A Novel Stratification Framework for Predicting Outcome in Patients with Prostate Cancer

Bogdan-Alexandru Luca[1,2,*], Vincent Moulton[2,#], Christopher Ellis[1,2], Dylan R Edwards[1], Colin Campbell[3], Rosalin A Cooper[4], Jeremy Clark[1], Daniel S Brewer[1,5,*,#], & Colin S Cooper[1,#,§]

[1] Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK.

[2] School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK.

[3] Intelligent Systems Laboratory, University of Bristol, Bristol, UK.

[4] Department of Pathology, University Hospital Southampton NHS Foundation Trust, Southampton, UK.

[5] The Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK.

[*] Contributed equally to this work.

[#] Jointly supervised this work.

[§] **Corresponding author**.

Professor Colin Cooper (colin.cooper@uea.ac.uk), University of East Anglia, Norwich Research Park, Norwich, NR4 7UG, UK

Running title: prostate cancer stratification

Abstract

Background

Unsupervised learning methods such as Hierarchical Cluster Analysis are commonly used for the analysis of genomic platform data. Unfortunately, such approaches ignore the well documented heterogeneous composition of prostate cancer samples. Our aim is to use more sophisticated analytical approaches to deconvolute the structure of prostate cancer transciptome data providing novel clinically actionable information for this disease.

Methods

We apply an unsupervised model called Latent Process Decomposition (LPD), which can handle heterogeneity within individual cancer samples, to genome-wide expression data from eight prostate cancer clinical series including 1,785 malignant samples with the clinical endpoints of PSA failure and metastasis.

Results

We show that PSA failure is correlated with the level of an expression signature called DESNT (HR = 1.52, 95% CI = [1.36, 1.7], $P = 9.0 \times 10^{-14}$, Cox model) and that patients with a majority DESNT signature have an increased metastatic risk (X^2-test, $P = 0.0017$, and $P = 0.0019$). Additionally, we develop a stratification framework that incorporates DESNT and identifies three novel molecular subtypes of prostate cancer.

Conclusions

These results highlight the importance of using more complex approaches for the analysis of genomic data, may assist drug targeting, and have allowed the construction of a nomogram combining DESNT with other clinical factors for use in clinical management.

Background

Driven by technological advances and decreased costs, a plethora of genomic datasets now exist. This is illustrated by the availability of expression data from over 1.3 million samples from the Gene Expression Omnibus [1] and DNA sequence data on 25,000 cases from the International Cancer Genome Consortium [2]. Such datasets have been used as the raw material for the discovery of disease sub-classes using a variety of mathematical approaches. Hierarchical clustering, k-means clustering, and self-organising maps have been applied to expression datasets leading, for example, to the discovery of five molecular breast cancer types (Basal, Luminal A, Luminal B, ERBB2-overexpressing, and Normal-like) [3]. The inherent shortcoming of this type of approach is the implicit assumption of sample assignment to a particular cluster or group. Such analyses are in complete contrast to the well documented heterogeneous composition of most individual cancer samples [4,5].

Unsupervised analysis of prostate cancer transcriptome profiles using the above approaches have failed to identify robust disease categories that have distinct clinical outcomes [6,7]. Noting that prostate cancer samples derived from genome wide studies frequently harbour multiple cancer lineages, and can have intra-tumour variations in genetic compositions [8–10], we applied an unsupervised learning method called Latent Process Decomposition (LPD) [11] that can take into account the issue of heterogeneity of composition within individual cancer samples. By heterogeneity we mean that an individual cancer sample can be made up of several different components that each has distinct properties We had previously used Latent Process Decomposition: (i) to confirm the presence of the basal and

ERBB2 overexpressing subtypes in breast cancer transcriptome datasets [12]; (ii) to demonstrate that data from the MammaPrint breast cancer recurrence assay would be optimally analyzed using four separate prognostic categories [12]; and (iii) to show that patients with advanced prostate cancer can be stratified into two clinically distinct categories based on expression profiles in blood [13]. LPD (closely related to Latent Dirichlet Allocation) is a mixed membership model in which the expression profile for a single cancer is represented as a combination of underlying latent (hidden) signatures. Each latent signature has a representative gene expression pattern. A given sample can be represented over a number of these underlying functional states, or just one such state. The appropriate number of signatures to use is determined using the LPD algorithm by maximising the probability of the model given the data.

The application of LPD to prostate cancer transcriptome datasets led to the discovery of an expression pattern, called DESNT, that was observed in all datasets examined [14]. Cancer samples were assigned as DESNT when this pattern was more common than any other signature, and this designation was associated with poor outcomes independently of other clinical parameters including Gleason, Clinical stage and PSA. In the current paper we test whether the presence of even a small proportion of the DESNT cancer signature confers poor outcome and use LPD to develop a new prostate cancer stratification framework.

Materials and Methods

Transcriptome datasets

Eight publically available transcriptome microarray datasets derived from prostatectomy samples from men with prostate cancer were used and are referred to as: Memorial Sloan Kettering Cancer Centre (MSKCC) [7], CancerMap [14], CamCap [6], Stephenson [15], TCGA

[16], Klein [17], Erho [18], and Karnes [19]. There were 1785 samples from primary malignant tissue and 173 from normal tissue (Table 1). MSKCC also had data from 19 metastatic cancer samples. The CamCap dataset was produced by combining two Illumina HumanHT-12 V4.0 expression beadchip datasets (GEO: GSE70768 and GSE70769) obtained from two prostatectomy series (Cambridge and Stockholm) [6]. The original CamCap [6] and CancerMap [14] datasets have 40 patients in common and thus 20 of the common samples were excluded at random from each dataset. Each Affymetrix Exon microarray dataset was normalised using the RMA algorithm [20] implemented in the Affymetrix Expression Console software. For CamCap and Stephenson previous normalised values were used. For the TCGA dataset, the counts per gene previously calculated were used [16] and transformed using the variance stabilising transformation implemented in the DESeq2 package[21]. For the CamCap and CancerMap datasets the ERG gene alterations had been scored by fluorescence in situ hybridization [6,14]. Only probes corresponding to genes measured by all platforms were retained. The ComBat algorithm from the sva R package and quantile transformation, was used to mitigate series-specific effects. Flow diagrams presenting each of the analyses performed in the current study, with the datasets used, are shown in the Supplementary Materials. The ethical approvals obtained for each dataset are listed in the original publications.

Latent Process Decomposition (LPD)

LPD [11,12] is an unsupervised Bayesian approach which breaks down (decomposes) each sample into component sub-elements (signatures). Each signature is a representative gene expression pattern. LPD is able to classify complex data based on the relative representation of these signatures in each sample. LPD can objectively assess the most likely number of signatures. We assessed the hold-out validation log-likelihood of the data computed at

various number of signatures and used a combination of both the uniform (equivalent to a maximum likelihood approach) and non-uniform (missed approach point) priors to choose the number of signatures. For input, each dataset was reduced to probes that detect the 500 genes with the greatest variance across the MSKCC dataset. For robustness, LPD is run 100 times with different seeds, for each dataset. Out of the 100 runs we selected the run with the survival log-rank *p*-value closest to the mode as a representative run that was used for subsequent analysis.

OAS-LPD (One Added Sample LPD)

The OAS-LPD algorithm is a modified version of the LPD algorithm in which new sample(s) are decomposed into LPD signatures, without retraining the model (i.e. without re-estimating the model parameters μ_{gk}, σ^2_{gk}, and α in Rogers et al. [11]). Only the variational parameters Q_{kga} and γ_{ak}, corresponding to the new sample(s), are iteratively updated until convergence, according to Eq. (6) and Eq. (7) from Rogers et al. 2005 [11]. LPD as presented by Rogers et al. [11] was first applied to the MSKCC dataset of 131 cancer and 29 normal samples, as described above. The model parameters μ_{gk}, σ^2_{gk}, and α, corresponding to the representative LPD run, were then used to classify additional expression profiles from all datasets, one sample at a time. A detailed description is provided in the Supplementary Methods.

Statistical tests

All statistical tests were performed in R version 3.3.1. For characterisation of signatures, each sample was assigned to the signature that had the largest gamma (γ) value for that sample.

Correlations

Pearson correlations between the expression profiles between the MSKCC and CancerMap were calculated for each of the eight signatures: (i) for each gene we select one corresponding probe at random; (ii) for each probe we transformed its distribution across all samples to a standard normal distribution; (iii) the mean expression for each gene across the samples assigned to signature *j* (gene subgroup mean) in each dataset was determined; (iv) the Pearson's correlation between the gene subgroup mean expression profile in MSKCC vs the gene subgroup mean expression profile in CancerMap is calculated for each signature.

Differentially expressed and methylated features

Differentially expressed probesets were identified for each signature using a moderated *t*-test implemented in the limma R package (Benjamin-Hocberg false discovery rate < 0.05, differentially expressed in at least 50/100 runs; samples assigned to the signature vs the rest).

Thus differential methylation was assigned at the probe level. Hypo and hypermethylated genes that are predictive of transcription were identified using the methylMix R package (functionally differentially methylated in at least 50/100 runs) using genes that are found to be differentially expressed in that signature as input. Datasets where there were <10 samples assigned to a signature were removed from the identification of intersection genes for that signature.

Survival analyses and nomogram

Survival analyses were performed using Cox proportional hazards models, the log-rank test, and Kaplan-Meier estimator, with biochemical recurrence after prostatectomy as the end point. For nomogram construction, the Cox proportional hazards model was fitted on the meta-dataset obtained by combining MSKCC, CancerMap, and Stephenson datasets, and validated on CamCap, using the rms R package. The Gleason grade was divided into <7, 3+4, 4+3, >7, the pathological stage in T1-T2 vs T3-T4, while DESNT percentage and PSA

were considered continuous covariates. The missing values for the predictors were imputed using the flexible additive models with predictive mean matching, implemented in the Hmisc R package. The linearity of the continuous covariates was assessed using the Martingale residuals [22]. The lack of collinearity between covariates was determined by calculating the variance inflation factors (VIF) (VIF values between 1.04 and 3.01) [23]. All covariates met the Cox proportional hazards assumption, as determined by the Schoenfeld residuals. The internal validation and calibration of the Cox model were performed by bootstrapping the training dataset 1,000 times. The calibration of the model was estimated by comparing the predicted and observed survival probabilities at five years. For comparing the discrimination accuracy of two non-nested Cox models the U-statistic calculated by the Hmisc rcorrp.cens function was used.

Detecting over-representation of genomic features

Mutated cancer genes identified by the Cancer Genome Atlas Research Network (2015) [16] were examined at the sample level. The under-/over-representation of these features in samples assigned to a particular LPD signature was determined using the χ^2 independence test.

Pathway over-representation analysis and signature correlation analysis

The GO biological process annotations were tested for over-representation (or under-representation) in the lists of differentially expressed genes in each signature, using clusterProfiler version 3.4.4. For a given pathway and a given sample the pathway activation score was calculated as indicated in Levine et al. [24]. Using the complete combined dataset of all 8 datasets, Z-scores were calculated for each sample for each of the 17,697 MSigDB v6.0 gene sets. These were correlated with DESNT γ values, and the top 20 sets with the highest absolute Pearson's correlation were selected. The resulting *p*-values from pathway over-representation analysis were adjusted for multiple testing using the false discovery rate.

Results

Presence of DESNT signature as a continuous variable is associated with poor clinical outcome

In our previous studies, LPD detected between three and eight underlying signatures (also called processes) in expression microarray datasets collected from prostate cancer samples after prostatectomy [14]. Decomposition of the MSKCC dataset [7] gave eight signatures [14]. Fig. 1a illustrates the proportion of the DESNT expression signature identified in each MSKCC sample, with individual cancer samples being assigned as a "DESNT cancer" when the DESNT signature was the most abundant as shown in Fig. 1a and Fig. 1c. Based on PSA failure, patients with DESNT cancer always exhibited poorer outcome relative to other cancer samples in the same dataset [14]. The implication is that it is the presence of regions of cancer containing the DESNT signature conferred poor outcome. If this idea is correct, we would predict that cancer samples containing a smaller contribution of DESNT signature, such as those shown in Fig. 1b for the MSKCC dataset, should also exhibit poorer outcome.

To increase the power to test this prediction we combined transcriptome data from the MSKCC [7], CancerMap [14], Stephenson [15], and CamCap [6] studies (n=503). There was a significant association with PSA recurrence when the proportion of expression assigned to the DESNT signature was treated as a continuous variable (HR = 1.52, 95% CI=[1.36, 1.7], $P = 9.0 \times 10^{-14}$, Cox proportional hazard regression model). Outcome became worse as the proportion of DESNT signature increased. For illustrative purposes cancer samples were divided into four groups based on the proportion of DESNT, with 47.4% of cancer samples contained at least some DESNT cancer (proportion greater than 0.001; Fig. 2a). PSA failure

free survival at 60 months is 82.5%, 67.4%, 59.5% and 44.9% for the proportion of DESNT signature being <0.001, 0.001 to 0.3, 0.3 to 0.6, and >0.6, respectively (Fig. 2b).

Nomogram for DESNT predicting PSA failure

The proportion of DESNT cancer was combined with other clinical variables (Gleason grade, PSA levels, pathological stage, and the surgical margins status) in a Cox proportional hazards model and fitted to a combined dataset of 318 cancer samples (MSKCC , CancerMap , and Stephenson); CamCap cancer samples (n=185) were used for external validation. The proportion of DESNT was an independent predictor of worse clinical outcome (HR = 1.33, 95% CI=[1.14, 1.56], $P = 3.0x10^{-4}$,) along with Gleason grade=4+3 (HR = 2.43, 95% CI=[1.10, 5.37], $P = 2.7x10^{-2}$), Gleason grade>7 (HR = 5.05, 95% CI=[2.35, 10.89], $P < 1x10^{-4}$), and positive surgical margins (HR = 1.65, 95% CI=[1.07, 2.56], $P = 2.2x10^{-2}$) (Fig. S1: Supplementary Figure 1). PSA level and pathological stage were below the threshold of statistical significance ($P = 0.09$, HR = 1.14, 95% CI=[0.97, 1.34]) and (P = 0.055, HR = 1.51, 95% CI=[0.99, 2.31]) respectively. At internal validation, the Cox model obtained a 1,000 bootstrap-corrected C-index of 0.747, and at external validation a C-index of 0.795. Using this model, a nomogram was constructed for use of DESNT cancer information in conjunction with clinical variables to predict the risk of biochemical recurrence at one, three, five, and, seven years following prostatectomy (Fig. 2c, Fig. S1).

LPD algorithm for detecting the presence of DESNT cancer in individual samples

The ability of LPD to detect structure is likely to be dependent on sample size, cohort composition, disease severity range and data quality. We observed optimal decompositions varying between three and eight underlying signatures in different datasets [14]. When we examined the two datasets that had an optimal eight underlying signatures (MSKCC and

CancerMap) we noted a striking relationship: based on correlations of expression profiles; all eight of the LPD signatures appeared to be common (Fig. S2; $R^2 > 0.5$). To provide a more consistent classification framework where the number of classes did not vary between datasets, we therefore used the MSKCC dataset and its decomposition into eight distinct signatures as a reference for identifying categories of prostate cancer type.

LPD is a computer intensive procedure and analyses can take days to run on a high-performance computing cluster. This would restrict ease of DESNT detection for clinical implementation. We therefore developed a variant of LPD called One Added Sample-LPD (OAS-LPD), where data from a single additional cancer sample could be decomposed into signatures, following normalisation, without repeating the entire LPD procedure. LPD model parameters [11] were first derived by decomposition of the MSKCC dataset into eight signatures. These signature parameters were then used as a framework for decomposition of additional data from single samples, selected in this case from a dataset, or in future from a patient undergoing assessment in the clinic. To test this procedure, we applied OAS-LPD individually to cancer samples from MSKCC, CancerMap, Stephenson, and CamCap (Fig. S3) and repeated Cox regression analysis and nomogram construction. Proportion of DESNT ($P = 0.0011$, HR = 1.53, 95% CI=[1.19, 1.98]), Gleason=4+3 ($P = 0.0061$, HR = 2.83, 95% CI=[1.35, 5.96]), Gleason>7 ($P < 1x10^{-4}$, HR = 5.39, 95% CI=[2.54, 11.44]) and surgical margin status (P = 0.0015, HR = 2.00, 95% CI=[1.30, 3.07]) remained independent predictors of clinical outcome (Fig. S4). Notably the performance of the Cox model (internal validation C-index=0.742; external validation C-index=0.786) was not significantly different to that of the original separate dataset Cox model (train dataset Z=-0.65, two-tailed $P = 0.52$; validation dataset $Z = 0.89$, two-tailed $P = 0.38$; U-statistic) and the nomogram (Fig. S5) had almost an identical presentation of parameters to that shown in Fig. 2c. This observation is consistent

the high degree of correlation between LPD and OAS-LPD DESNT gamma values across the MSKCC, CancerMap, Stephenson, and CamCap datasets $P = 2.39 \times 10^{-110}$)

New categories of prostate cancer

We wished to determine whether LPD signatures were characterized by particular clinical or molecular features indicating that they represented distinct categories of prostate cancer. OAS-LPD using the MSKCC derived model of gene signatures was applied to all datasets (n=1958, Table 1) and each sample was assigned to the signature that was the most abundant. Samples from non-cancerous (benign) prostate tissue were more frequently assigned to LPD2, LPD4, and LPD8 than to the other groups ($P < 0.05$, χ^2 test, Fig. S3, Table S1). When datasets with linked clinical data were combined (MSKCC, CancerMap, Stephenson, CamCap, Fig. 3a-c) primary cancers assigned to DESNT had worse outcome ($P = 3.4 \times 10^{-14}$, log-rank test, DESNT assigned samples vs the rest) while those assigned to LPD4 had improved outcome ($P = 0.0081$, log-rank test, LPD4 assigned samples vs the rest) as judged by PSA failure. Cancer samples with ERG-alterations assigned to signature LPD3 also exhibited better outcome ($P < 0.05$; log-rank test, comparison to all other ETS positive cancer samples) in all three datasets where ERG status was available (Fig. 4b-d).

To gain information about the new LPD categories we examined the distribution of genetic alterations in the decomposition of the TGCA dataset [16] (Fig. 4a), LPD3 cancer samples had over-representation of ETS and *PTEN* gene alterations, and under-representation of *CDH1* and *SPOP* gene alterations ($P < 0.05$, χ^2 test, Table 2). LPD5 cancer samples exhibited exactly the reverse pattern of genetic alteration: there was under-repression of ETS and *PTEN* gene alterations and over-representation of *SPOP* and *CHD1* alterations (Table 2). The statistically different distribution of ETS-gene alterations in samples assigned to LPD3, and

LPD5, observed in the TGCA dataset were confirmed in the CamCap and CancerMap dataset (Table 2). In summary we have identified three additional prostate cancer categories that have altered genetic and/or clinical associations: LPD3, LPD4, and LPD5 (Fig. 5) and that may be relevant for drug targeting.

Altered patterns of gene expression and DNA methylation

We examined samples assigned to each OAS-LPD signature for genes with significantly altered expression levels in all eight datasets ($P < 0.05$ after FDR correction, samples in LPD group vs all other LPD categories from the same dataset, Supplementary data 1). LPD3 cancers samples exhibited seven commonly overexpressed genes including ERG, GHR, and HDAC1. Pathway analysis suggested the involvement of Stat3 gene signalling (Fig. S6a, Supplementary data 2). LPD5 exhibited 47 significantly overexpressed gene and 13 under-expressed genes. Many of the genes had established roles in fatty acid metabolism and the control of secretion (Fig. S6b). LPD6-cancers and LPD8 cancers had failed to exhibit statistically significant changes in genetic alteration or clinical outcome in the current study but did have characteristic altered patterns of gene expression (Fig. S6c,e). The five genes commonly overexpressed in LPD6 cancers suggested involvement in metal ion homeostasis. 30 genes were overexpressed and 36 genes under expressed in in LPD8 cancers including several genes involved in extracellular matrix organisation. Cross referencing differential methylation data available for the TCGA dataset with gene associated with each LPD group indicated that many expression changes may be explained, at least in part, by changes in DNA methylation (Fig. 5, Fig. S7, Supplementary data 3).

DESNT as a signature of metastasis

The MSKCC study includes data from 19 metastatic cancer samples. For each metastatic sample, DESNT was the most abundant signature when OAS-LPD was applied (Fig. 3d). Two of the studied datasets (MSKCC and Erho) had publicly available annotations indicating that the patients from which primary cancer expression profiles were examined had progressed to develop metastasis after prostatectomy (Fig. S3). From nine cancer patients developing metastasis in the MSKCC dataset five occurred from samples in which the DESNT signature is most common (X^2-test, $P = 0.0017$) and of 212 cancer patients developing metastases in the Erho dataset 50 were from DESNT cancers (X^2-test, $P = 0.0019$) (Fig. S8). From these studies we concluded that DESNT cancers have an increased risk of developing metastasis, consistent with the higher risk of PSA failure. For the Erho dataset, membership of LPD1 was associated with lower risk of metastasis (X^2-test, $P = 0.026$, Fig. S8).

To further investigate the underlying nature of DESNT cancer we used the transcriptome profile for each primary prostate cancer sample to investigate associations with the 17,697 signatures and pathways annotated in the MSigDB database. The top 20 signatures where expression was associated with proportion of DESNT are shown in Table S2. The third most significant correlation was to genes downregulated in metastatic prostate cancer. This resulting data gives additional clues to the underlying biology of DESNT cancer including associations with genes altered in ductal breast cancer, in stem cells and during FGFR1 signaling.

Discussion

We have confirmed a key prediction of the DESNT cancer model by demonstrating that the presence of a small proportion of the DESNT cancer signature confers poorer outcome. The

proportion of DESNT signature can be considered a continuous variable such that as DESNT cancer content increases outcome became worse. This observation led to the development of nomograms for estimating PSA failure at three years, five years, and seven years following prostatectomy. The result provides an extension of previous studies in which nomograms incorporating Gleason score, Stage, and PSA value have been used to predict outcome following surgery [25].

The match between the eight underlying signatures detected for the MSKCC and CancerMap datasets was used as the basis for developing a novel classification framework for prostate cancer. A new algorithm called OAS-LPD was developed to allow rapid assessment of the presence of the signatures in individual cancer samples. In total four clinically and or genetically distinct subgroups were identified (DESNT, LPD3, LPD4, and LPD5, Fig. 5). The functional significance of the new disease groupings, for example in determining drug sensitivity, remains to be established. However, with the use of OAS-LPD it will be possible to undertake assessments of the response of patients in each of the groups DESNT, LPD3, LPD4, and LPD5 to drug treatments. There is limited overlap between the new classification and previously proposed subgroups based on genetic alterations [16,26–29].

Multiplatform data (expression, mutation, and methylation data from each cancer sample) are available for many cancer types, for example from The Cancer Genome Atlas. This has prompted the development of additional methods for sub-class discovery that can combine information from different platforms including the copula mixed model [30], Bayesian consensus clustering [31] and the iCluster model [16]. These approaches can suffer from the problem of sample assignment to a particular cluster or group, and the failure to take into

consideration the heterogeneous composition of individual cancer samples. These observations highlight the need to develop methods similar to LPD that can be applied to multiplatform data.

An important issue for patients diagnosed with prostate cancer is that clinical outcome is highly variable and precise prediction of the course of disease progression at the time of diagnosis is not possible [32]. In some studies, the use of population PSA screening can reduce mortality from prostate cancer by up to 21% [33]. However many, if not most, prostate cancers that are currently detected by PSA screening are clinically insignificant [34]. Over-diagnosis of clinically insignificant prostate cancer is a major issue and is set to increase still further [35]. There is therefore an urgent need for the identification of cancer categories that are associated with clinically aggressive or indolent disease to allow the targeting of radical therapies to the men that need them. For breast cancer unsupervised hierarchical clustering of transcriptome data resulted in a classification system that is routinely used to guide the management and treatment of this disease. Here we established a novel classification framework for the analysis of prostate cancer that has its origins in unsupervised analyses of transcriptome data. In future studies we plan to analyse the utility of DESNT and other LPD processes (particularly LPD3, LPD4 and LPD5) in managing prostate cancer patients, including predicting response to drug treatment. This will be performed through the assessment of LPD status in the contexts of established clinical trials. For evaluation we would plan to use each LPD assignment (eg DESNT, LPD3, LPD4 and LPD5) as a continuous variable as illustrated here by the development of a nomograms for the use of DESNT in predicting PSA failure. In conclusion our results highlight the

importance of devising and using more sophisticated approaches for the analysis of genomic datasets from all biological systems.

Additional Information

Acknowledgements

The research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia. We thank Shea Connell for useful comments and suggestions on the manuscript.

Authors' Contributions

CSC, DSB, and VM were involved in funding acquisition and supervised the project. CSC, DSB, B-AL, VM were involved in conceptualisation and planned the data analysis. B-AL, CE, DSB performed the majority of the analyses and investigations, with additional analysis and insight provided by VM, DRE, CC, RAC, JC, and CSC. B-AL, CE, CC, DSB, and CSC were involved in developing the methodology for the project. CSC, DSB, and B-AL wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All data were from other publications. The ethical approvals obtained for each dataset are listed in the original publications.

Consent for publication

All authors consent to publication.

Data availability

The datasets analysed during the current study are available (Table 1). The majority are available from the Gene Expression Omnibus repository:

- MSKCC[7] : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21034
- CancerMap[14] : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE94767
- Klein[17] : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62667
- CamCap[6] : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70768 and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70769
- Erho[18] : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46691
- Karnes[19] : https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62116
- Stephenson[15] : Data available from the corresponding author of this paper.
- TCGA[16]: Data available from the TCGA Data Portal https://portal.gdc.cancer.gov/projects/TCGA-PRAD

Competing Interests

C.S.C., D.S.B., B-A L., and V.M. are co-inventors on a patent application from the University of East Anglia on the detection of DESNT prostate cancer.

Funding

This work was funded by the Bob Champion Cancer Trust, The Masonic Charitable Foundation successor to The Grand Charity, The King Family, The Stephen Hargrave Foundation, and The University of East Anglia. We acknowledge support from Movember, from Prostate Cancer UK, Callum Barton, and from The Andy Ripley Memorial Fund.

References

1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.

2. Consortium ICG, Anderson W, Artez A, Bell C, Bernabé RR, Bhan MK, et al. International network of cancer genome projects. Nature. 2010;464(7291):993–8.
3. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003 Jul 8;100(14):8418–23.
4. Blanco-Calvo M, Concha Á, Figueroa A, Garrido F, Valladares-Ayerbes M. Colorectal cancer classification and cell heterogeneity: A systems oncology approach. Int J Mol Sci. 2015;16(6):13610–32.
5. Polyak K. Heterogeneity in breast cancer Review series introduction Heterogeneity in breast cancer. J Clin Invest. 2011;121(10):3786.
6. Ross-Adams H, Lamb ADD, Dunning MJJ, Halim S, Lindberg J, Massie CMM, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. EBioMedicine. 2015;2(9):1133–44.
7. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative Genomic Profiling of Human Prostate Cancer. Cancer Cell. 2010 Jun 13;18(1):11–22.
8. Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat Genet. 2015 Mar 2;47(4):367–72.
9. Boutros PC, Fraser M, Harding NJ, de Borja R, Trudel D, Lalonde E, et al. Spatial genomic heterogeneity within localized, multifocal prostate cancer. Nat Genet. 2015 May 25;47(7):736–45.
10. Tsourlakis M-C, Stender A, Quaas A, Kluth M, Wittmer C, Haese A, et al. Heterogeneity of ERG expression in prostate cancer: a large section mapping study of

entire prostatectomy specimens from 125 patients. BMC Cancer. 2016;16(1):641.

11. Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray data sets. IEEE/ACM Trans Comput Biol Bioinforma. 2005;2(2):143–56.

12. Carrivick L, Rogers S, Clark J, Campbell C, Girolami M, Cooper C. Identification of prognostic signatures in breast cancer microarray data using Bayesian techniques. J R Soc Interface. 2006;3(8):367–81.

13. Olmos D, Brewer D, Clark J, Danila DC, Parker C, Attard G, et al. Prognostic value of blood mRNA expression signatures in castration-resistant prostate cancer: a prospective, two-stage study. Lancet Oncol. 2012 Oct 8;2045(12):1–11.

14. Luca B, Brewer DS, Edwards DR, Edwards S, Whitaker HC, Merson S, et al. DESNT: A Poor Prognosis Category of Human Prostate Cancer. Eur Urol Focus. 2018 Dec 6;4(6):842–50.

15. Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE, Scardino PT, et al. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer. 2005;104(2):290–8.

16. Network CGAR, Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015 Nov 5;163(4):1011–25.

17. Klein EA, Yousefi K, Haddad Z, Choeurng V, Buerki C, Stephenson AJ, et al. A genomic classifier improves prediction of metastatic disease within 5 years after surgery in node-negative high-risk prostate cancer patients managed by radical prostatectomy without adjuvant therapy. Eur Urol. 2015;67(4):778–86.

18. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PLoS One. 2013;8(6):e66855.

19. Karnes RJ, Bergstralh EJ, Davicioni E, Ghadessi M, Buerki C, Mitra AP, et al. Validation of a Genomic Classifier that Predicts Metastasis Following Radical Prostatectomy in an At Risk Patient Population. J Urol. 2013;190(6):2047–53.

20. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.

21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

22. Therneau TM, GRAMBSCH PM, Fleming TR. Martingale-based residuals for survival models. Biometrika. 1990 Mar 1;77(1):147–60.

23. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. Multivariate data analysis. 1998.

24. Levine DM, Haynor DR, Castle JC, Stepaniants SB, Pellegrini M, Mao M, et al. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol. 2006;7(10):R93.

25. Shariat SF, Kattan MW, Vickers AJ, Karakiewicz PI, Scardino PT. Critical review of prostate cancer predictive tools. Future Oncol. 2009;5(10):1555–84.

26. Attard G, Clark J, Ambroisine L, Fisher G, Kovacs G, Flohr P, et al. Duplication of the fusion of TMPRSS2 to ERG sequences identifies fatal human prostate cancer. Oncogene. 2008;27(3):253–63.

27. Reid AHM, Attard G, Ambroisine L, Fisher G, Kovacs G, Brewer D, et al. Molecular characterisation of ERG, ETV1 and PTEN gene loci identifies patients at low and high risk of death from prostate cancer. Br J Cancer. 2010;102(4):678–84.

28. Mosquera JM, Beltran H, Park K, MacDonald TY, Robinson BD, Tagawa ST, et al. Concurrent AURKA and MYCN Gene Amplifications Are Harbingers of Lethal

TreatmentRelated Neuroendocrine Prostate Cancer. Neoplasia. 2013;15(1):1-IN4.

29. Rodrigues LU, Rider L, Nieto C, Romero L, Karimpour-Fard A, Loda M, et al. Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. Cancer Res. 2015 Mar 15;75(6):1021–34.

30. Rey M, Roth V. Copula Mixture Model for Dependency-seeking Clustering. 2012;

31. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013;29(20):2610–6.

32. Buyyounouski MK, Pickles T, Kestin LL, Allison R, Williams SG. Validating the Interval to Biochemical Failure for the Identification of Potentially Lethal Prostate Cancer. J Clin Oncol. 2016;30(15):1857–63.

33. Schröder FH, Hugosson J, Roobol MJ, Tammela TLJ, Zappa M, Nelen V, et al. Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. Lancet. 2014;384(9959):2027–35.

34. Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. Ann Intern Med. 2013 Jun 4;158(11):831–8.

35. Parker C, Emberton M. Screening for prostate cancer appears to work, but at what cost? BJU Int. 2009;104(3):290–2.

Legends

Figure 1. LPD decomposition of the MSKCC dataset. (a) DESNT bar chart from the LPD decomposition of the MSKCC dataset [14]. showing the number ID assigned to 23 examples samples that had some amount of DESNT signature. (b,c) Pie charts showing the relative proportions of the eight LPD signatures in 23 example samples. DESNT is in red, other LPD signatures are represented by different colours as indicated in the key. The number next to each pie chart indicates which cancer it represents from the bar chart above. Individual cancer samples were assigned as a "DESNT cancer" when the DESNT signature was the most abundant; examples are shown in the right-hand box (c, 'DESNT'). Many other cancer samples contained a smaller proportion of DESNT cancer and were associated with a poor outcome: examples shown in larger box (b, 'SOME DESNT').

Figure 2. Stratification of prostate cancer samples based on the percentage of DESNT cancer present. For these analyses the data from the MSKCC, CancerMap, CamCap, and Stephenson datasets were combined (n=503). (a) Plot showing the proportion of DESNT signature in each cancer sample and the division into four groups of increasing DESNT. Group 1 samples have a proportion of less than 0.001 of the DESNT signature. (b) Kaplan-Meier plot showing the Biochemical Recurrence (BCR) free survival based on proportion of DESNT cancer present as determined by LPD. Number of cancer patients in each Group are indicated (bottom right) and the number of PCR failures in each group are show in parentheses. The definition of Groups 1-4 is shown in Fig. 2a. Cancer samples with proportions up to 0.3 DESNT (Group 2) exhibited poorer clinical outcome (X^2-test, $P = 0.011$) compared to cancer samples lacking DESNT (<0.001). Cancer samples with the intermediate (0.3 to 0.6) and high (>0.6) proportions of DESNT also exhibited significantly worse outcome (respectively $P = 2.6 \times 10^{-5}$ and $P = 8.3 \times 10^{-9}$ compare to cancer samples

lacking DESNT. The combined log-rank $P = 1.3x10^{-8}$). (c) Nomogram model developed to predict PSA free survival at one, three, five, and seven years using proportion of DESNT. Assessing a single patient each clinical variable has a corresponding point score (top scales). The point scores for each variable are added to produce a total points score for each patient. The predicted probability of PSA free survival at one, three, five, and seven years can be determined by drawing a vertical line from the total points score to the probability scales below.

Figure 3. Prediction of clinical outcome according to OAS-LPD group. (a-c) Kaplan-Meier plots showing PSA free survival outcomes for the cancer patients assigned to LPD groups in analyses of the combine MSKCC, CancerMap, CamCap, and Stephenson datasets: (a) comparison of all LPD groups (LPD7 is DESNT); (b) cancer patients assigned to LPD4 compared to patients assigned to all other LPD groups; (c) cancer patients assigned to DESNT (LPD7) compared to cancers assigned to all other LPD groups. (d) OAS-LPD signature assignment proportions for the 19 metastatic tissue samples reported as part of the MSKCC dataset. In all cases DESNT (LPD7) was the dominant expression signature detected.

Figure 4. (a) OAS-LPD sub-groups in The Cancer Genome Atlas Dataset (n=333). Cancer samples were assigned to subgroups based on the most prominent signature as detected by OAS-LPD. The types of genetic alteration are shown for each gene (mutations, fusions, deletions, and over-expression). Clinical parameters including biochemical recurrence (BCR) are represented at the bottom together with groups for iCluster, methylation, somatic copy number alteration (SVNA), and messenger RNA (mRNA)[16]. Comparison of the frequency of genetic alterations present in each subgroup are shown in Table 2. (b-d) Kaplan-Meier

plots showing PSA free survival outcomes for ETS-rearrangement positive cancers in LPD3 compared to all other ETS-positive cancers for the CancerMap, CamCap, and TCGA datasets.

Figure 5. A classification framework for prostate cancer. Based on the analyses of genetic and clinical correlations we consider that there is good evidence for the existence of LPD3, LPD4, and LPD5 as separate cancer categories, moderate evidence of the existence of LPD6 and LPD8 (based on alteration of expression only), and weak evidence for LPD1. The methylation column list all genes that exhibit differential expression and that also contain at least one locus that is differentially methylated.

Table 1

Dataset	Primary	Normal	Type	Platform	Citation
MSKCC [7]	131	29	FF	Affymetrix Exon 1.0 ST v2	Taylor *et al.* 2010
CancerMap [14]	137	17	FF	Affymetrix Exon 1.0 ST v2	Luca *et al.* 2017
Stephenson [15]	78	11	FF	Affymetrix U133A	Stephenson *et al.* 2005
Klein [17]	182	0	FFPE	Affymetrix Exon 1.0 ST v2	Klein *et al.* 2015
CamCap [6]	147	73	FF	Illumina HT12 v4.0 BeadChip	Ross-Adams *et al.* 2015
TCGA [16]	333	43	FF	Illumina HiSeq 2000 RNA-Seq v2	TCGA network 2015
Erho [18]	545	0	FFPE	Affymetrix Exon 1.0 ST v2	Erho *et al.* 2013
Karnes [19]	232	0	FFPE	Affymetrix Exon 1.0 ST v2	Karnes *et al*. 2013

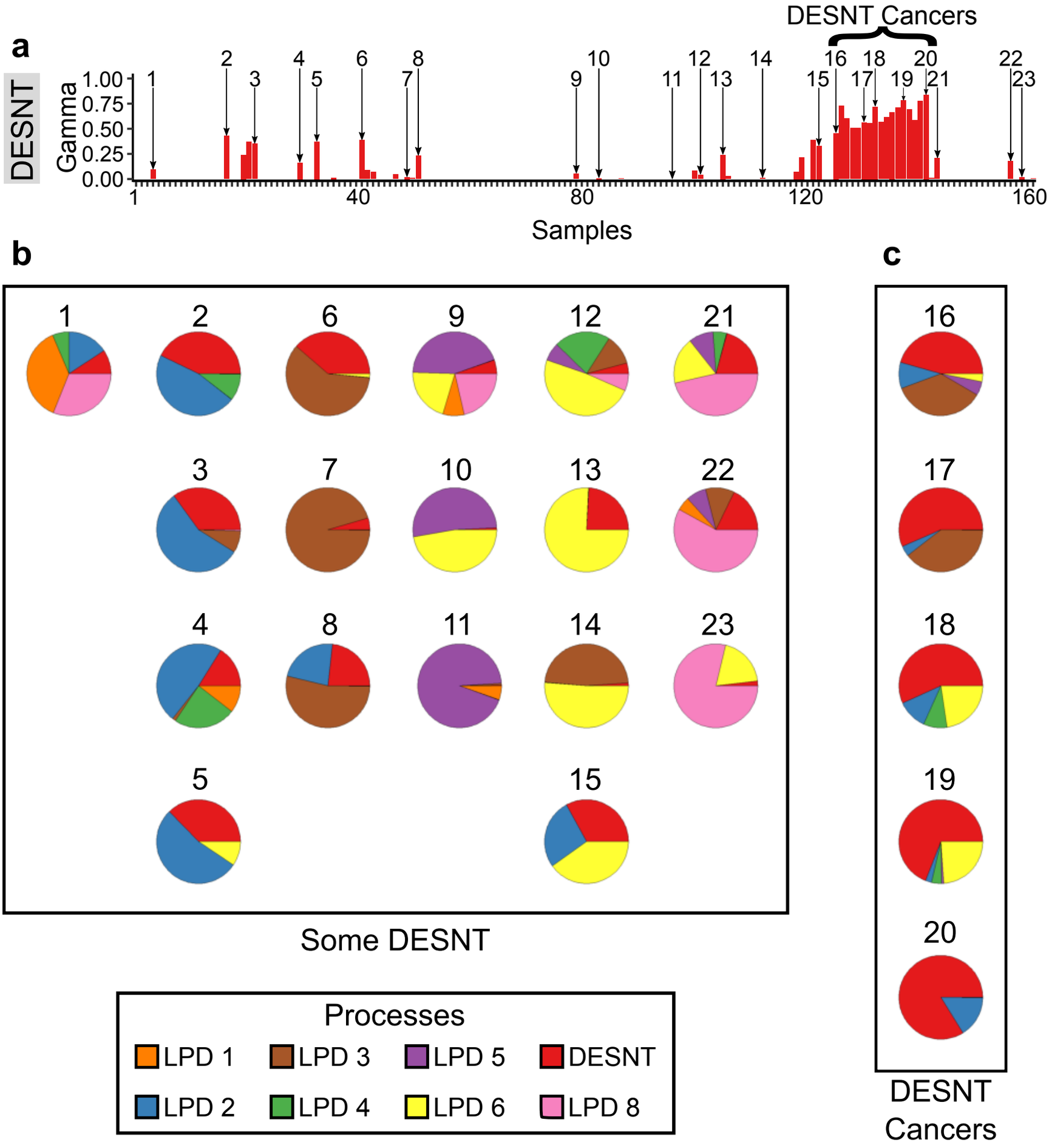
Table 1. Transcriptome datasets. The MSKCC study additionally reported expression profiles from 19 metastatic cancers. The ethical approvals obtained for each dataset are listed in the original publications.

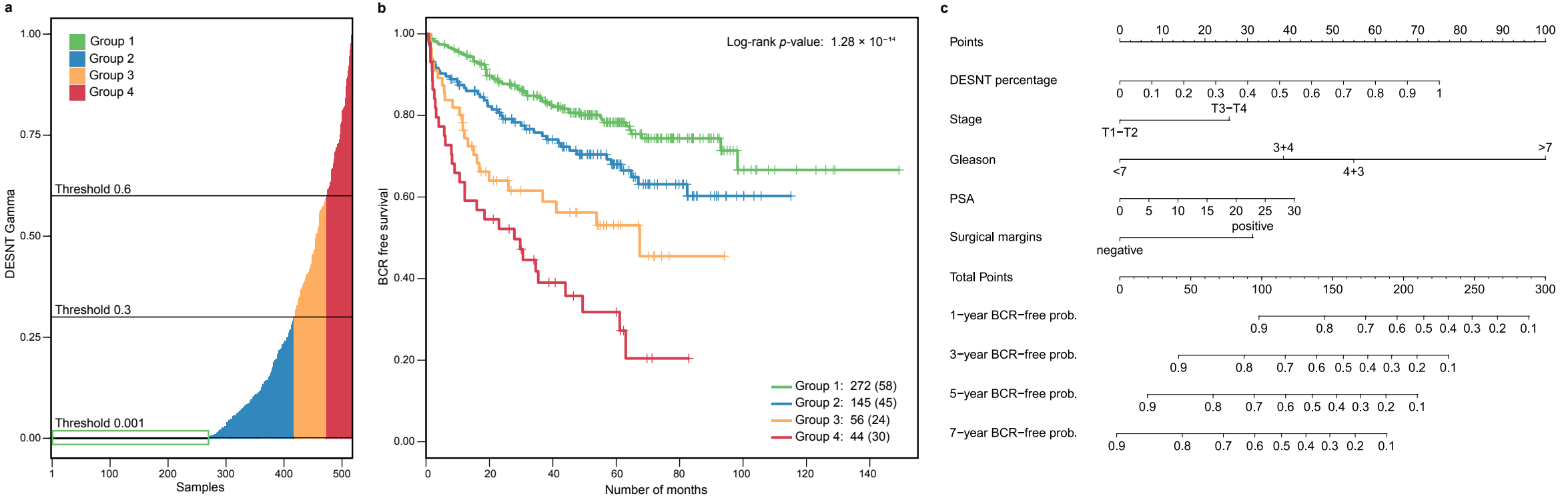
Table 2

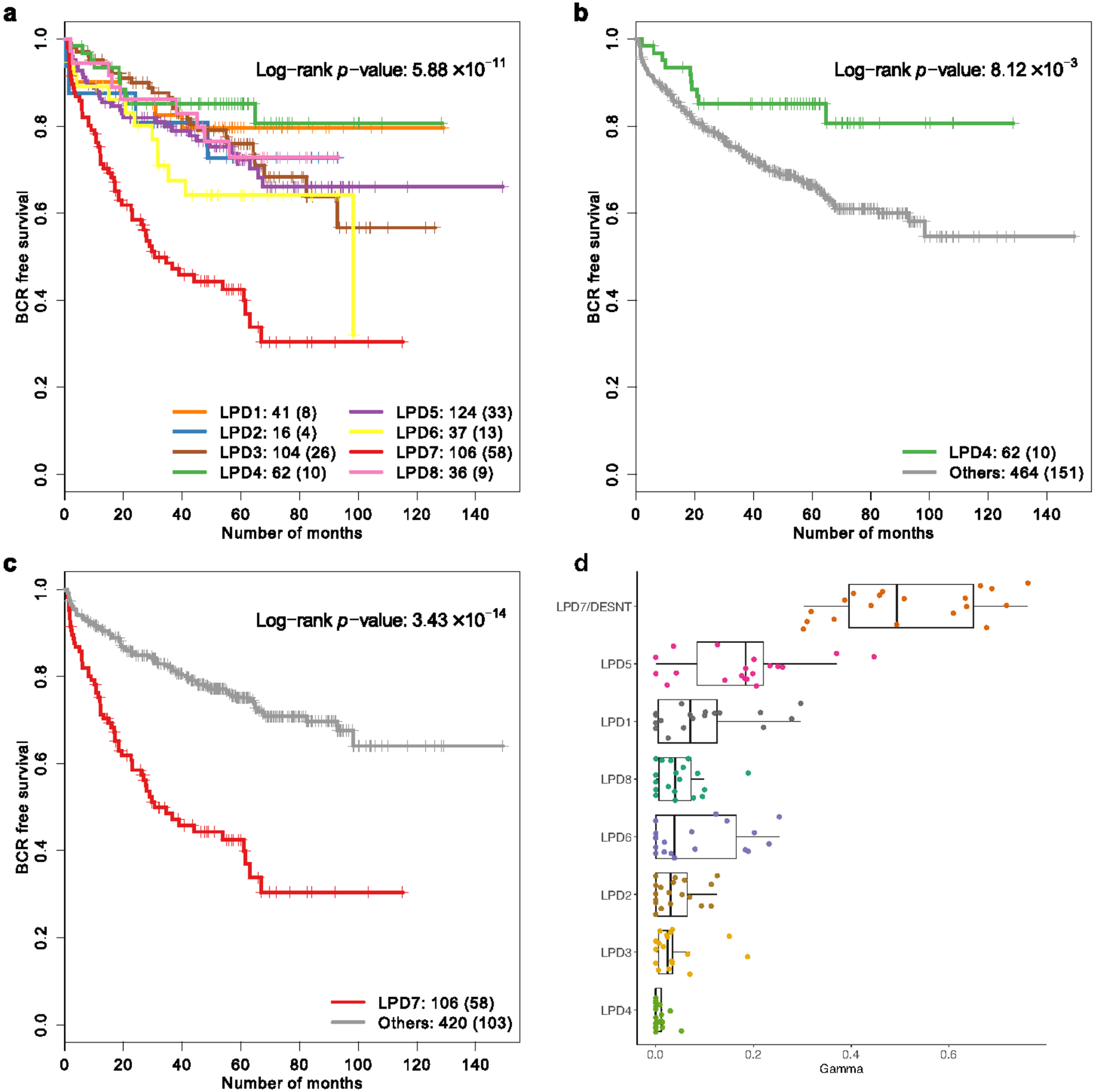
	TCGA			CancerMap			CamCap		
	ETS-	ETS+	χ^2 *P*-val	*ERG*-	*ERG*+	χ^2 *P*-val	*ERG*-	*ERG*+	χ^2 *P*-val
LPD1	8	3	0.0588	13	4	0.0851	0	3	0.235
LPD2	4	8	0.827	3	3	1	0	2	0.467
LPD3	9	67	1.45×10^{-08}	5	15	0.00977	4	17	0.00299
LPD4	14	21	1	14	15	0.619	1	2	0.987
LPD5	65	5	2.20×10^{-16}	19	1	0.000180	34	0	1.15×10^{-11}
LPD6	13	22	0.802	5	5	1	2	4	0.657
DESNT	13	66	1.17×10^{-06}	6	15	0.0207	9	24	0.00274
LPD8	9	6	0.193	8	4	0.540	4	1	0.371

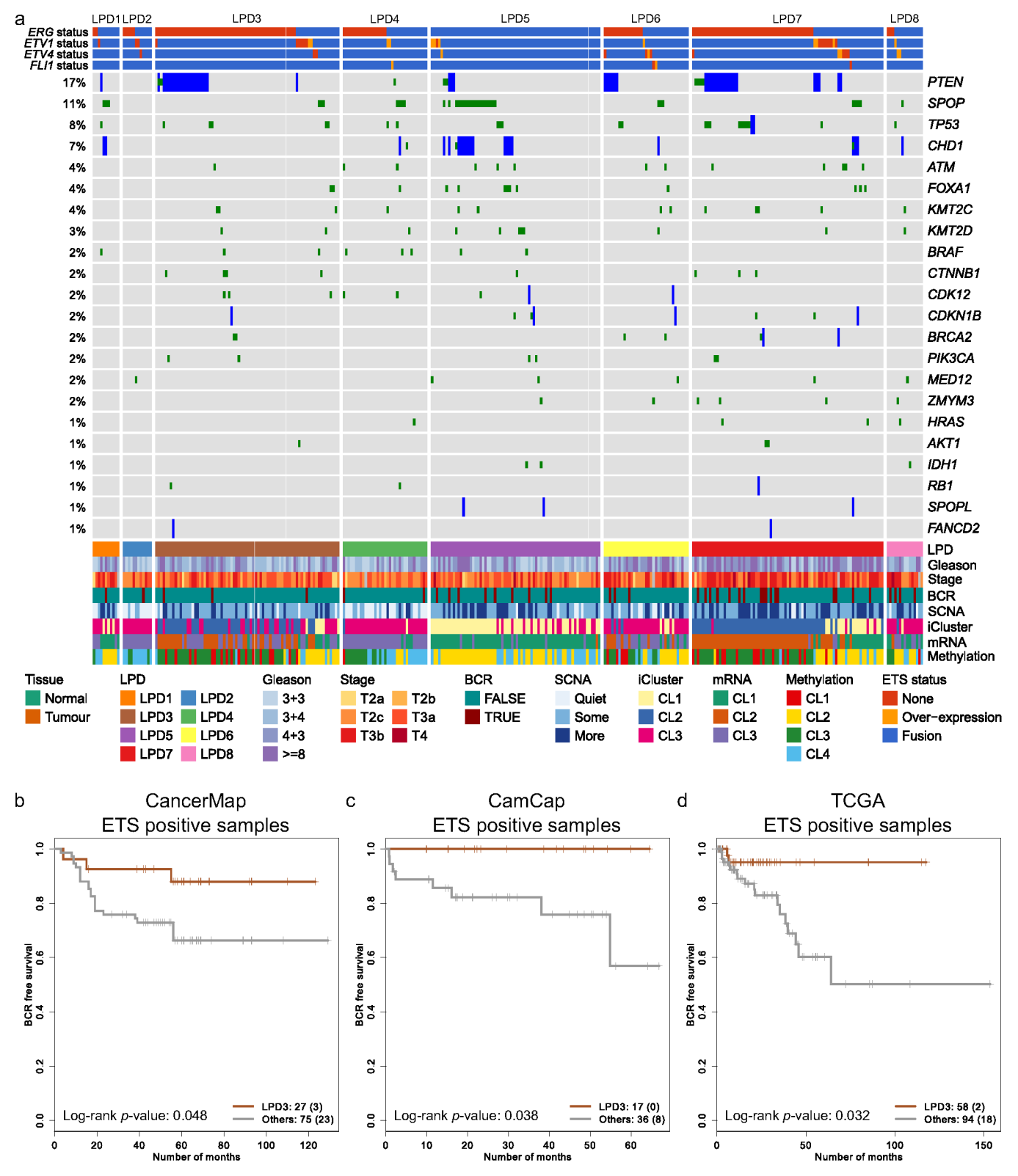
	PTEN			*SPOP*			*CHD1*		
	Non-homdel	Homdel	χ^2 *P*-val	Non-mut	Mut	χ^2 *P*-val	Non-homdel	Homdel	χ^2 *P*-val
LPD1	10	1	0.896	8	3	0.213	9	2	0.309
LPD2	12	0	0.284	12	0	0.436	12	0	0.756
LPD3	55	21	0.000894	73	3	0.0400	76	0	0.0211
LPD4	35	0	0.0174	31	4	1	34	1	0.603
LPD5	67	3	0.00830	51	19	4.46×10^{-06}	57	13	7.69×10^{-06}
LPD6	29	6	0.903	32	3	0.825	34	1	0.603
DESNT	60	19	0.0167	75	4	0.0795	76	3	0.432
LPD8	15	0	0.195	14	1	0.889	14	1	1

Table 2. Correlation of OAS-LPD subgroups with genetic alterations in The Cancer Genome Atlas Dataset. Statistically significant differences are highlighted in grey.


a
DESNT
Gamma
1.00
0.75
0.50
0.25
0.00
DESNT Cancers
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
1
40
80
120
160
Samples
b
1 2 6 9 12 21
3 7 10 13 22
4 8 11 14 23
5 15
Some DESNT
c
16
17
18
19
20
DESNT Cancers
Processes
LPD 1
LPD 2
LPD 3
LPD 4
LPD 5
LPD 6
DESNT
LPD 8


a
Group 1
Group 2
Group 3
Group 4
DESNT Gamma
Threshold 0.6
Threshold 0.3
Threshold 0.001
Samples
b
Log-rank p-value: 1.28 × 10⁻¹⁴
BCR free survival
Group 1: 272 (58)
Group 2: 145 (45)
Group 3: 56 (24)
Group 4: 44 (30)
Number of months
c
Points
DESNT percentage
Stage
T1−T2
T3−T4
Gleason
<7
3+4
4+3
>7
PSA
positive
Surgical margins
negative
Total Points
1−year BCR−free prob.
3−year BCR−free prob.
5−year BCR−free prob.
7−year BCR−free prob.


a
Log−rank p−value: 5.88 ×10−11
BCR free survival
Number of months
LPD1: 41 (8)
LPD2: 16 (4)
LPD3: 104 (26)
LPD4: 62 (10)
LPD5: 124 (33)
LPD6: 37 (13)
LPD7: 106 (58)
LPD8: 36 (9)
b
Log−rank p−value: 8.12 ×10−3
BCR free survival
Number of months
LPD4: 62 (10)
Others: 464 (151)
c
Log−rank p−value: 3.43 ×10−14
BCR free survival
Number of months
LPD7: 106 (58)
Others: 420 (103)
d
LPD7/DESNT
LPD5
LPD1
LPD8
LPD6
LPD2
LPD3
LPD4
Gamma


a
LPD1 LPD2 LPD3 LPD4 LPD5 LPD6 LPD7 LPD8
ERG status
ETV1 status
ETV4 status
FLI1 status
17% PTEN
11% SPOP
8% TP53
7% CHD1
4% ATM
4% FOXA1
4% KMT2C
3% KMT2D
2% BRAF
2% CTNNB1
2% CDK12
2% CDKN1B
2% BRCA2
2% PIK3CA
2% MED12
2% ZMYM3
1% HRAS
1% AKT1
1% IDH1
1% RB1
1% SPOPL
1% FANCD2
LPD
Gleason
Stage
BCR
SCNA
iCluster
mRNA
Methylation
Tissue
Normal
Tumour
LPD
LPD1 LPD2 LPD3 LPD4 LPD5 LPD6 LPD7 LPD8
Gleason
3+3 3+4 4+3 >=8
Stage
T2a T2b T2c T3a T3b T4
BCR
FALSE TRUE
SCNA
Quiet Some More
iCluster
CL1 CL2 CL3
mRNA
CL1 CL2 CL3
Methylation
CL1 CL2 CL3 CL4
ETS status
None Over−expression Fusion
b
CancerMap
ETS positive samples
BCR free survival
Log-rank p-value: 0.048
LPD3: 27 (3)
Others: 75 (23)
Number of months
c
CamCap
ETS positive samples
BCR free survival
Log-rank p-value: 0.038
LPD3: 17 (0)
Others: 36 (8)
Number of months
d
TCGA
ETS positive samples
BCR free survival
Log-rank p-value: 0.032
LPD3: 58 (2)
Others: 94 (18)
Number of months

	CLINICAL	GENE MUTATIONS	EXPRESSION	METHYLATION
LPD1	Less metastases in Erho *et al.* dataset.	No current evidence.		No current evidence.
LPD2 – NORMAL LIKE	Frequently contains normal samples. (Over-represented in 4/5 datasets)	No current evidence	▲2 genes ▼0 genes ▲Structural molecule activity ▲Protein-glutamine gammaglutamyl transferase activity	*KRT13, TGM4*
LPD3 – STAT SIGNALING	*ERG*+ cancers have better outcome. (2/3 datasets)	▲*ERG*/ETS ▲*PTEN* ▼*SPOP* ▼*CHD1*	▲7 genes ▼0 genes ▲Regulation of STAT proteins ▲Regulation of insulin secretion ▲Second-messenger-mediated signalling	*CSGALNACT1, ERG, GHR, GUCY1A3, HDAC1, ITPR3, PLA2G7*
LPD4 – IMPROVED SURVIVAL	Improved survival based on PSA failure. Frequently contains normal samples. (Over-represented in 2/5 datasets)	▼PTEN		No current evidence.
LPD5 – SPOP/CHD1	No current evidence.	▼*ERG*/ETS ▼*PTEN* ▲*SPOP* ▲*CHD1*	▲43 genes ▼13 genes ▲Fatty acid metabolism ▲Branched-chain aminoacid metabolism ▲Positive regulation of secretion	*ABHD2, ACAD8, ACLY, ALCAM, ALDH6A1, ALOX15B, ARHGEF7, AUH, BBS4, C1orf115, CAMKK2, COG5, CPEB3, CYP2J2, DHRS3, DHX32, EHHADH, ELOVL2, ERG, EXTL2, F3, FAM111A, GATA3, GLUD1, GNMT, HES1, HPGD, KHDRBS3, LAMB2, LAMC2, MIPEP, MON1B, NANS, NAT1, NCAPD3, PDE8B, PPFIBP2, PTK7, PTPN13, PTPRM, RAB27A, REPS2, RFX3, SCIN, SLC1A1, SLC4A4, SMPDL3A, SORL1, STXBP6, SYTL2, TBPL1, TFF3, TRIM29, TUBB2A, YIPF1, ZNF516*
LPD6 – TRANSITION METAL HOMEOSTASIS	No current evidence.	No current evidence.	▲9 genes ▼0 genes ▲Transition metal ion homeostasis ▲Viral genome replication ▲Humoral immune response	*CCL2, CFB, CFTR, CXCL2, IFI16, LCN2, LTF, LXN, TFRC*
LPD7 – DESNT POOR PROGNOSIS	Independent predictor of poor outcome based on PSA failure. More metastases in Erho *et al.* and MSKCC datasets. Dominant signature in metastatic cancers.	▲*ERG*/ETS ▲*PTEN*	▲2 genes ▼49 genes ▼Cell-substrate adhesion ▼Muscle contraction ▼Cell junction organisation	*ACTG2, ACTN1, ADAMTS1, ANPEP, ARMCX1, AZGP1, C7, CD44, CHRDL1, CNN1, CRISPLD2, CSRP1, CYP27A1, CYR61, DES, EGR1, ETS2, F5, FBLN1, FERMT2, FHL2, FLNA, FXYD6, FZD7, ITGA5, ITM2C, JAM3, JUN, KHDRBS3, LMOD1, LPHN2, MT1M, MYH11, MYL9, NFIL3, PARM1, PCP4, PDK4, PLAGL1, RAB27A, SERPINF1, SNAI2, SORBS1, SPARCL1, SPOCK3, SYNM, TAGLN, TCEAL2, TGFB3, TPM2, VCL*
LPD8 – NORMAL LIKE	Contains normal samples. (Over-represented in 5/5 datasets)	No current evidence.	▲29 genes ▼37 genes ▲Extracellular matrix organisation ▲Extracellular structure organisation	*ABCC4, ACAT2, ARHGEF6, ATP8A1, AXL, CANT1, CD83, CDH1, COL15A1, DCXR, DHCR24, DHRS7, DPYSL3, EPB41L3, FAM174B, FAM189A2, FBN1, FCHSD2, FHL1, FKBP4, FOXA1, FXYD5, GNAO1, GOLM1, GPX3, GTF3C1, HPN, IFI16, IRAK3, ITGA5, KIF5C, KLK3, LAPTM5, MAP7, MBOAT2, MFAP4, MFGE8, MIOS, MLPH, MMP2, MYO5C, NEDD4L, PART1, PARVA, PDIA5, PIGH, PLEKHO1, PLSCR4, PMEPA1, PRSS8, RFTN1, SAMD4A, SAMSN1, SEC23B, SERPINF1, SLC43A1, SPDEF, SPINT2, STEAP4, TMPRSS2, TRPM8, TSPAN1, VCAM1, WIPF1, XBP1, ZYX*