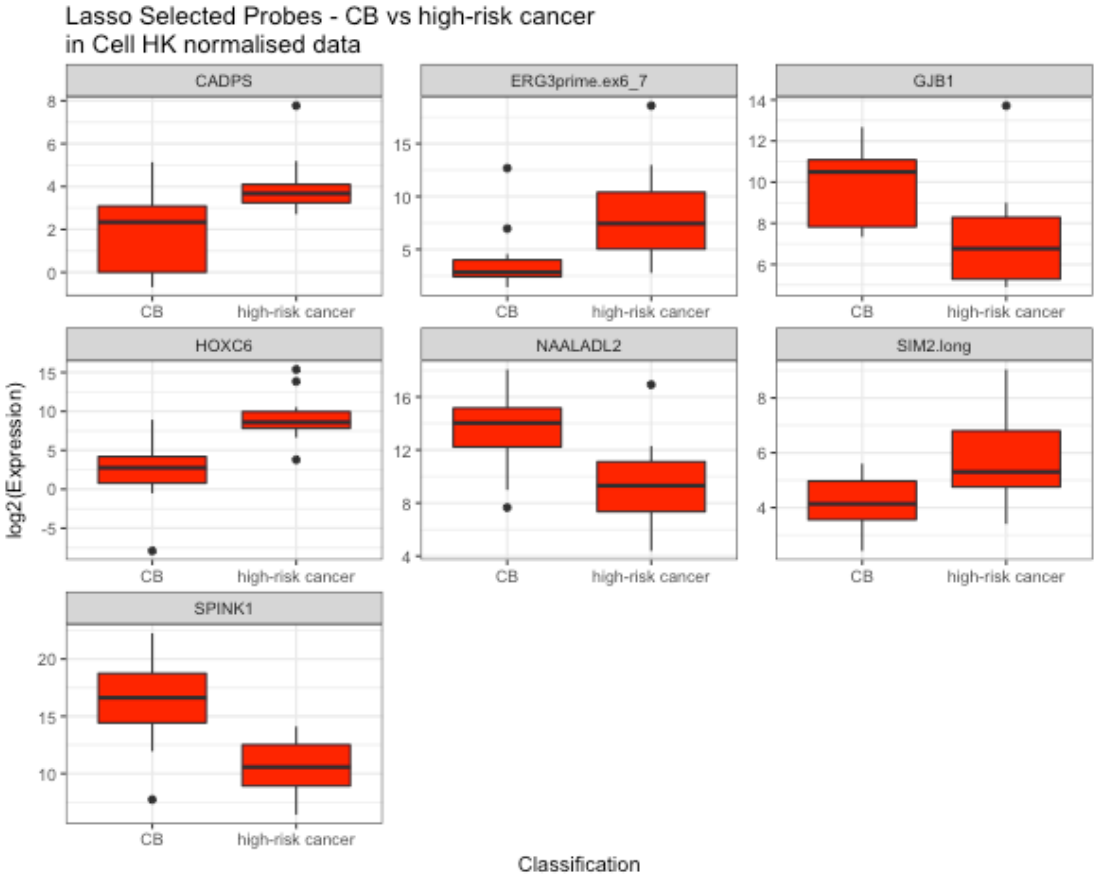
APOC1	***0.0173***	***1***			***0.3***
MSMB	***0.0356***	***1***			***-0.3***
PPAP2A	***0.0077***	***1***			***-0.3***
RAB17	***0.0356***	***1***			***-0.3***
RPS10	***0.0464***	***1***			***-0.3***
SPON2	***0.0464***	***1***			***-0.3***
STEAP2	***0.0173***	***1***			***-0.3***
VAX2	***0.0109***	***1***			***-0.3***
TMEM86A	***0.0464***	***1***			***0.2***
IFT57	***0.027***	***1***			***-0.2***
PTN	***0.031***	***1***	***0.0491***	***0.9765***	***-0.2***



Supplementary Figure 16 Boxplots showing all of the Lasso selected probes in CB Vs. high-risk cancer models in baseline normalised cell data.



Supplementary Figure 17 Boxplots showing all of the Lasso selected probes in CB Vs. high-risk cancer models in *KLK2* ratio cell data.



Supplementary Figure 18 Boxplots showing all of the Lasso selected probes in CB Vs. high-risk cancer models in HK normalised cell data.

Supplementary Table 40 Random Forest results for HR-Ca vs CBN.

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SPINK1	***0.70***	***167***	***CADPS***	***0.67***	***51***	***ERG3' exons 6-7***	***0.76***	***65***
CADPS	***0.69***	***166***	***ERG3' exons 6-7***	***0.66***	***50***	***CCDC88B***	***0.53***	***64***
NAALADL2	***0.67***	***165***	***CCDC88B***	***0.51***	***49***	***CADPS***	***0.51***	***63***
HOXC6	***0.64***	***164***	***RIOK3***	***0.34***	***48***	***B4GALNT4***	***0.29***	***62***
HPN	***0.57***	***163***	***HOXC6***	***0.27***	***47***	***HOXC6***	***0.27***	***61***
ERG3' exons 6-7	***0.49***	***162***	***B4GALNT4***	***0.26***	***46***	***TMPRSS2:ERG fusion***	***0.24***	***60***
UPK2	***0.47***	***161***	***TMPRSS2:ERG fusion***	***0.26***	***45***	***RIOK3***	***0.19***	***59***
TMPRSS2:ERG fusion	***0.43***	***160***	***SIM2 long***	***0.25***	***44***	***SIM2 long***	***0.19***	***58***
CP	***0.42***	***159***	***MIR4435_1HG***	***0.22***	***43***	***MIR4435_1HG***	***0.18***	***57***
TDRD	***0.39***	***158***	***SFRP4***	***0.21***	***42***	***NEAT1***	***0.18***	***56***
MNX1	***0.39***	***157***	***HPRT***	***0.17***	***41***	***AATF***	***0.15***	***55***
PCA3	***0.36***	***156***	***APOC1***	***0.16***	***40***	***SIRT1***	***0.13***	***54***
ST6GALNAC1	***0.34***	***155***	***AATF***	***0.16***	***39***	***APOC1***	***0.12***	***53***
TMCC2	***0.34***	***154***	***NEAT1***	***0.15***	***38***	***HPRT***	***0.11***	***52***
SIM2 long	***0.34***	***153***	***TMEM86A***	***0.13***	***37***	***MMP25***	***0.11***	***51***
APOC1	***0.34***	***152***	***MEX3A***	***0.13***	***36***	***TDRD***	***0.10***	***50***
CKAP2L	***0.31***	***151***	***SLC43A1***	***0.11***	***35***	***MCTP1***	***0.09***	***49***
PPFIA2	***0.29***	***150***	***HPN***	***0.11***	***34***	***TMEM86A***	***0.09***	***48***
SERPINB5	***0.27***	***149***	***SEC61A1***	***0.10***	***33***	***CLIC2***	***0.09***	***47***
TMEM45B	***0.27***	***148***	***SIRT1***	***0.10***	***32***	***SFRP4***	***0.09***	***46***
AGR2	***0.26***	***147***	***CLIC2***	***0.09***	***31***	***ERG5'***	***0.09***	***45***
EN2	***0.25***	***146***	***GCNT1***	***0.09***	***30***	***MEX3A***	***0.09***	***44***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ISX	***0.25***	***145***	***MCTP1***	***0.08***	***29***	***SLC43A1***	***0.08***	***43***
GCNT1	***0.25***	***144***	***MFSD2A***	***0.08***	***28***	***MFSD2A***	***0.08***	***42***
MFSD2A	***0.24***	***143***	***TDRD***	***0.08***	***27***	***SEC61A1***	***0.07***	***41***
DNAH5	***0.24***	***142***	***SULF2***	***0.06***	***26***	***MAK***	***0.07***	***40***
SFRP4	***0.24***	***141***	***PTPRC***	***0.06***	***25***	***HPN***	***0.07***	***39***
SLC43A1	***0.24***	***140***	***GAPDH***	***0.06***	***24***	***SULF2***	***0.07***	***38***
B4GALNT4	***0.24***	***139***	***ISX***	***0.06***	***23***	***GCNT1***	***0.06***	***37***
PTN	***0.24***	***138***	***ANKRD34B***	***0.05***	***22***	***EN2***	***0.06***	***36***
GJB1	***0.23***	***137***	***MMP25***	***0.05***	***21***	***SPINK1***	***0.06***	***35***
MMP25	***0.23***	***136***	***ITPR1***	***0.04***	***20***	***PTPRC***	***0.06***	***34***
Timp4	***0.22***	***135***	***CACNA1D***	***0.04***	***19***	***ANKRD34B***	***0.05***	***33***
RIOK3	***0.21***	***134***	***MXI1***	***0.04***	***18***	***IGFBP3***	***0.05***	***32***
MDK	***0.21***	***133***	***SRSF3***	***0.04***	***17***	***UPK2***	***0.05***	***31***
CLU	***0.20***	***132***	***LASS1***	***0.03***	***16***	***AURKA***	***0.05***	***30***
LASS1	***0.20***	***131***	***B2M***	***0.03***	***15***	***SNCA***	***0.05***	***29***
MMP11	***0.20***	***130***	***SLC12A1***	***0.03***	***14***	***CACNA1D***	***0.05***	***28***
ERG3' exons 4-5	***0.19***	***129***	***GABARAPL2***	***0.03***	***13***	***LASS1***	***0.05***	***27***
VAX2	***0.19***	***128***	***ERG3' exons 4-5***	***0.03***	***12***	***GAPDH***	***0.05***	***26***
SPON2	***0.18***	***127***	***EIF2D***	***0.03***	***11***	***CAMKK2***	***0.04***	***25***
PPAP2A	***0.18***	***126***	***MAPK8IP2***	***0.02***	***10***	***B2M***	***0.04***	***24***
TMEM47	***0.17***	***125***	***FOLH1***	***0.02***	***9***	***ERG3' exons 4-5***	***0.04***	***23***
CLIC2	***0.17***	***124***	***CAMKK2***	***0.02***	***8***	***SLC12A1***	***0.04***	***22***
SLC12A1	***0.17***	***123***	***ANPEP***	***0.02***	***7***	***ITPR1***	***0.04***	***21***
COL9A2	***0.17***	***122***	***CDC20***	***0.01***	***6***	***MAPK8IP2***	***0.03***	***20***
ANKRD34B	***0.17***	***121***	***CKAP2L***	***0.01***	***5***	***SRSF3***	***0.03***	***19***
TWIST1	***0.17***	***120***	***CDKN3***	***0.01***	***4***	***ISX***	***0.03***	***18***
SSPO	***0.17***	***119***	***TERF2IP***	***0.01***	***3***	***FOLH1***	***0.03***	***17***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
MYOF	***0.17***	***118***	***CD10***	***0.01***	***2***	***EIF2D***	***0.03***	***16***
CCDC88B	***0.16***	***117***	***NLRP3***	***0.01***	***1***	***CDC20***	***0.03***	***15***
SIM2 short	***0.16***	***116***				***GABARAPL2***	***0.02***	***14***
DLX1	***0.16***	***115***				***MXI1***	***0.02***	***13***
CAMKK2	***0.16***	***114***				***AMH***	***0.02***	***12***
IGFBP3	***0.15***	***113***				***TBP***	***0.01***	***11***
IFT57	***0.15***	***112***				***PDLIM5***	***0.01***	***10***
MMP26	***0.15***	***111***				***ARHGEF25***	***0.01***	***9***
SNORA20	***0.15***	***110***				***ACTR5***	***0.01***	***8***
RNF157	***0.14***	***109***				***NLRP3***	***0.01***	***7***
TMEM86A	***0.14***	***108***				***CD10***	***0.01***	***6***
MSMB	***0.14***	***107***				***TERF2IP***	***0.01***	***5***
P712P	***0.14***	***106***				***ANPEP***	***0.00***	***4***
PALM3	***0.14***	***105***				***MIC1***	***0.00***	***3***
SLC4A1.S	***0.14***	***104***				***CASKIN1***	***0.00***	***2***
MAPK8IP2	***0.14***	***103***				***SACM1L***	***0.00***	***1***
MCTP1	***0.14***	***102***						
ERG5'	***0.14***	***101***						
FOLH1	***0.14***	***100***						
AMH	***0.13***	***99***						
SEC61A1	***0.13***	***98***						
AR.ex9	***0.13***	***97***						
ABCB9	***0.13***	***96***						
MIR146A	***0.13***	***95***						
RPS11	***0.12***	***94***						
RAB17	***0.12***	***93***						
OR52A2	***0.12***	***92***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ACTR5	***0.11***	***91***						
RPS10	***0.11***	***90***						
Met	***0.11***	***89***						
OGT	***0.11***	***88***						
STEAP2	***0.11***	***87***						
MEX3A	***0.11***	***86***						
ITGBL1	***0.11***	***85***						
PECI	***0.10***	***84***						
SSTR1	***0.10***	***83***						
HIST1H1E	***0.10***	***82***						
HIST1H2BG	***0.10***	***81***						
MGAT5B	***0.10***	***80***						
SULF2	***0.10***	***79***						
HMBS	***0.10***	***78***						
MAK	***0.10***	***77***						
AR exons 4-8	***0.10***	***76***						
SMAP1 exons 7-8	***0.10***	***75***						
CDC37L1	***0.09***	***74***						
RPLP2	***0.09***	***73***						
AMACR	***0.09***	***72***						
NEAT1	***0.09***	***71***						
STEAP4	***0.09***	***70***						
MED4	***0.09***	***69***						
AURKA	***0.09***	***68***						
NKAIN1	***0.08***	***67***						
GOLM1	***0.08***	***66***						
CD10	***0.08***	***65***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ZNF577	***0.08***	***64***						
GAPDH	***0.08***	***63***						
KLK3 exons 1-2	***0.08***	***62***						
IMPDH2	***0.08***	***61***						
MXI1	***0.08***	***60***						
RPL23AP53	***0.08***	***59***						
FDPS	***0.08***	***58***						
ALAS1	***0.08***	***57***						
PPP1R12B	***0.08***	***56***						
PCSK6	***0.08***	***55***						
NLRP3	***0.08***	***54***						
MCM7	***0.07***	***53***						
DPP4	***0.07***	***52***						
ARHGEF25	***0.07***	***51***						
SRSF3	***0.07***	***50***						
STOM	***0.07***	***49***						
PTPRC	***0.07***	***48***						
VPS13A	***0.07***	***47***						
CACNA1D	***0.07***	***46***						
ANPEP	***0.07***	***45***						
MIC1	***0.07***	***44***						
CAMK2N2	***0.06***	***43***						
AATF	***0.06***	***42***						
KLK4	***0.06***	***41***						
HIST1H1C	***0.06***	***40***						
TRPM4	***0.06***	***39***						
KLK3 exons 2-3	***0.06***	***38***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
PVT1	***0.06***	***37***						
BTG2	***0.06***	***36***						
TERT	***0.06***	***35***						
SIRT1	***0.06***	***34***						
HPRT	***0.06***	***33***						
MIATNB	***0.05***	***32***						
KLK2	***0.05***	***31***						
MEMO1	***0.05***	***30***						
RPL18A	***0.05***	***29***						
COL10A1	***0.05***	***28***						
RP11_97O12.7	***0.05***	***27***						
GABARAPL2	***0.05***	***26***						
LBH	***0.04***	***25***						
MKi67	***0.04***	***24***						
EIF2D	***0.04***	***23***						
SULT1A1	***0.04***	***22***						
HOXC4	***0.04***	***21***						
CDC20	***0.04***	***20***						
HIST3H2A	***0.04***	***19***						
CDKN3	***0.04***	***18***						
CASKIN1	***0.03***	***17***						
MARCH5	***0.03***	***16***						
BRAF	***0.03***	***15***						
HIST1H2BF	***0.03***	***14***						
PSTPIP1	***0.03***	***13***						
ITPR1	***0.03***	***12***						
TFDP1	***0.03***	***11***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 51)*			*Transcripts identified by Mann Whitney U (n = 65)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
TERF2IP	***0.03***	***10***						
TBP	***0.03***	***9***						
MIR4435 1HG	***0.03***	***8***						
SYNM	***0.03***	***7***						
SACM1L	***0.03***	***6***						
SChLAP1	***0.03***	***5***						
SNCA	***0.02***	***4***						
SMIM1	***0.02***	***3***						
PDLIM5	***0.02***	***2***						
B2M	***0.01***	***1***						

Supplementary Table 41 Random Forest results for comparing high-risk cancer samples with clinically benign samples in *KLK2* factorised cell data.

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HOXC6	***1.66***	***166***	***CLIC2***	***0.80***	***12***	***HOXC6***	***1.73***	***25***
TDRD	***0.33***	***165***	***TDRD***	***0.70***	***11***	***FOLH1***	***0.48***	***24***
SLC43A1	***0.30***	***164***	***SLC43A1***	***0.68***	***10***	***CADPS***	***0.39***	***23***
FOLH1	***0.30***	***163***	***SIM2 long***	***0.65***	***9***	***TDRD***	***0.36***	***22***
CADPS	***0.27***	***162***	***ERG3' exons 6-7***	***0.58***	***8***	***SIM2 long***	***0.35***	***21***
SIM2 long	***0.26***	***161***	***PCA3***	***0.56***	***7***	***CLIC2***	***0.33***	***20***
CLIC2	***0.24***	***160***	***B4GALNT4***	***0.45***	***6***	***SLC43A1***	***0.30***	***19***
ERG3' exons 6-7	***0.21***	***159***	***HPN***	***0.40***	***5***	***ERG3' exons 6-7***	***0.25***	***18***
HPN	***0.19***	***158***	***TMPRSS2:ERG fusion***	***0.34***	***4***	***PCA3***	***0.24***	***17***
PCA3	***0.14***	***157***	***GCNT1***	***0.32***	***3***	***APOC1***	***0.22***	***16***
B4GALNT4	***0.13***	***156***	***LASS1***	***0.23***	***2***	***SLC12A1***	***0.22***	***15***
APOC1	***0.09***	***155***	***ERG3' exons 4-5***	***0.12***	***1***	***B4GALNT4***	***0.20***	***14***
NAALADL2	***0.09***	***154***				***GCNT1***	***0.11***	***13***
SLC12A1	***0.08***	***153***				***TMPRSS2:ERG fusion***	***0.10***	***12***
LASS1	***0.07***	***152***				***TMEM86A***	***0.10***	***11***
TMEM86A	***0.07***	***151***				***SEC61A1***	***0.10***	***10***
GCNT1	***0.07***	***150***				***HPN***	***0.10***	***9***
ISX	***0.05***	***149***				***CCDC88B***	***0.09***	***8***
HIST1H2BG	***0.05***	***148***				***ISX***	***0.09***	***7***
DLX1	***0.05***	***147***				***MCTP1***	***0.08***	***6***
Timp4	***0.04***	***146***				***ERG5'***	***0.07***	***5***
CAMKK2	***0.04***	***145***				***LASS1***	***0.07***	***4***
TMPRSS2:ERG fusion	***0.04***	***144***				***ERG3' exons 4-5***	***0.06***	***3***

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
GJB1	***0.04***	***143***				***SFRP4***	***0.03***	***2***
TMEM45B	***0.04***	***142***				***CAMKK2***	***0.03***	***1***
HIST1H1C	***0.04***	***141***						
SMIM1	***0.04***	***140***						
MMP25	***0.03***	***139***						
VAX2	***0.03***	***138***						
UPK2	***0.03***	***137***						
SULT1A1	***0.03***	***136***						
ABCB9	***0.03***	***135***						
SEC61A1	***0.03***	***134***						
RNF157	***0.03***	***133***						
CKAP2L	***0.03***	***132***						
AR exons 4-8	***0.03***	***131***						
AURKA	***0.03***	***130***						
IGFBP3	***0.02***	***129***						
P712P	***0.02***	***128***						
SIM2 short	***0.02***	***127***						
SFRP4	***0.02***	***126***						
GOLM1	***0.02***	***125***						
SPINK1	***0.02***	***124***						
ERG3' exons 4-5	***0.02***	***123***						
CD10	***0.02***	***122***						
ERG5'	***0.02***	***121***						
MGAT5B	***0.02***	***120***						
STEAP2	***0.02***	***119***						
ANKRD34B	***0.02***	***118***						
CP	***0.02***	***117***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SLC4A1.S	***0.02***	***116***						
MNX1	***0.02***	***115***						
ST6GALNAC1	***0.02***	***114***						
LBH	***0.02***	***113***						
COL9A2	***0.02***	***112***						
NKAIN1	***0.02***	***111***						
SRSF3	***0.02***	***110***						
SERPINB5	***0.02***	***109***						
KLK3 exons 2-3	***0.02***	***108***						
PPP1R12B	***0.01***	***107***						
ACTR5	***0.01***	***106***						
SPON2	***0.01***	***105***						
SULF2	***0.01***	***104***						
RPL23AP53	***0.01***	***103***						
CAMK2N2	***0.01***	***102***						
CDC37L1	***0.01***	***101***						
HIST1H2BF	***0.01***	***100***						
MIR146A	***0.01***	***99***						
TERT	***0.01***	***98***						
SACM1L	***0.01***	***97***						
ALAS1	***0.01***	***96***						
OR52A2	***0.01***	***95***						
HIST3H2A	***0.01***	***94***						
RPS11	***0.01***	***93***						
KLK3 exons 1-2	***0.01***	***92***						
NLRP3	***0.01***	***91***						
TMEM47	***0.01***	***90***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
MEX3A	***0.01***	***89***						
MKi67	***0.01***	***88***						
RIOK3	***0.01***	***87***						
PSTPIP1	***0.01***	***86***						
BRAF	***0.01***	***85***						
SSPO	***0.01***	***84***						
MDK	***0.01***	***83***						
ITGBL1	***0.01***	***82***						
AMACR	***0.01***	***81***						
VPS13A	***0.01***	***80***						
RAB17	***0.01***	***79***						
MIC1	***0.01***	***78***						
PPAP2A	***0.01***	***77***						
KLK4	***0.01***	***76***						
SNORA20	***0.01***	***75***						
PECI	***0.01***	***74***						
PTN	***0.01***	***73***						
RPS10	***0.01***	***72***						
MFSD2A	***0.01***	***71***						
CACNA1D	***0.01***	***70***						
PALM3	***0.01***	***69***						
MCTP1	***0.01***	***68***						
CCDC88B	***0.01***	***67***						
AMH	***0.01***	***66***						
STOM	***0.01***	***65***						
AGR2	***0.01***	***64***						
DNAH5	***0.01***	***63***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HOXC4	***0.01***	***62***						
TWIST1	***0.01***	***61***						
PDLIM5	***0.01***	***60***						
AATF	***0.01***	***59***						
PVT1	***0.004***	***58***						
B2M	***0.004***	***57***						
HPRT	***0.004***	***56***						
DPP4	***0.004***	***55***						
RPLP2	***0.004***	***54***						
MEMO1	***0.004***	***53***						
MSMB	***0.004***	***52***						
PPFIA2	***0.004***	***51***						
COL10A1	***0.004***	***50***						
ZNF577	***0.004***	***49***						
TRPM4	***0.004***	***48***						
MIATNB	***0.004***	***47***						
SChLAP1	***0.004***	***46***						
GAPDH	***0.004***	***44.5***						
RPL18A	***0.004***	***44.5***						
TMCC2	***0.003***	***43***						
MCM7	***0.003***	***42***						
NEAT1	***0.003***	***41***						
HIST1H1E	***0.003***	***40***						
CLU	***0.002***	***39***						
MYOF	***0.002***	***38***						
BTG2	***0.002***	***37***						
ITPR1	***0.002***	***36***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
CDC20	***0.002***	***35***						
STEAP4	***0.002***	***34***						
Met	***0.002***	***33***						
EN2	***0.002***	***32***						
SMAP1 exons 7-8	***0.002***	***31***						
SSTR1	***0.002***	***30***						
MAPK8IP2	***0.002***	***29***						
MAK	***0.002***	***28***						
GABARAPL2	***0.002***	***27***						
CASKIN1	***0.002***	***26***						
MED4	***0.002***	***25***						
IFT57	***0.002***	***24***						
AR.ex9	***0.002***	***23***						
TFDP1	***2.17604×10^{-17}***	***22***						
RP11_97O12.7	***2.13163×10^{-17}***	***21***						
CDKN3	***2.04281×10^{-17}***	***19.5***						
HMBS	***2.04281×10^{-17}***	***19.5***						
SNCA	***1.77636×10^{-17}***	***18***						
ARHGEF25	***1.73195×10^{-17}***	***17***						
OGT	***1.59872×10^{-17}***	***16***						
MXI1	***1.55431×10^{-17}***	***15***						
MARCH5	***1.46549×10^{-17}***	***14***						
MMP11	***1.42109×10^{-17}***	***13***						
TERF2IP	***1.33227×10^{-17}***	***12***						
SYNM	***1.24345×10^{-17}***	***11***						
ANPEP	***1.19904×10^{-17}***	***9.5***						
IMPDH2	***1.19904×10^{-17}***	***9.5***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 12)*			*Transcripts identified by Mann Whitney U (n = 25)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
MIR4435_1HG	***1.02141x10^{-17}***	***8***						
MMP26	***9.32587x10^{-18}***	***6***						
PCSK6	***9.32587x10^{-18}***	***6***						
TBP	***9.32587x10^{-18}***	***6***						
FDPS	***8.43769x10^{-18}***	***4***						
EIF2D	***7.10543x10^{-18}***	***3***						
SIRT1	***3.9968x10^{-18}***	***2***						
PTPRC	***3.55271x10^{-18}***	***1***						

Supplementary Table 42 Random Forest results when comparing clinically benign samples to high risk cancer samples using the *RPLP2* and *TWIST1* normalised data.

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
CADPS	***0.59***	***167***	***HOXC6***	***0.87***	***20***	***SPINK1***	***0.69***	***35***
SPINK1	***0.58***	***166***	***CADPS***	***0.77***	***19***	***HOXC6***	***0.68***	***34***
HOXC6	***0.50***	***165***	***SPINK1***	***0.77***	***18***	***CADPS***	***0.63***	***33***
ST6GALNAC1	***0.36***	***164***	***ERG3' exons 6-7***	***0.47***	***17***	***ST6GALNAC1***	***0.27***	***32***
VAX2	***0.27***	***163***	***ST6GALNAC1***	***0.39***	***16***	***VAX2***	***0.26***	***31***
ERG3' exons 6-7	***0.19***	***162***	***NAALADL2***	***0.34***	***15***	***ERG3' exons 6-7***	***0.25***	***30***
NAALADL2	***0.18***	***161***	***SLC43A1***	***0.25***	***14***	***B4GALNT4***	***0.23***	***29***
SLC43A1	***0.15***	***160***	***CLIC2***	***0.24***	***13***	***NAALADL2***	***0.22***	***28***
HPN	***0.13***	***159***	***TDRD***	***0.23***	***12***	***PPAP2A***	***0.20***	***27***
PPAP2A	***0.12***	***158***	***UPK2***	***0.22***	***11***	***SLC43A1***	***0.19***	***26***
UPK2	***0.12***	***157***	***PTN***	***0.21***	***10***	***HPN***	***0.16***	***25***
TDRD	***0.12***	***156***	***SERPINB5***	***0.20***	***9***	***UPK2***	***0.15***	***24***
PTN	***0.11***	***155***	***PALM3***	***0.17***	***8***	***CLIC2***	***0.14***	***23***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
TMPRSS2:ERG fusion	***0.10***	***154***	***GJB1***	***0.16***	***7***	***APOC1***	***0.14***	***22***
IFT57	***0.10***	***153***	***HPN***	***0.16***	***6***	***TDRD***	***0.13***	***21***
SERPINB5	***0.09***	***152***	***TMPRSS2:ERG fusion***	***0.15***	***5***	***ISX***	***0.13***	***20***
SFRP4	***0.08***	***151***	***CCDC88B***	***0.13***	***4***	***IFT57***	***0.12***	***19***
GJB1	***0.08***	***150***	***SIM2 long***	***0.13***	***3***	***TMEM86A***	***0.12***	***18***
CLIC2	***0.08***	***149***	***LASS1***	***0.11***	***2***	***MSMB***	***0.12***	***17***
MAPK8IP2	***0.08***	***148***	***GCNT1***	***0.09***	***1***	***SERPINB5***	***0.11***	***16***
B4GALNT4	***0.07***	***147***				***TMPRSS2:ERG fusion***	***0.11***	***15***
ERG5'	***0.06***	***146***				***MMP25***	***0.11***	***14***
COL9A2	***0.06***	***145***				***SPON2***	***0.09***	***13***
APOC1	***0.06***	***144***				***STEAP2***	***0.08***	***12***
SIM2 long	***0.06***	***143***				***PALM3***	***0.08***	***11***
PECI	***0.06***	***142***				***GCNT1***	***0.08***	***10***
ISX	***0.06***	***141***				***RAB17***	***0.08***	***9***
MSMB	***0.06***	***140***				***PTN***	***0.08***	***8***
CP	***0.06***	***139***				***SIM2 long***	***0.07***	***7***
TMEM86A	***0.05***	***138***				***RPS10***	***0.07***	***6***
MNX1	***0.05***	***137***				***GJB1***	***0.07***	***5***
PALM3	***0.05***	***136***				***CCDC88B***	***0.07***	***4***
IGFBP3	***0.04***	***135***				***MCTP1***	***0.05***	***3***
ANKRD34B	***0.04***	***134***				***LASS1***	***0.05***	***2***
LASS1	***0.04***	***133***				***IGFBP3***	***0.05***	***1***
CCDC88B	***0.04***	***132***						
TMCC2	***0.04***	***131***						
GCNT1	***0.04***	***130***						
FOLH1	***0.04***	***129***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
DLX1	***0.03***	***128***						
MMP25	***0.03***	***127***						
RAB17	***0.03***	***126***						
RPL18A	***0.03***	***125***						
MDK	***0.03***	***124***						
RPS10	***0.03***	***123***						
EN2	***0.03***	***122***						
RIOK3	***0.03***	***121***						
MFSD2A	***0.03***	***120***						
KLK3 exons 2-3	***0.03***	***119***						
CKAP2L	***0.03***	***118***						
PCA3	***0.03***	***117***						
PPFIA2	***0.02***	***116***						
MCTP1	***0.02***	***115***						
MYOF	***0.02***	***114***						
RNF157	***0.02***	***113***						
CDC37L1	***0.02***	***112***						
AMACR	***0.02***	***111***						
Timp4	***0.02***	***110***						
CDC20	***0.02***	***109***						
SEC61A1	***0.02***	***108***						
STEAP2	***0.02***	***107***						
SRSF3	***0.02***	***106***						
STOM	***0.02***	***105***						
SPON2	***0.02***	***104***						
MKi67	***0.02***	***103***						
SMIM1	***0.02***	***102***						
ITGBL1	***0.02***	***101***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HOXC4	***0.02***	***100***						
PPP1R12B	***0.01***	***99***						
KLK4	***0.01***	***98***						
ACTR5	***0.01***	***97***						
CLU	***0.01***	***96***						
AR exon 9	***0.01***	***95***						
RP11_97O12.7	***0.01***	***94***						
CAMKK2	***0.01***	***93***						
TRPM4	***0.01***	***92***						
MIR146A	***0.01***	***91***						
SIRT1	***0.01***	***90***						
GOLM1	***0.01***	***89***						
SLC4A1.S	***0.01***	***88***						
ZNF577	***0.01***	***87***						
RPS11	***0.01***	***86***						
PTPRC	***0.01***	***85***						
NLRP3	***0.01***	***84***						
TMEM47	***0.01***	***83***						
CACNA1D	***0.01***	***82***						
HMBS	***0.01***	***81***						
ABCB9	***0.01***	***80***						
PVT1	***0.01***	***79***						
SSPO	***0.01***	***78***						
ITPR1	***0.01***	***77***						
KLK3 exons 1-2	***0.01***	***76***						
STEAP4	***0.01***	***75***						
PCSK6	***0.01***	***74***						
AURKA	***0.01***	***73***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
TBP	***0.01***	***72***						
SChLAP1	***0.01***	***71***						
VPS13A	***0.01***	***70***						
NKAIN1	***0.01***	***69***						
MIATNB	***0.01***	***68***						
FDPS	***0.01***	***67***						
OR52A2	***0.01***	***66***						
RPL23AP53	***0.01***	***65***						
HIST1H2BF	***0.01***	***64***						
CAMK2N2	***0.01***	***63***						
DPP4	***0.01***	***62***						
SMAP1 exons 7-8	***0.01***	***61***						
HIST1H1C	***0.01***	***60***						
ALAS1	***0.01***	***59***						
TMEM45B	***0.005***	***58***						
TWIST1	***0.005***	***57***						
HIST1H1E	***0.005***	***56***						
MMP26	***0.004***	***55***						
SNCA	***0.004***	***54***						
BRAF	***0.004***	***53***						
GABARAPL2	***0.004***	***52***						
RPLP2	***0.004***	***51***						
MIR4435_1HG	***0.004***	***50***						
ERG3' exons 4-5	***0.004***	***49***						
AMH	***0.004***	***48***						
ANPEP	***0.004***	***47***						
SACM1L	***0.004***	***46***						
AGR2	***0.004***	***45***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
MEX3A	***0.004***	***44***						
MXI1	***0.004***	***43***						
MARCH5	***0.004***	***42***						
CD10	***0.003***	***41***						
EIF2D	***0.003***	***40***						
ARHGEF25	***0.003***	***39***						
NEAT1	***0.003***	***38***						
IMPDH2	***0.003***	***37***						
Met	***0.002***	***36***						
PSTPIP1	***0.002***	***35***						
P712P	***0.002***	***34***						
DNAH5	***0.002***	***33***						
MAK	***0.002***	***32***						
SIM2 short	***0.002***	***31***						
SYNM	***0.002***	***30***						
MCM7	***0.002***	***29***						
TERT	***0.002***	***28***						
AR exons 4-8	***0.002***	***27***						
PDLIM5	***0.002***	***26***						
B2M	***0.002***	***25***						
COL10A1	***0.002***	***24***						
LBH	***2.53×10^{-17}***	***23***						
SULF2	***2.18×10^{-17}***	***22***						
MMP11	***2.13×10^{-17}***	***20.5***						
SULT1A1	***2.13×10^{-17}***	***20.5***						
MGAT5B	***1.95×10^{-17}***	***19***						
CASKIN1	***1.87×10^{-17}***	***18***						
SLC12A1	***1.82×10^{-17}***	***17***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 20)*			*Transcripts identified by polr (n = 35)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HIST3H2A	$1.78x10^{-17}$	***15.5***						
MIC1	$1.78x10^{-17}$	***15.5***						
KLK2	$1.69x10^{-17}$	***14***						
BTG2	$1.64x10^{-17}$	***13***						
SNORA20	$1.60x10^{-17}$	***11.5***						
TFDP1	$1.60x10^{-17}$	***11.5***						
HIST1H2BG	$1.24x10^{-17}$	***10***						
HPRT	$1.15x10^{-17}$	***8.5***						
SSTR1	$1.15x10^{-17}$	***8.5***						
AATF	$1.02x10^{-17}$	***7***						
TERF2IP	$9.77x10^{-18}$	***6***						
CDKN3	$9.33x10^{-18}$	***5***						
MED4	$8.88x10-^{18}$	***4***						
MEMO1	$8.44x10^{-18}$	***3***						
GAPDH	$7.99x10^{-18}$	***2***						
OGT	$5.33x10^{-18}$	***1***						

6.20 CB- L-I-H Trend

Supplementary Table 43 Transcripts that have significant expression trend (using polr and glm) across clinically benign, low-risk, intermediate-risk and high-risk cancer samples in the baseline normalised cell NanoString data.

Transcript	*Glm p-value*	*glm Adjusted p-value*	*Polr p-value*	*Polr adjusted p-value*
AATF	***0.0002***	***0.035***	***0.0023***	***0.3326***
ACTR5	***0.0114***	***0.9827***	***0.0204***	***0.9931***
AMH	***0.0387***	***0.9827***		
ANKRD34B	***0.0033***	***0.4275***	***0.0065***	***0.878***
ANPEP	***0.0007***	***0.1031***	***0.0033***	***0.4854***
APOC1	***$2.90x10^{-06}$***	***0.0005***	***0.0001***	***0.0246***
ARHGEF25	***0.026***	***0.9827***	***0.0251***	***0.9931***
AURKA	***0.0066***	***0.7921***	***0.0232***	***0.9931***
B2M	***0.0022***	***0.2821***	***0.0092***	***0.9931***
B4GALNT4	***$7.83x10^{-06}$***	***0.0013***	***$9.95x10^{-05}$***	***0.0166***
BTG2	***0.0121***	***0.9827***	***0.0398***	***0.9931***
CACNA1D	***0.003***	***0.3935***	***0.0184***	***0.9931***
CADPS	***0.0001***	***0.019***	***0.0017***	***0.261***
CAMKK2	***$5.44x10^{-05}$***	***0.0087***	***0.0012***	***0.1813***
CCDC88B	***0.0008***	***0.1059***	***0.0043***	***0.6079***
CD10	***0.0011***	***0.158***	***0.0047***	***0.6525***
CDC20	***0.018***	***0.9827***	***0.041***	***0.9931***
CDKN3	***0.0218***	***0.9827***	***0.0342***	***0.9931***
CKAP2L	***0.0063***	***0.7631***	***0.0163***	***0.9931***
CLIC2	***0.0012***	***0.1699***	***0.0071***	***0.9579***
COL9A2	***0.0273***	***0.9827***	***0.031***	***0.9931***
DPP4	***0.0385***	***0.9827***		
EIF2D	***0.0038***	***0.4774***	***0.0197***	***0.9931***
EN2	***0.0141***	***0.9827***	***0.0189***	***0.9931***
ERG3' exons 4-5	***0.0065***	***0.7819***	***0.0083***	***0.9931***
ERG3' exons 6-7	***$5.87x10^{-06}$***	***0.001***	***0.0001***	***0.0191***
FDPS	***0.0231***	***0.9827***		
FOLH1	***0.0031***	***0.3935***	***0.0055***	***0.7553***
GABARAPL2	***0.0039***	***0.4932***	***0.0317***	***0.9931***
GAPDH	***0.0004***	***0.0542***	***0.0022***	***0.3326***
GCNT1	***0.0002***	***0.025***	***0.0008***	***0.1318***
HIST1H2BF	***0.0265***	***0.9827***	***0.0481***	***0.9931***
HIST1H2BG	***0.0092***	***0.9827***	***0.026***	***0.9931***
HOXC6	***$1.85x10^{-05}$***	***0.003***	***0.0001***	***0.0191***
HPN	***$6.51x10^{-05}$***	***0.0102***	***0.0007***	***0.1079***
HPRT	***0.0015***	***0.1973***	***0.0107***	***0.9931***
ISX	***$3.91x10^{-05}$***	***0.0063***	***0.0004***	***0.066***
ITPR1	***0.0016***	***0.2177***	***0.0118***	***0.9931***
LASS1	***0.0002***	***0.0331***	***0.0019***	***0.2854***
MAK	***0.0482***	***0.9827***		

Transcript	*Glm p-value*	*glm Adjusted p-value*	*Polr p-value*	*Polr adjusted p-value*
MAPK8IP2	***0.036***	***0.9827***		
MCTP1	***4.10x10^{-06}***	***0.0007***	***0.0003***	***0.0481***
MED4	***0.0335***	***0.9827***		
MEMO1	***0.0368***	***0.9827***		
MEX3A	***0.0046***	***0.5676***	***0.0073***	***0.9736***
MFSD2A	***0.0003***	***0.0379***	***0.0035***	***0.4969***
MGAT5B	***0.0254***	***0.9827***	***0.0352***	***0.9931***
MIC1	***0.0484***	***0.9827***	***0.0359***	***0.9931***
MIR146A	***0.021***	***0.9827***		
MIR4435_1HG	***0.0018***	***0.2367***	***0.0131***	***0.9931***
MMP11	***0.0231***	***0.9827***	***0.0497***	***0.9931***
MMP25	***0.0002***	***0.0277***	***0.0014***	***0.2178***
MMP26	***0.0317***	***0.9827***	***0.032***	***0.9931***
MXI1	***0.0006***	***0.0855***	***0.0057***	***0.7873***
NEAT1	***0.0002***	***0.0265***	***0.0003***	***0.047***
NLRP3	***0.0012***	***0.1632***	***0.0042***	***0.6027***
PCA3	***0.0154***	***0.9827***		
PDLIM5	***0.0101***	***0.9827***	***0.0468***	***0.9931***
PSTPIP1	***0.0149***	***0.9827***	***0.0217***	***0.9931***
PTPRC	***0.0003***	***0.0504***	***0.0036***	***0.5209***
RIOK3	***6.64x10^{-06}***	***0.0011***	***0.0003***	***0.0515***
RPL18A	***0.04***	***0.9827***		
RPS11	***0.0436***	***0.9827***		
SACM1L	***0.0239***	***0.9827***	***0.0396***	***0.9931***
SEC61A1	***0.001***	***0.1335***	***0.0074***	***0.9864***
SFRP4	***0.0054***	***0.6665***	***0.0235***	***0.9931***
SIM2 long	***0.0004***	***0.061***	***0.0022***	***0.3237***
SIM2 short	***0.0139***	***0.9827***	***0.0253***	***0.9931***
SIRT1	***0.0018***	***0.2367***	***0.0176***	***0.9931***
SLC12A1	***0.0317***	***0.9827***		

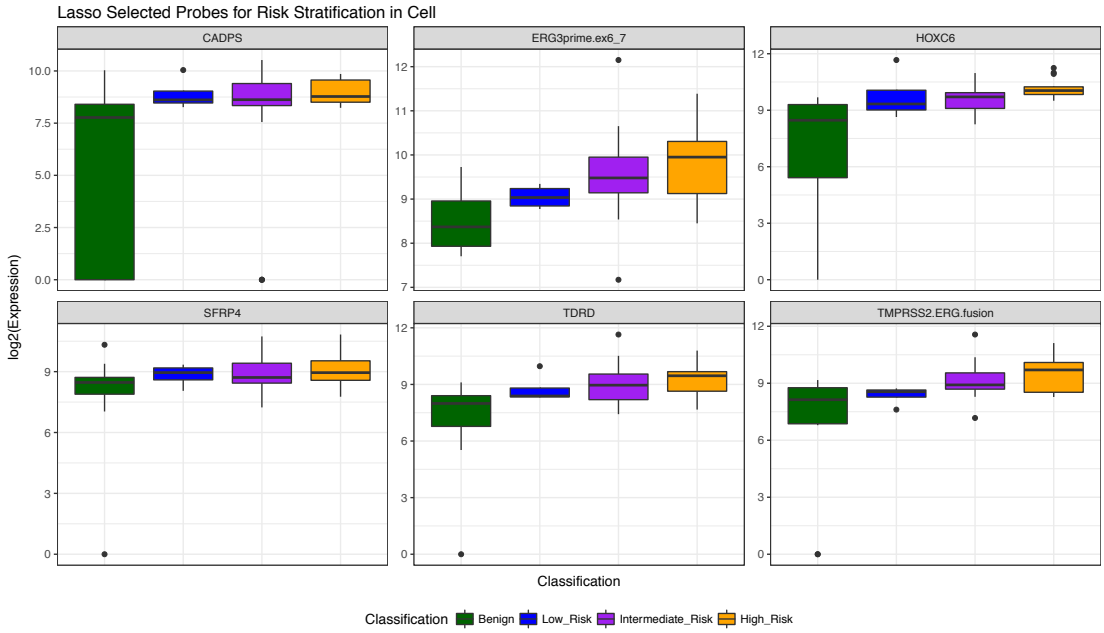
Supplementary Table 44 Transcripts that have significant expression trend (using polr and glm) across clinically benign, low-risk, intermediate-risk and high-risk cancer samples in the *KLK2* ratio cell NanoString data.

Transcript	*Glm p-value*	*glm Adjusted p-value*	*Polr p-value*	*Polr adjusted p-value*
ANKRD34B	***0.0112***	***0.998***	***0.0374***	***0.994***
APOC1	***0.0178***	***0.998***		
B4GALNT4	***0.0216***	***0.998***		
CADPS	***0.0052***	***0.8154***	***0.0112***	***0.994***
CAMKK2	***0.0385***	***0.998***		
CCDC88B	***0.0175***	***0.998***		
CKAP2L	***0.0046***	***0.731***	***0.0187***	***0.994***
CLIC2	***0.0191***	***0.998***	***0.0397***	***0.994***
ERG3' exons 4-5	***0.0275***	***0.998***	***0.0243***	***0.994***
ERG3' exons 6-7	***2.18 x10^{-05}***	***0.0036***	***0.0002***	***0.0283***
FOLH1	***0.0124***	***0.998***	***0.0348***	***0.994***
GCNT1	***0.0066***	***0.998***	***0.0104***	***0.994***
HOXC6	***1.36x10^{-05}***	***0.0023***	***0.0002***	***0.0406***
HPN	***0.0027***	***0.4419***	***0.0081***	***0.994***
ISX	***0.0022***	***0.3607***	***0.0367***	***0.994***
LASS1	***0.005***	***0.7872***	***0.0209***	***0.994***
MAPK8IP2	***0.0215***	***0.998***		
MCTP1	***0.0262***	***0.998***	***0.0467***	***0.994***
MEX3A	***0.0227***	***0.998***	***0.0412***	***0.994***
MFSD2A	***0.0103***	***0.998***		
MIR146A	***0.0494***	***0.998***		
MMP25	***0.0274***	***0.998***	***0.0463***	***0.994***
NLRP3	***0.0386***	***0.998***		
PCA3	***0.0182***	***0.998***		
PSTPIP1	***0.0274***	***0.998***		
RIOK3	***0.0386***	***0.998***		
SEC61A1	***0.0337***	***0.998***		
SFRP4	***0.0156***	***0.998***	***0.0494***	***0.994***
SIM2 long	***0.0031***	***0.5021***	***0.0056***	***0.9028***
SLC43A1	***0.0199***	***0.998***		
SNORA20	***0.045***	***0.998***		
SULF2	***0.0153***	***0.998***		
TDRD	***0.0002***	***0.0263***	***0.0011***	***0.1757***
TMCC2	***0.0331***	***0.998***		
TMEM86A	***0.0079***	***0.998***	***0.0184***	***0.994***
TMPRSS2:ERG	***6.86 x10^{-05}***	***0.0112***	***0.0007***	***0.1136***

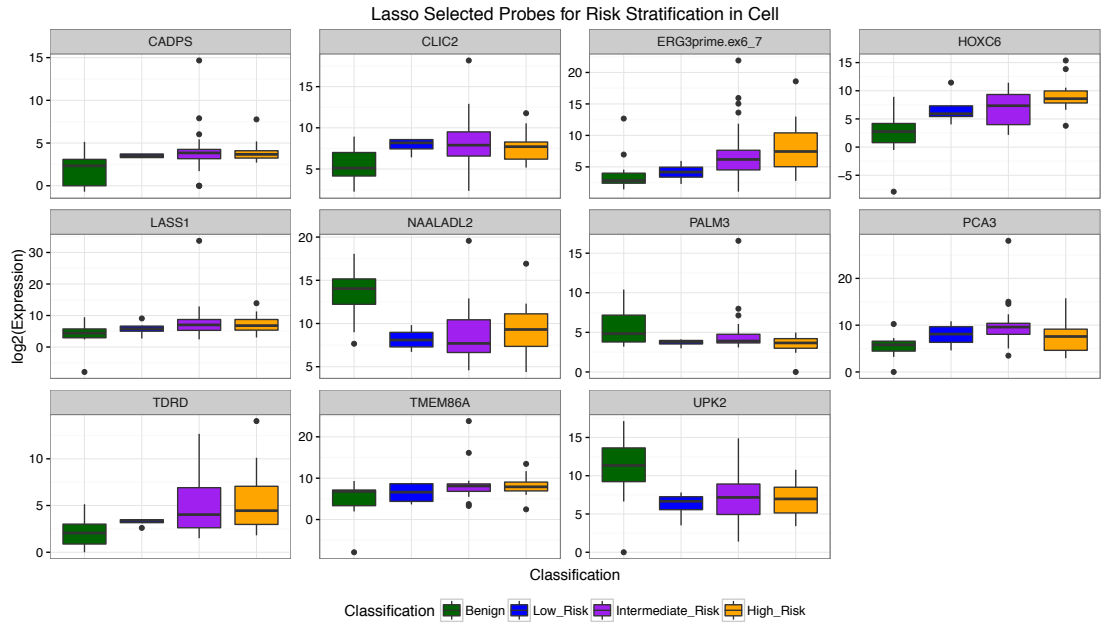
Supplementary Table 45 Transcripts that have significant expression trend (using polr and glm) across clinically benign, low-risk, intermediate-risk and high-risk cancer samples in the HK normalised cell NanoString data.

Transcript	*Glm p-value*	*glm Adjusted p-value*	*Polr p-value*	*Polr adjusted p-value*
CADPS	***0.0213***	***0.9941***		
CLIC2	***0.0333***	***0.9941***		
EN2			***0.0463***	***0.9994***
ERG 3' exons 6-7	***0.0043***	***0.6877***	***0.0098***	***0.9994***
FOLH1	***0.0174***	***0.9941***	***0.0191***	***0.9994***
GJB1	***0.0215***	***0.9941***		
HOXC6	***4.54 x10^{-6}***	***0.0008***	***6.37x10^{-05}***	***0.0106***
LASS1	***0.0287***	***0.9941***		
MEX3A	***0.0243***	***0.9941***	***0.0337***	***0.9994***
MSMB	***0.0334***	***0.9941***		
NAALADL2	***0.0018***	***0.2913***	***0.0098***	***0.9994***
PALM3	***0.027***	***0.9941***	***0.0461***	***0.9994***
SERPINB5	***0.0162***	***0.9941***	***0.0425***	***0.9994***
SIM2 long	***0.0032***	***0.5147***	***0.0043***	***0.7056***
SLC43A1	***0.0011***	***0.1895***	***0.006***	***0.978***
ST6GALNAC1	***0.0049***	***0.7755***	***0.0179***	***0.9994***
TDRD	***0.0012***	***0.2024***	***0.0034***	***0.564***
TMEM86A	***0.0107***	***0.9941***	***0.0337***	***0.9994***
TMPRSS2:ERG fusion	***0.004***	***0.6414***	***0.0127***	***0.9994***
UPK2	***0.0028***	***0.4609***	***0.0077***	***0.9994***

Supplementary Figure 19 Boxplots of the Lasso identified transcripts for modeling between clinically benign, low-risk, intermediate-risk and high-risk cancer categories in the baseline normalised data.



Supplementary Figure 20 Boxplots showing the Lasso selected transcripts for CB-L-I-H trend in *KLK2* ratio data.



Supplementary Figure 21 Transcripts selected by Lasso for showing trend of expression levels across clinical categories: clinically benign, low-risk, intermediate-risk and high-risk cancer in the HK normalised cell data.

Supplementary Table 46 Random Forest rankings for three subsets of transcripts (all 167, the 87 chosen by glm, and the 70 chosen by polr*), for groups CB, L, I and H. (*All 70 transcripts identified by polr are were also common to those identified by glm) in the baseline normalization data.

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ERG3' exons 6-7	***6.60***	***167***	***ERG3' exons 6-7***	***8.10***	***87***	***ERG3' exons 6-7***	***5.00***	***70***
TMPRSS2:ERG fusion	***4.31***	***166***	***APOC1***	***3.84***	***86***	***TMPRSS2:ERG fusion***	***4.76***	***69***
RIOK3	***2.55***	***165***	***SPINK1***	***2.94***	***85***	***NEAT1***	***3.46***	***68***
NEAT1	***2.32***	***164***	***CCDC88B***	***2.58***	***84***	***RIOK3***	***3.08***	***67***
CADPS	***1.74***	***163***	***CADPS***	***2.36***	***83***	***APOC1***	***2.26***	***66***
APOC1	***1.51***	***162***	***B4GALNT4***	***2.09***	***82***	***SIM2 long***	***2.14***	***65***
SIM2 long	***1.51***	***161***	***CAMKK2***	***2.06***	***81***	***CCDC88B***	***1.84***	***64***
SPINK1	***1.42***	***160***	***GAPDH***	***1.79***	***80***	***MCTP1***	***1.61***	***63***
MCTP1	***1.34***	***159***	***CKAP2L***	***1.45***	***79***	***GCNT1***	***1.55***	***62***
CCDC88B	***1.11***	***158***	***TDRD***	***1.24***	***78***	***CADPS***	***1.49***	***61***
MFSD2A	***1.02***	***157***	***CP***	***1.19***	***77***	***HOXC6***	***1.49***	***60***
CAMKK2	***1.01***	***156***	***ISX***	***1.07***	***76***	***MFSD2A***	***1.45***	***59***
GCNT1	***1.00***	***155***	***CD10***	***1.01***	***75***	***SPINK1***	***1.43***	***58***
MXI1	***0.90***	***154***	***HPN***	***0.99***	***74***	***CAMKK2***	***1.40***	***57***
TMEM86A	***0.89***	***153***	***UPK2***	***0.97***	***73***	***SIRT1***	***1.27***	***56***
HOXC6	***0.86***	***152***	***SLC4A1 S***	***0.93***	***72***	***SULT1A1***	***1.18***	***55***
CP	***0.86***	***151***	***PCA3***	***0.89***	***71***	***CKAP2L***	***1.14***	***54***
SLC43A1	***0.83***	***150***	***CACNA1D***	***0.89***	***70***	***LASS1***	***0.93***	***53***
SFRP4	***0.82***	***149***	***AATF***	***0.83***	***69***	***TMEM86A***	***0.91***	***52***
B4GALNT4	***0.78***	***148***	***SNORA20***	***0.81***	***68***	***SLC43A1***	***0.88***	***51***
SIRT1	***0.72***	***147***	***IGFBP3***	***0.75***	***67***	***AURKA***	***0.87***	***50***
MIR4435_1HG	***0.72***	***146***	***ANKRD34B***	***0.70***	***66***	***HPRT***	***0.85***	***49***
CKAP2L	***0.71***	***145***	***CLIC2***	***0.69***	***65***	***MIR4435_1HG***	***0.80***	***48***
SULT1A1	***0.71***	***144***	***SMAP1 exons 7-***	***0.69***	***64***	***B4GALNT4***	***0.76***	***47***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
			8					
LASS1	***0.67***	***143***	***DLX1***	***0.66***	***63***	***CD10***	***0.72***	***46***
MYOF	***0.56***	***142***	***AURKA***	***0.65***	***62***	***MIC1***	***0.69***	***45***
HPN	***0.54***	***141***	***MYOF***	***0.64***	***61***	***MMP25***	***0.69***	***44***
PCA3	***0.54***	***140***	***ERG3' exons 4-5***	***0.62***	***60***	***SFRP4***	***0.58***	***43***
SNORA20	***0.52***	***139***	***AMH***	***0.59***	***59***	***ANKRD34B***	***0.57***	***42***
SLC4A1 S	***0.50***	***138***	***TERF2IP***	***0.58***	***58***	***SRSF3***	***0.56***	***41***
CD10	***0.49***	***137***	***AMACR***	***0.58***	***57***	***HIST1H2BF***	***0.51***	***40***
SMAP1 exons 7-8	***0.49***	***136***	***ERG5'***	***0.55***	***56***	***SULF2***	***0.50***	***39***
MMP25	***0.45***	***135***	***MAK***	***0.48***	***55***	***TDRD***	***0.50***	***38***
AMH	***0.42***	***134***	***ANPEP***	***0.46***	***54***	***CDKN3***	***0.49***	***37***
TDRD	***0.41***	***133***	***DPP4***	***0.44***	***53***	***EN2***	***0.48***	***36***
SULF2	***0.40***	***132***	***NAALADL2***	***0.44***	***52***	***MXI1***	***0.46***	***35***
OR52A2	***0.39***	***131***	***MEX3A***	***0.43***	***51***	***STEAP4***	***0.46***	***34***
MIC1	***0.39***	***130***	***EN2***	***0.43***	***50***	***ITPR1***	***0.45***	***33***
HPRT	***0.38***	***129***	***RNF157***	***0.41***	***49***	***ERG3' exons 4-5***	***0.45***	***32***
HIST1H1E	***0.37***	***128***	***DNAH5***	***0.40***	***48***	***SNCA***	***0.44***	***31***
MAPK8IP2	***0.37***	***127***	***HIST1H2BG***	***0.39***	***47***	***MEX3A***	***0.44***	***30***
UPK2	***0.37***	***126***	***ACTR5***	***0.38***	***46***	***PSTPIP1***	***0.44***	***29***
GAPDH	***0.36***	***125***	***TFDP1***	***0.37***	***45***	***GAPDH***	***0.41***	***28***
AURKA	***0.33***	***124***	***CDKN3***	***0.37***	***44***	***CACNA1D***	***0.40***	***27***
ANKRD34B	***0.32***	***123***	***MCM7***	***0.37***	***43***	***ISX***	***0.39***	***26***
STOM	***0.29***	***122***	***CDC20***	***0.36***	***42***	***TMCC2***	***0.38***	***25***
Timp4	***0.29***	***121***	***AR exon 9***	***0.34***	***41***	***PTPRC***	***0.38***	***24***
DLX1	***0.28***	***120***	***PALM3***	***0.33***	***40***	***AATF***	***0.38***	***23***
EN2	***0.27***	***119***	***SSTR1***	***0.31***	***39***	***MMP26***	***0.38***	***22***
RPL23AP53	***0.27***	***118***	***AGR2***	***0.31***	***38***	***CLIC2***	***0.36***	***21***
ITPR1	***0.26***	***117***	***ABCB9***	***0.30***	***37***	***COL9A2***	***0.34***	***20***

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ISX	***0.26***	***116***	***CAMK2N2***	***0.29***	***36***	***TERF2IP***	***0.33***	***19***
RNF157	***0.25***	***115***	***CLU***	***0.28***	***35***	***MMP11***	***0.31***	***18***
TMEM45B	***0.24***	***114***	***BTG2***	***0.27***	***34***	***ACTR5***	***0.30***	***17***
FOLH1	***0.24***	***113***	***ALAS1***	***0.27***	***33***	***ANPEP***	***0.30***	***16***
CDKN3	***0.24***	***112***	***NLRP3***	***0.27***	***32***	***SACM1L***	***0.30***	***15***
AATF	***0.23***	***111***	***TMCC2***	***0.26***	***31***	***HIST1H2BG***	***0.30***	***14***
SLC12A1	***0.23***	***110***	***OGT***	***0.25***	***30***	***PDLIM5***	***0.28***	***13***
CACNA1D	***0.23***	***109***	***EIF2D***	***0.24***	***29***	***NLRP3***	***0.27***	***12***
SACM1L	***0.23***	***108***	***ARHGEF25***	***0.24***	***28***	***HPN***	***0.27***	***11***
IGFBP3	***0.22***	***107***	***FOLH1***	***0.24***	***27***	***MGAT5B***	***0.26***	***10***
ACTR5	***0.21***	***106***	***LBH***	***0.23***	***26***	***FOLH1***	***0.26***	***9***
MEX3A	***0.20***	***105***	***TMEM47***	***0.22***	***25***	***EIF2D***	***0.21***	***8***
DPP4	***0.20***	***104***	***ST6GALNAC1***	***0.22***	***24***	***SIM2 short***	***0.20***	***7***
SRSF3	***0.20***	***103***	***B2M***	***0.22***	***23***	***CDC20***	***0.20***	***6***
AR.ex9	***0.20***	***102***	***Met***	***0.21***	***22***	***BTG2***	***0.19***	***5***
ANPEP	***0.19***	***101***	***RP11_97O12.7***	***0.19***	***21***	***GABARAPL2***	***0.19***	***4***
VAX2	***0.19***	***100***	***PPAP2A***	***0.19***	***20***	***SEC61A1***	***0.16***	***3***
ERG5'	***0.19***	***99***	***COL9A2***	***0.18***	***19***	***B2M***	***0.15***	***2***
COL9A2	***0.19***	***98***	***COL10A1***	***0.18***	***18***	***ARHGEF25***	***0.15***	***1***
AGR2	***0.18***	***97***	***GOLM1***	***0.17***	***17***			
CLIC2	***0.18***	***96***	***PECI***	***0.16***	***16***			
ABCB9	***0.18***	***95***	***AR exons 4-8***	***0.15***	***15***			
DNAH5	***0.18***	***94***	***CDC37L1***	***0.15***	***14***			
MSMB	***0.18***	***93***	***SIM2 short***	***0.15***	***13***			
STEAP4	***0.17***	***92***	***TBP***	***0.14***	***12***			
SSTR1	***0.17***	***91***	***MIATNB***	***0.14***	***11***			
HIST1H2BF	***0.16***	***90***	***PCSK6***	***0.14***	***10***			
HMBS	***0.15***	***89***	***MARCH5***	***0.13***	***9***			

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SChLAP1	***0.15***	***88***	***BRAF***	***0.13***	***8***			
LBH	***0.15***	***87***	***HOXC4***	***0.13***	***7***			
PPFIA2	***0.15***	***86***	***FDPS***	***0.12***	***6***			
NLRP3	***0.15***	***85***	***KLK2***	***0.12***	***5***			
PTPRC	***0.15***	***84***	***RPS11***	***0.11***	***4***			
SPON2	***0.14***	***83***	***CASKIN1***	***0.11***	***3***			
AMACR	***0.14***	***82***	***GABARAPL2***	***0.10***	***2***			
MCM7	***0.14***	***81***	***STEAP2***	***0.08***	***1***			
MIR146A	***0.14***	***80***						
KLK4	***0.14***	***79***						
ALAS1	***0.13***	***78***						
MMP26	***0.13***	***77***						
PPAP2A	***0.13***	***76***						
MNX1	***0.13***	***75***						
ERG3' exons 4-5	***0.13***	***74***						
CDC37L1	***0.13***	***73***						
NAALADL2	***0.13***	***72***						
PSTPIP1	***0.12***	***71***						
SSPO	***0.11***	***70***						
EIF2D	***0.11***	***69***						
CDC20	***0.11***	***68***						
CLU	***0.11***	***67***						
PALM3	***0.11***	***66***						
KLK3 exons 2-3	***0.11***	***65***						
TRPM4	***0.11***	***64***						
MKi67	***0.11***	***63***						
TERF2IP	***0.10***	***62***						
HOXC4	***0.10***	***61***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
COL10A1	***0.10***	***60***						
RP11_97O12.7	***0.10***	***59***						
SEC61A1	***0.10***	***58***						
RAB17	***0.10***	***57***						
NKAIN1	***0.10***	***56***						
MDK	***0.10***	***55***						
SNCA	***0.09***	***54***						
MGAT5B	***0.09***	***53***						
VPS13A	***0.09***	***52***						
MED4	***0.09***	***51***						
ARHGEF25	***0.09***	***50***						
MAK	***0.09***	***49***						
PPP1R12B	***0.09***	***48***						
TBP	***0.09***	***47***						
SERPINB5	***0.08***	***46***						
GJB1	***0.08***	***45***						
BTG2	***0.08***	***44***						
MEMO1	***0.08***	***43***						
HIST3H2A	***0.08***	***42***						
TERT	***0.08***	***41***						
PVT1	***0.08***	***40***						
TFDP1	***0.07***	***39***						
P712P	***0.07***	***38***						
ZNF577	***0.07***	***37***						
Met	***0.07***	***36***						
OGT	***0.07***	***35***						
AR exons 4-8	***0.07***	***34***						
ITGBL1	***0.07***	***33***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
GOLM1	***0.07***	***32***						
FDPS	***0.07***	***31***						
MIATNB	***0.07***	***30***						
B2M	***0.07***	***29***						
RPS10	***0.06***	***28***						
ST6GALNAC1	***0.06***	***27***						
RPL18A	***0.06***	***26***						
IMPDH2	***0.06***	***25***						
SMIM1	***0.05***	***24***						
HIST1H2BG	***0.05***	***23***						
TMCC2	***0.05***	***22***						
STEAP2	***0.05***	***21***						
RPS11	***0.05***	***20***						
IFT57	***0.05***	***19***						
BRAF	***0.05***	***18***						
TWIST1	***0.05***	***17***						
CAMK2N2	***0.05***	***16***						
SIM2 short	***0.05***	***15***						
MMP11	***0.04***	***14***						
HIST1H1C	***0.04***	***13***						
PCSK6	***0.04***	***12***						
PECI	***0.04***	***11***						
PDLIM5	***0.04***	***10***						
MARCH5	***0.04***	***9***						
CASKIN1	***0.04***	***8***						
TMEM47	***0.04***	***7***						
RPLP2	***0.04***	***6***						
KLK2	***0.03***	***5***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 87)*			*Transcripts identified by polr (n = 70)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
GABARAPL2	***0.03***	***4***						
PTN	***0.03***	***3***						
KLK3 exons 1-2	***0.03***	***2***						
SYNM	***0.01***	***1***						

Supplementary Table 47 Random Forest results for trend across clinical categories: CBN-L-I-H in *KLK2* factorised data.

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
PCA3	***1.19***	***166***	***PCA3***	***3.36***	***36***	***HOXC6***	***3.58***	***20***
HOXC6	***1.14***	***165***	***HOXC6***	***2.56***	***35***	***ERG3' exons 6-7***	***2.98***	***19***
TMPRSS2:ERG fusion	***0.89***	***164***	***ERG3' exons 6-7***	***1.85***	***34***	***TMPRSS2:ERG fusion***	***2.97***	***18***
ERG3' exons 6-7	***0.88***	***163***	***TMPRSS2:ERG fusion***	***1.81***	***33***	***FOLH1***	***2.30***	***17***
SLC12A1	***0.59***	***162***	***FOLH1***	***1.42***	***32***	***TMEM86A***	***2.09***	***16***
NAALADL2	***0.59***	***161***	***TDRD***	***1.24***	***31***	***HPN***	***2.01***	***15***
APOC1	***0.53***	***160***	***APOC1***	***1.20***	***30***	***CKAP2L***	***1.97***	***14***
FOLH1	***0.50***	***159***	***SLC43A1***	***1.16***	***29***	***GCNT1***	***1.82***	***13***
CP	***0.49***	***158***	***TMEM86A***	***1.15***	***28***	***CADPS***	***1.79***	***12***
OR52A2	***0.49***	***157***	***GCNT1***	***1.11***	***27***	***TDRD***	***1.77***	***11***
SIM2 long	***0.49***	***156***	***CKAP2L***	***1.09***	***26***	***MMP25***	***1.71***	***10***
TDRD	***0.47***	***155***	***SIM2 long***	***1.08***	***25***	***SIM2 long***	***1.54***	***9***
PALM3	***0.45***	***154***	***HPN***	***1.07***	***24***	***CLIC2***	***1.52***	***8***
SERPINB5	***0.44***	***153***	***B4GALNT4***	***1.07***	***23***	***ISX***	***1.48***	***7***
AR exons 4-8	***0.41***	***152***	***CADPS***	***1.02***	***22***	***ANKRD34B***	***1.47***	***6***
TMEM86A	***0.40***	***151***	***MAPK8IP2***	***0.95***	***21***	***MCTP1***	***1.46***	***5***
MSMB	***0.40***	***150***	***ANKRD34B***	***0.91***	***20***	***LASS1***	***1.44***	***4***
MDK	***0.39***	***149***	***MMP25***	***0.90***	***19***	***SFRP4***	***1.40***	***3***
CKAP2L	***0.38***	***148***	***SEC61A1***	***0.90***	***18***	***MEX3A***	***1.30***	***2***
DLX1	***0.37***	***147***	***LASS1***	***0.87***	***17***	***ERG3' exons 4-5***	***1.08***	***1***
HPN	***0.36***	***146***	***CLIC2***	***0.85***	***16***			
SLC43A1	***0.36***	***145***	***SULF2***	***0.82***	***15***			
STEAP2	***0.35***	***144***	***TMCC2***	***0.80***	***14***			
IGFBP3	***0.34***	***143***	***CCDC88B***	***0.77***	***13***			

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
SPON2	***0.34***	***142***	***MEX3A***	***0.75***	***12***			
LASS1	***0.34***	***141***	***CAMKK2***	***0.74***	***11***			
TMEM47	***0.34***	***140***	***SNORA20***	***0.72***	***10***			
AGR2	***0.33***	***139***	***SFRP4***	***0.72***	***9***			
CADPS	***0.32***	***138***	***MFSD2A***	***0.69***	***8***			
MMP11	***0.31***	***137***	***MCTP1***	***0.68***	***7***			
GJB1	***0.30***	***136***	***ERG3' exons 4-5***	***0.66***	***6***			
SSTR1	***0.30***	***135***	***NLRP3***	***0.62***	***5***			
TMCC2	***0.30***	***134***	***PSTPIP1***	***0.59***	***4***			
AMACR	***0.30***	***133***	***MIR146A***	***0.58***	***3***			
B4GALNT4	***0.29***	***132***	***RIOK3***	***0.54***	***2***			
SULF2	***0.29***	***131***	***ISX***	***0.45***	***1***			
GCNT1	***0.29***	***130***						
ZNF577	***0.28***	***129***						
ANKRD34B	***0.28***	***128***						
HIST1H2BG	***0.27***	***127***						
SPINK1	***0.27***	***126***						
MMP25	***0.27***	***125***						
HIST3H2A	***0.26***	***124***						
TRPM4	***0.26***	***123***						
SLC4A1.S	***0.25***	***122***						
SULT1A1	***0.25***	***121***						
CDKN3	***0.25***	***120***						
Timp4	***0.25***	***119***						
ST6GALNAC1	***0.25***	***118***						
SNORA20	***0.25***	***117***						
EN2	***0.25***	***116***						
AR exon 9	***0.24***	***115***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ITGBL1	***0.24***	***114***						
UPK2	***0.24***	***113***						
MKi67	***0.24***	***112***						
SChLAP1	***0.24***	***111***						
AMH	***0.23***	***110***						
MCTP1	***0.23***	***109***						
SFRP4	***0.23***	***108***						
MFSD2A	***0.23***	***107***						
SIM2 short	***0.23***	***106***						
PPP1R12B	***0.23***	***105***						
TERT	***0.23***	***104***						
RAB17	***0.22***	***103***						
NKAIN1	***0.22***	***102***						
SMIM1	***0.22***	***101***						
P712P	***0.22***	***100***						
ERG3' exons 4-5	***0.22***	***99***						
PECI	***0.22***	***98***						
ERG5'	***0.22***	***97***						
VAX2	***0.22***	***96***						
CLIC2	***0.22***	***95***						
RNF157	***0.21***	***94***						
CDC37L1	***0.21***	***93***						
CCDC88B	***0.21***	***92***						
CLU	***0.20***	***91***						
MIC1	***0.20***	***90***						
TMEM45B	***0.20***	***89***						
MNX1	***0.20***	***88***						
ISX	***0.20***	***87***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HIST1H1C	***0.19***	***86***						
KLK4	***0.18***	***85***						
LBH	***0.18***	***84***						
COL10A1	***0.18***	***83***						
MED4	***0.18***	***82***						
HIST1H2BF	***0.18***	***81***						
PPAP2A	***0.18***	***80***						
ABCB9	***0.17***	***79***						
STOM	***0.17***	***78***						
DNAH5	***0.17***	***77***						
DPP4	***0.17***	***76***						
MMP26	***0.17***	***75***						
HOXC4	***0.16***	***74***						
MGAT5B	***0.16***	***73***						
MIR146A	***0.16***	***72***						
PCSK6	***0.16***	***71***						
CAMKK2	***0.16***	***70***						
MARCH5	***0.15***	***69***						
RPL23AP53	***0.15***	***68***						
IMPDH2	***0.15***	***67***						
HPRT	***0.15***	***66***						
ACTR5	***0.15***	***65***						
MAPK8IP2	***0.15***	***64***						
SNCA	***0.15***	***63***						
SYNM	***0.15***	***62***						
PSTPIP1	***0.15***	***61***						
CACNA1D	***0.14***	***60***						
PVT1	***0.14***	***59***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
HMBS	***0.14***	***58***						
SACM1L	***0.14***	***57***						
KLK3 exons 2-3	***0.14***	***56***						
COL9A2	***0.14***	***55***						
SRSF3	***0.14***	***54***						
KLK3 exons 1-2	***0.13***	***53***						
RPS10	***0.13***	***52***						
NLRP3	***0.13***	***51***						
RP11_97O12.7	***0.13***	***50***						
PPFIA2	***0.13***	***49***						
SMAP1 exons 7-8	***0.13***	***48***						
MAK	***0.13***	***47***						
AATF	***0.13***	***46***						
CDC20	***0.13***	***45***						
MXI1	***0.13***	***44***						
SSPO	***0.13***	***43***						
MEX3A	***0.13***	***42***						
MCM7	***0.12***	***41***						
PDLIM5	***0.12***	***40***						
OGT	***0.12***	***39***						
GOLM1	***0.12***	***38***						
MYOF	***0.12***	***37***						
VPS13A	***0.12***	***36***						
CASKIN1	***0.12***	***35***						
RPS11	***0.11***	***34***						
RIOK3	***0.11***	***33***						
B2M	***0.11***	***32***						
FDPS	***0.11***	***31***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
ANPEP	***0.11***	***30***						
CD10	***0.11***	***29***						
RPL18A	***0.11***	***28***						
ITPR1	***0.11***	***27***						
SEC61A1	***0.11***	***26***						
EIF2D	***0.11***	***25***						
TFDP1	***0.10***	***24***						
TWIST1	***0.10***	***23***						
MEMO1	***0.10***	***22***						
RPLP2	***0.10***	***21***						
HIST1H1E	***0.10***	***20***						
Met	***0.10***	***19***						
GABARAPL2	***0.10***	***18***						
AURKA	***0.10***	***17***						
MIATNB	***0.10***	***16***						
ALAS1	***0.09***	***15***						
PTN	***0.09***	***14***						
STEAP4	***0.09***	***13***						
GAPDH	***0.09***	***12***						
TERF2IP	***0.08***	***11***						
IFT57	***0.08***	***10***						
MIR4435_1HG	***0.08***	***9***						
TBP	***0.08***	***8***						
BRAF	***0.07***	***7***						
BTG2	***0.07***	***6***						
CAMK2N2	***0.07***	***5***						
ARHGEF25	***0.07***	***4***						
NEAT1	***0.06***	***3***						

All Transcripts (n = 166)			*Tanscripts identified by glm (n = 36)*			*Transcripts identified by polr (n = 20)*		
Transcript	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*	*Transcript*	*IncNodePurity*	*Rank*
PTPRC	***0.05***	***2***						
SIRT1	***0.05***	***1***						

Supplementary Table 48 Random Forest results for CB, low-risk, intermediate-risk and high-risk cancer trend using the *RPLP2* and *TWIST1* normalised data.

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
PCA3	***1.20***	***167***	***HOXC6***	***3.24***	***19***	***HOXC6***	***3.53***	***15***
HOXC6	***0.81***	***166***	***NAALADL2***	***2.78***	***18***	***NAALADL2***	***3.35***	***14***
CP	***0.79***	***165***	***ERG3' exons 6-7***	***2.75***	***17***	***TMPRSS2:ERG fusion***	***2.84***	***13***
PALM3	***0.70***	***164***	***PALM3***	***2.47***	***16***	***UPK2***	***2.76***	***12***
ERG3' exons 6-7	***0.67***	***163***	***UPK2***	***2.34***	***15***	***PALM3***	***2.75***	***11***
NAALADL2	***0.61***	***162***	***TMPRSS2:ERG fusion***	***2.28***	***14***	***ERG3' exons 6-7***	***2.64***	***10***
UPK2	***0.60***	***161***	***ST6GALNAC1***	***2.11***	***13***	***SIM2 long***	***2.53***	***9***
TMPRSS2:ERG fusion	***0.57***	***160***	***TMEM86A***	***2.06***	***12***	***TMEM86A***	***2.47***	***8***
OR52A2	***0.52***	***159***	***CADPS***	***2.00***	***11***	***TDRD***	***2.44***	***7***
SPINK1	***0.51***	***158***	***SIM2 long***	***1.97***	***10***	***ST6GALNAC1***	***2.43***	***6***
VAX2	***0.49***	***157***	***SERPINB5***	***1.79***	***9***	***EN2***	***2.30***	***5***
TDRD	***0.48***	***156***	***GJB1***	***1.76***	***8***	***SERPINB5***	***2.14***	***4***
CKAP2L	***0.44***	***155***	***TDRD***	***1.75***	***7***	***FOLH1***	***1.94***	***3***
CADPS	***0.43***	***154***	***LASS1***	***1.66***	***6***	***SLC43A1***	***1.85***	***2***
HPN	***0.42***	***153***	***CLIC2***	***1.47***	***5***	***MEX3A***	***1.68***	***1***
AMH	***0.42***	***152***	***SLC43A1***	***1.43***	***4***			
HMBS	***0.38***	***151***	***MSMB***	***1.40***	***3***			
APOC1	***0.37***	***150***	***FOLH1***	***1.22***	***2***			
PPAP2A	***0.37***	***149***	***MEX3A***	***1.18***	***1***			
TMEM47	***0.37***	***148***						
LASS1	***0.34***	***147***						
SIM2 long	***0.34***	***146***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
ST6GALNAC1	***0.33***	***145***						
ISX	***0.33***	***144***						
DLX1	***0.32***	***143***						
MAPK8IP2	***0.31***	***142***						
AR.ex9	***0.30***	***141***						
MKi67	***0.30***	***140***						
TMEM45B	***0.30***	***139***						
TMEM86A	***0.29***	***138***						
PTN	***0.29***	***137***						
TERT	***0.28***	***136***						
EN2	***0.28***	***135***						
B4GALNT4	***0.28***	***134***						
CAMKK2	***0.28***	***133***						
ERG5'	***0.27***	***132***						
IGFBP3	***0.27***	***131***						
GCNT1	***0.27***	***130***						
MMP11	***0.27***	***129***						
AGR2	***0.27***	***128***						
MFSD2A	***0.26***	***127***						
SFRP4	***0.26***	***126***						
NKAIN1	***0.26***	***125***						
MDK	***0.26***	***124***						
DNAH5	***0.26***	***123***						
Timp4	***0.25***	***122***						
SLC4A1.S	***0.24***	***121***						
SPON2	***0.24***	***120***						
GJB1	***0.24***	***119***						
KLK4	***0.24***	***118***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
PPP1R12B	***0.24***	***117***						
AMACR	***0.24***	***116***						
SLC43A1	***0.24***	***115***						
TMCC2	***0.24***	***114***						
HOXC4	***0.23***	***113***						
ANKRD34B	***0.23***	***112***						
SERPINB5	***0.23***	***111***						
SChLAP1	***0.23***	***110***						
SLC12A1	***0.23***	***109***						
MMP25	***0.23***	***108***						
CLU	***0.23***	***107***						
TWIST1	***0.23***	***106***						
MYOF	***0.22***	***105***						
Met	***0.22***	***104***						
MARCH5	***0.22***	***103***						
MIR146A	***0.22***	***102***						
FOLH1	***0.21***	***101***						
CCDC88B	***0.21***	***100***						
COL9A2	***0.21***	***99***						
HIST1H2BG	***0.20***	***98***						
MNX1	***0.20***	***97***						
PCSK6	***0.20***	***96***						
AATF	***0.20***	***95***						
SMIM1	***0.20***	***94***						
PDLIM5	***0.20***	***93***						
HPRT	***0.20***	***92***						
ACTR5	***0.19***	***91***						
KLK3 exons 2-3	***0.19***	***90***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
HIST1H2BF	***0.19***	***89***						
ZNF577	***0.19***	***88***						
SRSF3	***0.19***	***87***						
ANPEP	***0.19***	***86***						
CLIC2	***0.18***	***85***						
MAK	***0.18***	***84***						
RIOK3	***0.18***	***83***						
SIRT1	***0.18***	***82***						
SMAP1 exons 7-8	***0.18***	***81***						
VPS13A	***0.18***	***80***						
PPFIA2	***0.18***	***79***						
ERG3' exons 4-5	***0.17***	***78***						
IMPDH2	***0.17***	***77***						
IFT57	***0.17***	***76***						
GOLM1	***0.17***	***75***						
LBH	***0.17***	***74***						
TFDP1	***0.17***	***73***						
CDKN3	***0.17***	***72***						
ITGBL1	***0.17***	***71***						
RP11_97O12.7	***0.17***	***70***						
BTG2	***0.17***	***69***						
CACNA1D	***0.16***	***68***						
HIST1H1C	***0.16***	***67***						
MIC1	***0.16***	***66***						
CASKIN1	***0.16***	***65***						
CDC37L1	***0.16***	***64***						
PECI	***0.16***	***63***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
MMP26	***0.16***	***62***						
MCTP1	***0.16***	***61***						
MGAT5B	***0.15***	***60***						
HIST3H2A	***0.15***	***59***						
TRPM4	***0.15***	***58***						
HIST1H1E	***0.15***	***57***						
RNF157	***0.15***	***56***						
ARHGEF25	***0.15***	***55***						
SNORA20	***0.14***	***54***						
STEAP2	***0.14***	***53***						
MEX3A	***0.14***	***52***						
CD10	***0.14***	***51***						
RAB17	***0.14***	***50***						
MCM7	***0.14***	***49***						
PTPRC	***0.14***	***48***						
PSTPIP1	***0.14***	***47***						
SULF2	***0.14***	***46***						
SSTR1	***0.14***	***45***						
SACM1L	***0.14***	***44***						
RPLP2	***0.13***	***43***						
KLK3 exons 1-2	***0.13***	***42***						
KLK2	***0.13***	***41***						
P712P	***0.13***	***40***						
SIM2 short	***0.13***	***39***						
MSMB	***0.13***	***38***						
SEC61A1	***0.13***	***37***						
AR.ex4_8	***0.13***	***36***						
SULT1A1	***0.13***	***35***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
SSPO	***0.13***	***34***						
OGT	***0.13***	***33***						
ALAS1	***0.12***	***32***						
RPL23AP53	***0.12***	***31***						
STEAP4	***0.12***	***30***						
SYNM	***0.12***	***29***						
COL10A1	***0.12***	***28***						
AURKA	***0.12***	***27***						
ABCB9	***0.12***	***26***						
NEAT1	***0.12***	***25***						
PVT1	***0.12***	***24***						
RPS11	***0.12***	***23***						
DPP4	***0.12***	***22***						
SNCA	***0.12***	***21***						
CAMK2N2	***0.11***	***20***						
STOM	***0.10***	***19***						
RPL18A	***0.10***	***18***						
MED4	***0.10***	***17***						
GABARAPL2	***0.10***	***16***						
RPS10	***0.10***	***15***						
FDPS	***0.10***	***14***						
CDC20	***0.10***	***13***						
MXI1	***0.10***	***12***						
ITPR1	***0.09***	***11***						
TBP	***0.09***	***10***						
MIR4435_1HG	***0.09***	***9***						
TERF2IP	***0.09***	***8***						
BRAF	***0.09***	***7***						

All Transcripts (n = 167)			*Tanscripts identified by glm (n = 19)*			*Transcripts identified by polr (n = 15)*		
Transcript	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*	*Transcript*	*MeanDecreaseGini*	*Rank*
MIATNB	***0.09***	***6***						
NLRP3	***0.08***	***5***						
EIF2D	***0.08***	***4***						
GAPDH	***0.07***	***3***						
B2M	***0.07***	***2***						
MEMO1	***0.06***	***1***						

6.21 Cell vs EV fraction

Supplementary Table 49 The 129 transcripts that are significantly (post multiple testing correction) different between the cell and microvesicular fraction.

Transcript	*p-value*	*Adjusted p-value*	*Log2 Fold Change*
NEAT1	***1.77E-16***	***2.94E-14***	***-0.88***
PTPRC	***1.77E-16***	***2.94E-14***	***-1.97***
MMP25	***2.17E-16***	***3.58E-14***	***-1.43***
SULF2	***2.48E-16***	***4.07E-14***	***-1.69***
HIST1H1C	***2.83E-16***	***4.59E-14***	***0.25***
MCTP1	***2.83E-16***	***4.59E-14***	***-1.09***
IFT57	***2.93E-16***	***4.66E-14***	***0.43***
CCDC88B	***2.93E-16***	***4.66E-14***	***-1.27***
STOM	***2.93E-16***	***4.66E-14***	***-1.73***
MFSD2A	***3.24E-16***	***5.12E-14***	***-1.66***
B2M	***3.46E-16***	***5.44E-14***	***-0.36***
PSTPIP1	***3.96E-16***	***6.17E-14***	***-1.44***
APOC1	***6.72E-16***	***1.03E-13***	***-1.07***
NLRP3	***6.72E-16***	***1.03E-13***	***-1.64***
MIR4435_1HG	***6.94E-16***	***1.06E-13***	***-0.43***
MSMB	***7.18E-16***	***1.09E-13***	***0.29***
KLK2	***8.19E-16***	***1.24E-13***	***0.55***
AR.ex4_8	***9.33E-16***	***1.40E-13***	***0.51***
KLK3.ex2_3	***1.03E-15***	***1.53E-13***	***0.51***
KLK4	***1.10E-15***	***1.63E-13***	***0.41***
PTN	***1.17E-15***	***1.73E-13***	***0.76***
BTG2	***1.34E-15***	***1.95E-13***	***-0.34***
DPP4	***1.52E-15***	***2.20E-13***	***0.47***
CLIC2	***1.52E-15***	***2.20E-13***	***-1.35***
STEAP2	***1.57E-15***	***2.25E-13***	***0.52***
MIR146A	***1.74E-15***	***2.47E-13***	***-0.96***
PECI	***2.04E-15***	***2.88E-13***	***0.30***
IMPDH2	***2.65E-15***	***3.70E-13***	***0.30***
RPLP2	***3.77E-15***	***5.24E-13***	***0.11***
P712P	***6.29E-15***	***8.69E-13***	***0.67***
TWIST1	***8.38E-15***	***1.15E-12***	***0.34***
RP11_97O12.7	***8.65E-15***	***1.18E-12***	***0.22***
RPS11	***9.51E-15***	***1.27E-12***	***0.10***
TMEM86A	***9.51E-15***	***1.27E-12***	***-1.19***
MAK	***1.39E-14***	***1.85E-12***	***-1.15***
ZNF577	***1.43E-14***	***1.89E-12***	***0.38***
PPAP2A	***1.68E-14***	***2.20E-12***	***0.29***
HIST1H2BF	***2.16E-14***	***2.80E-12***	***0.24***
TERT	***2.22E-14***	***2.87E-12***	***0.51***
05-Mar	***3.04E-14***	***3.89E-12***	***0.26***
PCA3	***7.66E-14***	***9.73E-12***	***0.58***
SERPINB5	***9.78E-14***	***1.23E-11***	***0.79***
NKAIN1	***1.07E-13***	***1.34E-11***	***0.57***

Transcript	*p-value*	*Adjusted p-value*	*Log2 Fold Change*
SMIM1	***1.69E-13***	***2.09E-11***	***0.62***
SSPO	***1.85E-13***	***2.27E-11***	***0.38***
FOLH1	***2.28E-13***	***2.78E-11***	***0.51***
RPL18A	***2.57E-13***	***3.11E-11***	***0.17***
FDPS	***3.27E-13***	***3.92E-11***	***0.19***
AMACR	***3.37E-13***	***4.01E-11***	***0.50***
OR52A2	***3.57E-13***	***4.22E-11***	***0.77***
MIC1	***4.15E-13***	***4.85E-11***	***-0.67***
GABARAPL2	***6.86E-13***	***7.95E-11***	***0.15***
PDLIM5	***7.06E-13***	***8.12E-11***	***0.19***
RPS10	***7.27E-13***	***8.29E-11***	***0.14***
KLK3.ex1_2	***7.71E-13***	***8.71E-11***	***0.46***
PPFIA2	***2.07E-12***	***2.32E-10***	***0.65***
SEC61A1	***2.40E-12***	***2.66E-10***	***-0.46***
MNX1	***3.48E-12***	***3.82E-10***	***0.40***
CD10	***4.24E-12***	***4.63E-10***	***0.30***
NAALADL2	***5.18E-12***	***5.59E-10***	***0.38***
CAMK2N2	***6.68E-12***	***7.14E-10***	***0.53***
TFDP1	***9.35E-12***	***9.91E-10***	***0.18***
Met	***9.62E-12***	***1.01E-09***	***-0.88***
SIM2.long	***1.68E-11***	***1.74E-09***	***0.65***
COL10A1	***2.09E-11***	***2.15E-09***	***-0.59***
SSTR1	***3.16E-11***	***3.22E-09***	***0.24***
CP	***4.49E-11***	***4.54E-09***	***-0.99***
PCSK6	***6.05E-11***	***6.05E-09***	***0.40***
Timp4	***6.74E-11***	***6.67E-09***	***0.61***
VAX2	***1.09E-10***	***1.06E-08***	***0.36***
CACNA1D	***1.09E-10***	***1.06E-08***	***0.19***
HOXC6	***1.12E-10***	***1.07E-08***	***0.81***
SPON2	***2.40E-10***	***2.28E-08***	***0.34***
AMH	***2.60E-10***	***2.44E-08***	***0.30***
ARHGEF25	***6.56E-10***	***6.10E-08***	***0.58***
EIF2D	***6.90E-10***	***6.35E-08***	***0.12***
SChLAP1	***8.67E-10***	***7.89E-08***	***0.67***
GJB1	***1.01E-09***	***9.08E-08***	***0.49***
AURKA	***1.17E-09***	***1.04E-07***	***-0.35***
HIST3H2A	***2.08E-09***	***1.83E-07***	***0.37***
RAB17	***2.47E-09***	***2.15E-07***	***0.38***
HMBS	***2.80E-09***	***2.41E-07***	***0.25***
MKi67	***3.85E-09***	***3.27E-07***	***-1.18***
DNAH5	***5.02E-09***	***4.22E-07***	***0.57***
CKAP2L	***8.93E-09***	***7.41E-07***	***-0.50***
CASKIN1	***1.05E-08***	***8.65E-07***	***0.24***
SULT1A1	***1.08E-08***	***8.75E-07***	***-0.18***
MXI1	***1.57E-08***	***1.24E-06***	***0.13***
ITPR1	***1.57E-08***	***1.24E-06***	***-0.14***
MMP11	***1.94E-08***	***1.51E-06***	***0.30***
HPRT	***3.16E-08***	***2.43E-06***	***0.18***
SIM2.short	***3.30E-08***	***2.51E-06***	***0.35***
PALM3	***3.62E-08***	***2.72E-06***	***0.31***

Transcript	*p-value*	*Adjusted p-value*	*Log2 Fold Change*
AGR2	***4.06E-08***	***3.00E-06***	***0.32***
SYNM	***4.87E-08***	***3.55E-06***	***0.54***
MDK	***1.42E-07***	***1.02E-05***	***0.21***
EN2	***2.66E-07***	***1.89E-05***	***0.36***
MED4	***3.15E-07***	***2.21E-05***	***0.09***
RNF157	***3.58E-07***	***2.47E-05***	***0.58***
MGAT5B	***7.01E-07***	***4.76E-05***	***0.28***
LBH	***1.01E-06***	***6.80E-05***	***0.28***
IGFBP3	***1.06E-06***	***6.98E-05***	***-0.56***
TMEM45B	***1.30E-06***	***8.42E-05***	***-0.28***
HOXC4	***1.82E-06***	***0.0001***	***0.37***
CLU	***2.99E-06***	***0.0002***	***0.61***
SNCA	***2.99E-06***	***0.0002***	***0.15***
MYOF	***4.76E-06***	***0.0003***	***0.12***
CDC37L1	***5.24E-06***	***0.0003***	***0.11***
GOLM1	***7.66E-06***	***0.0005***	***0.39***
SACM1L	***1.11E-05***	***0.0006***	***0.11***
SFRP4	***1.27E-05***	***0.0007***	***0.35***
ERG3prime.ex4_5	***2.31E-05***	***0.001***	***0.64***
LASS1	***2.31E-05***	***0.001***	***-0.46***
B4GALNT4	***2.71E-05***	***0.001***	***-0.45***
MEX3A	***2.86E-05***	***0.002***	***0.39***
STEAP4	***4.06E-05***	***0.002***	***-0.11***
HPN	***4.66E-05***	***0.002***	***0.21***
MAPK8IP2	***4.74E-05***	***0.002***	***-0.45***
TRPM4	***7.91E-05***	***0.004***	***0.37***
ANPEP	***9.04E-05***	***0.004***	***-0.18***
TERF2IP	***9.50E-05***	***0.004***	***0.04***
SRSF3	***9.99E-05***	***0.005***	***-0.19***
HIST1H1E	***0.0002***	***0.007***	***0.09***
ANKRD34B	***0.0003***	***0.012***	***-0.35***
PPP1R12B	***0.0003***	***0.015***	***0.12***
HIST1H2BG	***0.0005***	***0.021***	***0.15***
CDC20	***0.0009***	***0.039***	***0.22***
TMEM47	***0.0011***	***0.043***	***0.61***
SMAP1.ex7_8	***0.0013***	***0.049***	***0.14***