DESNT: a Poor Prognosis Category of Human Prostate Cancer

Bogdan-Alexandru Luca[a,b,§], Daniel S Brewer[b,c,§,¶,*], Dylan R Edwards[2], Sandra Edward[d], Hayley C Whitaker[e], Sue Merson[d], Nening Dennis[d], Rosalin A Cooper[f], Steven Hazell[g], Anne Y Warren[h], The CancerMap Group[i], Rosalind Eeles[d,g], Andy G Lynch[e], Helen Ross-Adams[e], Alastair D Lamb[e,j], David E Neal[e,j], Krishna Sethia[k], Robert D Mills[k], Richard Y Ball[l], Helen Curley[b], Jeremy Clark[b], Vincent Moulton[a,¶,*], Colin S Cooper[b,¶,*]

[a]*School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, UK;* [b]*Norwich Medical School, University of East Anglia, Norwich Research Park, Norwich, UK;* [c]*The Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK;* [d]*Division of Genetics and Epidemiology, The Institute Of Cancer Research, Sutton, UK;* [e]*Urological Research Laboratory, Cancer Research UK Cambridge Research Institute, University of Cambridge, Cambridge, UK;* [f]*Department of Pathology, University Hospital Southampton NHS Foundation Trust, Southampton, UK;* [g]*Royal Marsden NHS Foundation Trust, London and Sutton, UK;* [h]*Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK;* [i] *A list of participants and their affiliations appears in the Supplemental Information;* [j]*Department of Surgical Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK;* [k]*Department of Urology, Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, UK;* [l]*Department of Histopathology, Norfolk and Norwich University Hospitals NHS Foundation Trust, Norwich, UK*.

[§]These authors contributed equally to this work;
[¶]These authors jointly supervised this work.
*Corresponding Authors. Professor Vincent Moulton (V.Moulton@uea.ac.uk) , Dr Daniel Brewer (Dan.Brewer@uea.ac.uk) and Professor Colin Cooper (Colin.Cooper@uea.ac.uk) University of East Anglia, Norwich Research Park, Norwich, NR4 7UG, UK

Abstract

Background: A critical problem in the clinical management of prostate cancer is that it is highly heterogeneous. Accurate prediction of individual cancer behaviour is therefore not achievable at the time of diagnosis leading to substantial overtreatment. It remains an enigma that, in contrast to breast cancer, unsupervised analyses of global expression profiles has not currently defined robust categories of prostate cancer with distinct clinical outcomes.

Objective: To devise a novel classification framework for human prostate cancer based on unsupervised mathematical approaches.

Design, Setting, and Participants: Our analyses are based on the hypothesis that previous attempts to classify prostate cancer have been unsuccessful because individual samples of prostate cancer frequently have heterogeneous compositions. To address this issue we applied an unsupervised Bayesian procedure called Latent Process Decomposition to four independent prostate cancer transcriptome datasets obtained using samples from prostatectomy patients and containing between 78 and 182 participants.

Outcome Measurements and Statistical Analysis: Biochemical failure was assessed using log-rank analysis and Cox regression analysis.

Results and Limitations: Application of LPD identified a common process in all four independent datasets examined. Cancers assigned to this process (designated DESNT cancers) are characterized by low expression of a core set of 45 genes, many encoding proteins involved in the cytoskeleton machinery, ion transport and cell adhesion. For the three datasets with linked PSA failure data following prostatectomy, patients with DESNT cancer exhibited poor outcome relative to other patients ($P = 2.65 \times 10^{-5}$, $P = 4.28 \times 10^{-5}$, and $P = 2.98 \times 10^{-8}$). When these three datasets were combined the independent predictive value of DESNT membership was $P = 1.61 \times 10^{-7}$ compared to $P = 1.00 \times 10^{-5}$ for Gleason sum. A limitation of the study is that only prediction of PSA failure was examined.

Conclusions: Our results demonstrate the existence of a novel poor prognosis category of human prostate cancer and will assist in the targeting of therapy, helping avoid treatment-associated morbidity in men with indolent disease.

Patient Summary: Prostate cancer, unlike breast cancer, does not have a robust classification framework. We propose that this failure has occurred because prostate cancer samples selected for analysis frequently have heterozygous compositions (individual samples are made up of many different parts that each have different characteristics). Applying a mathematical approach that can overcome this problem we identify a novel poor prognosis category of human prostate cancer called DESNT.

Keywords: poor prognosis category; novel prostate cancer classification; DESNT prostate cancer; Latent Process Decomposition

Words 3383

1. Introduction

Risk categories based on PSA, Gleason score and Clinical Stage that predict PSA failure[1] underpin the treatment of localized prostate cancer, as illustrated, for example, by the UK National Institute for Health and Care Excellence guidelines[2]. Attempts to improved risk stratification have been made by the development of prognostic tests, such as Prolaris[3], Oncotype DX[4] and Decipher[5]. Most such expression-based prognostic signatures for prostate cancer have in common that they were derived using supervised steps, involving either comparisons of aggressive and non-aggressive disease[5,6] or the selection of genes representing specific biological functions[3,7,8]. Alternatively expression biomarkers may be linked to the presence of somatic copy number variations[9]. In contrast, for breast cancer, unsupervised analysis of trancriptome profiles, using approaches such as hierarchical clustering has identified robust disease categories that have distinct clinical outcomes and that require different treatment strategies[10].

Our hypothesis is that completely unsupervised classification of prostate cancer based on transcriptome data has not been successful previously[9,11] because individual samples of prostate cancer can contain more than one contributing lineage[12,13] and frequently have heterogeneous compositions[14-16]. To test this idea, in the current study, we applied Latent Process Decomposition[17,18] (LPD). Based on the latent Dirichlet allocation method[19], LPD assesses the structure of a dataset in the absence of knowledge of clinical outcome or biological role[17]. In contrast to standard unsupervised clustering models (e.g. *k*-means and hierarchical clustering), individual cancers are not assigned to a single cluster: instead gene expression levels in each cancer are modeled via combinations of latent processes. We previously used LPD to confirm the presence of basal and *ERBB2* overexpressing categories in breast cancer datasets[17], and to show that, based on blood expression profiles, patients with advanced prostate cancer can be stratified into two clinically distinct groups[20].

2. Materials and Methods

2.1 The CancerMap dataset

Fresh prostate cancer specimens were obtained and processed from a systematic series of patients who had undergone a prostatectomy at the Royal Marsden NHS Foundation Trust and Addenbrooke's Hospital, Cambridge as previously described[9,21,22]. The relevant local Research Ethics Committee approved was obtained. Expression profiles were determined and data was processed as previously described[22] using 1.0 Human Exon ST arrays (Affymetrix, Santa Clara, CA, USA) according to the manufacturer's instructions. Data are available from the Gene Expression Omnibus: GSE (data to be released on publication). CancerMap patients did not receive neo-adjuvant treatment.

2.2 Additional Transcriptome Datasets

We analysed five prostate cancer microarray datasets that will be referred to as: MSKCC, CancerMap, CamCap, Stephenson and Klein. The data used, platforms and location of clinical data are presented in Fig. 1b. Each dataset was obtained using samples from prostatectomy patients. CamCap dataset used in our study was produced combining Illumina HumanHT-12 V4.0 expression beadchip (bead microarray) datasets (GEO: GSE70768 and GSE70769) obtained from two prostatectomy series (Cambridge and Stockholm) and consisted of 147 cancer and 73 normal samples[9]. The CamCap and CancerMap datasets have in common 40 patients and thus are not independent. One RNAseq dataset consisting of 333 prostate cancers from The Cancer Genome Atlas was analysed which we refer to as TCGA[13]. The counts per gene supplied by TCGA were used.

2.3 Latent Process Decomposition

Latent process decomposition (LPD) [17,18], an unsupervised Bayesian approach, was used to classify samples into subgroups called processes. We selected the 500 probesets with greatest variance across the MSKCC dataset for use in LPD. These probesets map to 492 genes. For each dataset all probesets that map to these genes were used in LPD analyses (CancerMap: 507 probesets, CamCap:483, Stephenson: 609).

LPD can objectively assess the most likely number of processes. We assessed the hold-out validation log-likelihood of the data computed at various number of processes and used a combination of both the uniform (equivalent to a maximum likelihood approach) and non-uniform (MAP approach) priors to choose the number of processes. For robustness, we restarted LPD 100 times with different seeds, for each dataset. Out of the 100 runs we selected a representative run that was used for subsequent analysis. The representative run, was the run with the survival log-rank *p*-value closest to the mode. For the Klein dataset, for which we do not have clinical data, we used the hold-out log-likelihood from LPD instead.

2.4 Statistical Tests

All statistical tests were performed in R version 3.2.2 (https://www.r-project.org/). Correlations between the expression profiles between two datasets for a particular gene set and sample subgroup were calculated as follows:

1. For each gene we select one probeset at random;
2. for each probeset we transformed its distribution across all samples to a standard normal distribution;
3. the average expression for each probeset across the samples in the subgroup is determined, to obtain an expression profile for the subgroup.
4. the Pearson's correlation between the expression profiles of the subgroups in the two datasets is determined.

Differentially expressed probesets were identified using a moderated *t*-test implemented in the *limma* R package[23]. Genes are considered significantly differentially expressed if the adjusted *p*-value was below 0.01 (*p* values adjusted using the False Discovery Rate).

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. When several samples per patient were available, only the sample with the highest proportion of tumour tissue was used. Multivariate survival analyses were performed with the clinical covariates Gleason grade (≤7 and >7), pathological stage (*T*1/*T*2 and *T*3/*T*4) and PSA levels (≤10 and >10). We modelled the variables that did not satisfy the proportional hazards assumption (T-stage in MSKCC), as a product of the variable with the heavyside function:

$$g(t) = \begin{cases} 1, \text{ if } t \geq t_0 \\ 0, \text{ otherwise} \end{cases}$$

where t_0 is a time threshold. The multiplication of a predictor with the heavyside function, divides the predictor into time intervals for which the extended Cox model computes different hazard ratios. Before carrying out multivariate analyses we assessed collinearity between the DESNT predictor and the other traditional indicators. To do this we calculated the variance inflation factor (VIF) for each covariate in each model. VIF varied between 1.005241 and 1.461661, suggesting a very weak correlation between the predictors.

2.5 Driving an optimal predictor of DESNT membership

To derive an optimal predictor of DESNT membership the datasets were prepared so that they were comparable: probes were only retained if the associated gene was found in every microarray platform, only one randomly chosen probe was retained per gene and the batch effects adjusted using the ComBat algorithm[24]. The MSKCC dataset was used as the training set and other datasets as test sets. Gene selection was performed using regularized general linear model approach (LASSO) implemented in the glmnet R package[25], starting with all genes that were significantly up or down regulated in DESNT in at least two of the total of five microarray dataset (1669 genes). LASSO was run 100 times and only genes that were selected in at least 25% of runs were retained. The optimal predictor was then derived using the random forest model[26] implemented in the randomForest R package[27]. Default parameters were used, apart from the number of trees were set to 10001 and the class size imbalance was adjusted for by down-sampling the majority class to the frequency of the minority class

3. Results

3.1 Identification of the DESNT cancer category

Four independent transcriptome datasets (designated MSKCC[11], CancerMap, Klein[28], and Stephenson[29], Fig. 1b) obtained from prostatectomy specimens were analyzed. LPD was performed using between 3 and 8 underlying latent processes contributing to the overall expression profile as indicated from log-likelihood plots (Fig. 1b, Supplemental Fig. 1). Following the independent decomposition of each dataset, cancers were assigned to individual processes based on their highest p_i value yielding the results shown in Fig. 1a and Supplemental Fig. 2. p_i is the contribution of each process "*i*" to the expression profile of an individual cancer: sum of p_i over all processes=1.

Searching for relationships between the decompositions, a single process was identified that, based on correlations of gene expression levels, appeared to be common across all four datasets (Fig. 1c). To further investigate this association, for each dataset, we identified genes that were expressed at significantly lower or higher levels ($P < 0.01$ after correction for False Discovery Rate) in the cancers assigned to this process compared to all other cancers from the same dataset. This unveiled a shared set of 45 genes, all with lower expression (Fig. 2a, Supplemental Table 1). Many of the proteins encoded by these 45 core genes are components of the cytoskeleton or regulate its

dynamics, while others are involved in cell adhesion and ion transport (Fig. 2b). Eleven of the 45 genes were members of published prognostic signatures for prostate cancer (Fig. 2c, Supplemental Data File 1). For example *MYLK*, *ACTG2*, and *CNN1* are down-regulated in a signature for cancer metastasis[30], while lower expression of *TPM2* is associated with poorer outcome as part of the Oncotype DX signature[4]. The cancers assigned to this common process are referred to as "DESNT" (*latin* DEScenduNT, they descend).

3.2 Patients with DESNT cancers exhibit poor prognosis

Using linked clinical data available for the MSKCC expression dataset we found that patients with DESNT cancer exhibited poor outcome when compared to patients assigned to other processes ($P = 2.65x10^{-5}$, Log-rank test, Fig. 1d). Validation was provided in two further datasets where PSA failure data following prostatectomy were available (Fig. 1d): for both the Stephenson and CancerMap datasets patients with DESNT cancer exhibited poor outcome ($P = 4.28x10^{-5}$ and $P = 2.98x10^{-8}$ respectively). The number of cancers in each group is indicated in the bottom right corner of each Kaplan-Meier plot. The number of patients with PSA failure is indicated in parentheses. In multivariate analysis, including Gleason sum, Stage and PSA, assignment as a DESNT cancer was an independent predictor of poor outcome in the Stephenson and CancerMap datasets ($P = 1.83x10^{-4}$ and $P = 3.66x10^{-3}$, Cox regression model) but not in the MSKCC dataset ($P = 0.327$) (Table 1, Supplemental Fig. 3). When the three datasets were combined the independent predictive value of DESNT membership was $P = 1.61x10^{-7}$ (Supplemental Fig. 3), compared to $P = 1.00x10^{-5}$ for Gleason sum. Including surgical margin status in the multivariate analysis had little influence on these values giving $P = 3.63x10^{-7}$ for DESNT compared to $P = 1.80x10^{-5}$ for Gleason Sum. The combined multivariate model is a significant improvement over a baseline Cox proportional hazard ratio model containing Gleason, PSA and Clinical Stage ($p=9.528x10^{-7}$; likelihood ratio test). The poor prognosis DESNT process was also identified in the CamCap dataset[9] (Table 1, Supplemental Fig. 3 and 4), which was excluded from the above analysis because it was not independent: there is a substantial overlap with cancers included in CancerMap (Fig. 1b).

3.3 A random forest classifier for identifying DESNT cancer

We wished to develop a classifier that, unlike LPD, was not computer processing intensive and that could be applied both to a wider range of datasets and to individual cancers. 1669 genes with significantly altered expression between DESNT and non-DESNT cancers in at least two datasets were selected for analysis. A LASSO logistic regression model was used to identify genes that were the best predictors of DESNT membership in the MSKCC dataset leading to the selection of a set of 20 genes (Supplemental Table 2), which had a one gene overlap (*ACTG2*) to the 45 genes with significantly lower expression in DESNT cancers. Using random forest (RF) classification these 20 genes provided high specificity and sensitivity for predicting that individual cancers were DESNT in both the MSKCC training dataset and in three validation datasets (Supplemental Fig. 5). For the two validation datasets (Stephenson and CancerMap)

with linked PSA failure data the predicted cancer subgroup exhibited poorer clinical outcome in both univariate and multivariate analyses, in agreement with the results observed using LPD (Table 1, Fig. 3).

3.4 DESNT cancers in the The Cancer Genome Atlas dataset

When RF classification was applied to RNAseq data from 333 prostate cancers described by The Cancer Genome Atlas (TCGA)[13] a patient subgroup was identified that was confirmed as DESNT based on: (i) correlations of gene expression levels with DESNT cancer groups in other datasets (Supplemental Fig 6); (ii) demonstration of overlaps of differentially expressed genes between DESNT and non-DESNT cancers with the core down-regulated gene set (45/45 genes); and (iii) its poorer clinical outcome based on PSA failure ($P = 5.4 \times 10^{-4}$) compared to non-DESNT patients (Table 1, Fig. 3e).

For the TCGA dataset, we failed to find correlations between assignment as a DESNT cancer and the presence of any specific genetic alteration ($P > 0.05$ after correction for False Discovery Rate, χ^2 test, Fig. 4). Of particular note there was no correlation to ETS-gene status (P, = 0.136, χ^2 test, Fig. 4). A lack of correlation between DESNT cancers and *ERG*-gene rearrangement, determined using the fluorescence *in situ* hybridization break-apart assay[31], was confirmed using CancerMap samples (LPD-DESNT, $P = 0.549$; RF-DESNT, $P = 0.2623$, χ^2 test: DESNT cancers identified by LPD and by RF approaches are referred to respectively as LPD-DESNT and RF-DESNT). These observations are consistent with the lack of correlation between *ERG* status and clinical outcome[32], although different views on the relationship between *ERG*-gene status and clinical outcome have been expressed[33]. Since ETS-gene alteration, found in around half of prostate cancers[13,31], is considered to be an early step in prostate cancer development[15,34] it is likely that changes involved in the generation of DESNT cancer represent a later event that is common to both ETS-positive and ETS-negative cancers.
For RF-DESNT cancers in the TGCA series many of the 45 core genes exhibited altered levels of CpG gene methylation compared to non-RF-DESNT cancers (Supplemental Table 3) suggesting a possible role in controlling gene expression. Supporting this idea, for sixteen of the 45 core genes epigenetic down-regulation in human cancer has been previously reported, including six genes in prostate cancer (*CLU*, *DPYSL3*, *GSTP1*, *KCNMA1*, *SNAI2*, and *SVIL*) (Fig 2b, Supplemental Table 1). CpG methylation of five of the genes (*FBLN1*, *GPX3*, *GSTP1*, *KCNMA1, TIMP3)* has previously been linked to cancer aggression.

4. Discussion

Evidence from The European Randomized study of Screening for Prostate Cancer demonstrates that PSA screening can reduce mortality from prostate cancer by 21%[35]. However, a critical problem is that the progression of prostate cancer is highly heterogeneous[36,37] and PSA screening leads to the detection of up to 50% of cancers that are clinically irrelevant[38,39]: that is cancers that would never have caused symptoms in a man's lifetime in the absence of screening. Unsupervised analyses of breast cancer datasets using hierarchical clustering previously revealed the existence

basal, *ERBB2*-overexpressing and luminal cancer categories[10]. This mathematical approach has not proven successful when applied to prostate cancer microarray datasets[9,11]. However in our study the use of LPD, an unsupervised method that takes into account the issue of cancer heterogeneity, has revealed the existence of a novel category of prostate cancer, designated DESNT, common across all datasets. The subsequent linking to clinical data revealed that DESNT cancers exhibit poor prognosis.
It was notable that membership of the DESNT cancer groups was not an independent predictor of clinical outcome in the MSKCC dataset. It is possible that the difference may simply reflect statistical variation since the size of the DESNT group in several datasets was small (MSKCC, 13%; CancerMap, 8%; Stephenson, 31%; Klein, 23%). Critically, however, when the datasets with linked clinical data were combined DESNT membership remained an independent predictor of clinical outcome. We failed to detect systematic differences between MSKCC and other datasets used in multivariate analyses (Supplemental Fig. 3h).

We have not, in this study, investigated the biological function and mechanisms of alterations of expression of the 45 core genes. However gene down-regulation mediated by CpG methylation is well documented in human cancer, as is the association of CpG methylation of single genes with aggressive cancer behavior (Supplemental Table 1). The results found for DESNT cancers are consistent with these observations, but would suggest that it is the combine under expression of multiple genes that represents a critical determinant of cancer progression and aggression. Several of the genes found to have lower expression in DESNT cancer (*ACTA2*, *CNN1*, *LMOD1*) encode proteins primarily expressed in smooth muscle cells or myofibroblast, indicative of an altered tumour-stromal environment. We failed to find a correlation between stromal content and clinical outcome in the CamCap and CancerMap datasets (Fig. 2). However this does not exclude the possibility that DESNT cancers themselves may have lower stromal content, in part explaining the lower expression of these genes.

Other under-expressed genes encode components of the actin cytoskeleton or regulate its dynamics (e.g. *MLCK*, *MYL9*, *ACTN1*, and *TNS1*). Increased malignancy may correlate with increased cell migratory behaviour, which in turn can involve deployment of particular types of cell adhesion and cytoskeletal machinery[40]. A high dependency on actomyosin contractility is recognised as a hallmark of amoeboid movement. Down-regulation of these genes in DESNT cancers would argue against its involvement. The lower expression of focal adhesion components such as integrin α5 (ITGA5), vinculin (VCL) and integrin-linked kinase (ILK), would also argue against involvement of "mesenchymal" type migration, which is dependent on these classes of genes[40]. It is thus possible that the observed alterations may support involvement of collective migration or expansive growth phenotypes[40].

Notably, we failed to find any relationship between DESNT cancers and either CNV (copy number variant) signatures (Lalonde *et al.* and Ross-Adams *et al.* in Fig. 2c) or DNA repair gene alterations (Fig. 4). Assignment of cancers within the DESNT classification

framework together with the use of standard clinical indicators (Stage, Gleason sum, PSA), CNV signatures[11], expression biomarkers such as Prolaris[3], Decipher[5], and Oncotype DX[4] identified in supervised analyses and urine biomarkers[41], should significantly enhance the ability identify patients whose cancers should be targeted by radical therapies, avoiding the side effects of treatment, including impotence, in men with non-aggressive disease. In future studies we are focusing on the development of both LPD and RF based tests that can be used to detect DESNT cancer in biopsy tissue in a clinical setting.

References

[1] D'Amico AV. Biochemical Outcome After Radical Prostatectomy, External Beam Radiation Therapy, or Interstitial Radiation Therapy for Clinically Localized Prostate Cancer. Jama 1998;280:969–74. doi:10.1001/jama.280.11.969.

[2] Graham J, Kirkbride P, Cann K, Hasler E, Prettyjohns M. Prostate cancer: summary of updated NICE guidance. Bmj 2014;348:f7524–4. doi:10.1136/bmj.f7524.

[3] Cuzick J, Swanson GP, Fisher G, Brothman AR, Berney DM, Reid JE, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol 2011;12:245–55. doi:10.1016/S1470-2045(10)70295-3.

[4] Klein EA, Cooperberg MR, Magi-Galluzzi C, Simko JP, Falzarano SM, Maddala T, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. Eur Urol 2014;66:550–60. doi:10.1016/j.eururo.2014.05.004.

[5] 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:e66855. doi:10.1371/journal.pone.0066855.

[6] Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 2004;113:913–23. doi:10.1172/JCI20032.

[7] Tomlins SA, Alshalalfa M, Davicioni E, Erho N, Yousefi K, Zhao S, et al. Characterization of 1577 primary prostate cancers reveals novel biological and clinicopathologic insights into molecular subtypes. Eur Urol 2015;68:555–67. doi:10.1016/j.eururo.2015.04.033.

[8] You S, Knudsen BS, Erho N, Alshalalfa M, Takhar M, Al-Deen Ashab H, et al. Integrated Classification of Prostate Cancer Reveals a Novel Luminal Subtype with Poor Outcome. Cancer Res 2016;76:4948–58. doi:10.1158/0008-5472.CAN-16-0902.

[9] Ross-Adams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, Massie CM, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. EBioMedicine 2015;2:1133–44. doi:10.1016/j.ebiom.2015.07.017.

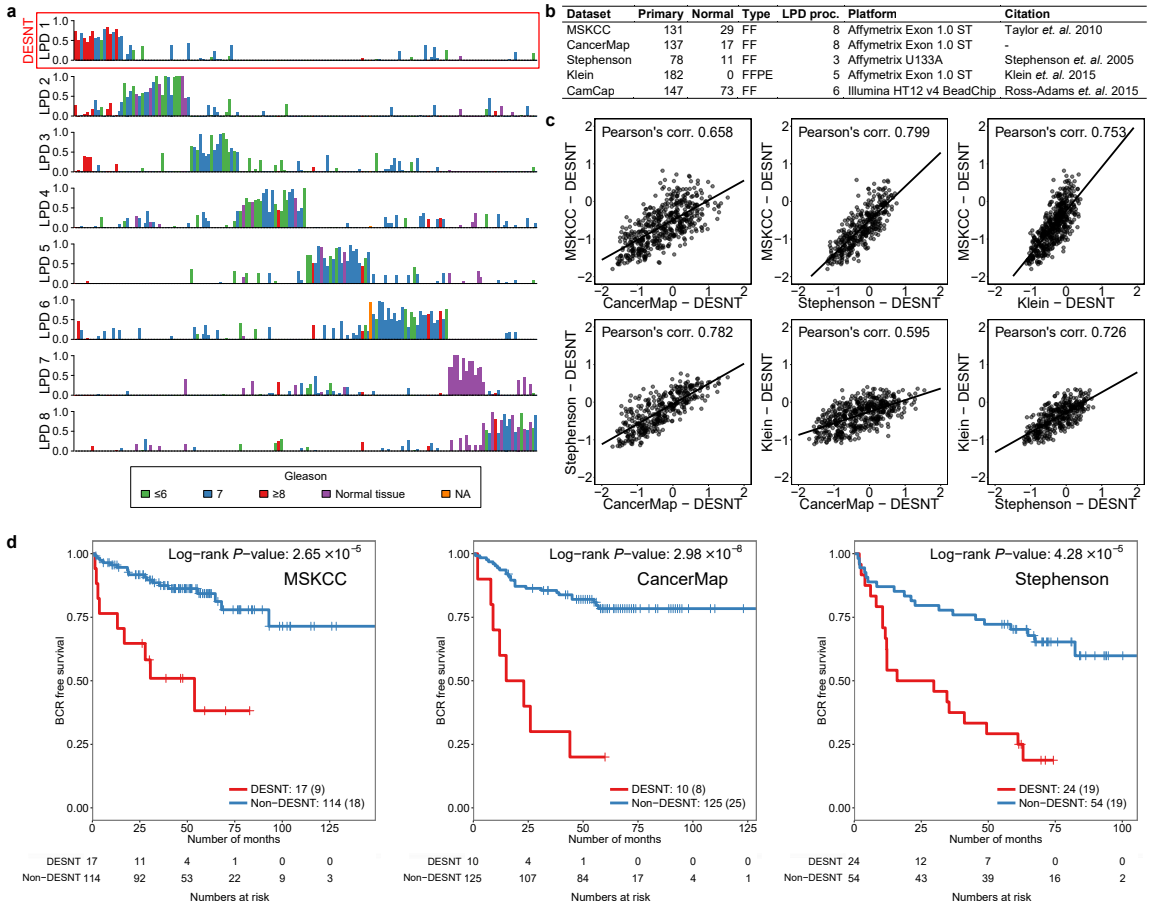
[10] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USa 2003;100:8418–23. doi:10.1073/pnas.0932692100.
[11] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010;18:11–22. doi:10.1016/j.ccr.2010.05.026.
[12] 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;47:367–72. doi:10.1038/ng.3221.
[13] Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell 2015;163:1011–25. doi:10.1016/j.cell.2015.10.025.
[14] 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;47:736–45. doi:10.1038/ng.3315.
[15] Clark J, Attard G, Jhavar S, Flohr P, Reid A, De-Bono J, et al. Complex patterns of ETS gene alteration arise during cancer development in the human prostate. Oncogene 2008;27:1993–2003. doi:10.1038/sj.onc.1210843.
[16] 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;16:641. doi:10.1186/s12885-016-2674-6.
[17] 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;3:367–81. doi:10.1098/rsif.2005.0093.
[18] Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray data sets. IEEE/ACM Trans Comput Biol Bioinform 2005;2:143–56. doi:10.1109/TCBB.2005.29.
[19] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003;3:993–1022.
[20] 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;13:1114–24. doi:10.1016/S1470-2045(12)70372-8.
[21] Warren AY, Whitaker HC, Haynes B, Sangan T, McDuffus L-A, Kay JD, et al. Method for sampling tissue for research which preserves pathological data in radical prostatectomy. Prostate 2013;73:194–202. doi:10.1002/pros.22556.
[22] Jhavar S, Reid A, Clark J, Kote-Jarai Z, Christmas T, Thompson A, et al. Detection of TMPRSS2-ERG translocations in human prostate cancer by expression profiling using GeneChip Human Exon 1.0 ST arrays. J Mol Diagn 2008;10:50–7. doi:10.2353/jmoldx.2008.070085.
[23] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47–7. doi:10.1093/nar/gkv007.

[24] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118–27. doi:10.1093/biostatistics/kxj037.
[25] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010;33:1–22. doi:10.1109/TPAMI.2005.127.
[26] Breiman L. Random Forests. Machine Learning 2001;45:5–32. doi:10.1023/A:1010933404324.
[27] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002.
[28] 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;67:778–86. doi:10.1016/j.eururo.2014.10.036.
[29] 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;104:290–8. doi:10.1002/cncr.21157.
[30] Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33:49–54. doi:10.1038/ng1060.
[31] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8. doi:10.1126/science.1117679.
[32] Weischenfeldt J, Simon R, Feuerbach L, Schlangen K, Weichenhan D, Minner S, et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell 2013;23:159–70. doi:10.1016/j.ccr.2013.01.002.
[33] Clark JP, Cooper CS. ETS gene fusions in prostate cancer. Nat Rev Urol 2009;6:429–39. doi:10.1038/nrurol.2009.127.
[34] Park K, Dalton JT, Narayanan R, Barbieri CE, Hancock ML, Bostwick DG, et al. TMPRSS2:ERG gene fusion predicts subsequent detection of prostate cancer in patients with high-grade prostatic intraepithelial neoplasia. J Clin Oncol 2014;32:206–11. doi:10.1200/JCO.2013.49.8386.
[35] 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;384:2027–35. doi:10.1016/S0140-6736(14)60525-0.
[36] D'Amico AV. Cancer-Specific Mortality After Surgery or Radiation for Patients With Clinically Localized Prostate Cancer Managed During the Prostate-Specific Antigen Era. Journal of Clinical Oncology 2003;21:2163–72. doi:10.1200/JCO.2003.01.075.
[37] 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 2012;30:1857–63. doi:10.1200/JCO.2011.35.1924.
[38] Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever E, Gulati R, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J Natl Cancer Inst 2009;101:374–83. doi:10.1093/jnci/djp001.
[39] Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of Study Features and Methods on Overdiagnosis Estimates in Breast and Prostate Cancer Screening. Annals of Internal Medicine 2013;158:831–8. doi:10.7326/0003-4819-158-11-201306040-00008.
[40] Friedl P, Locker J, Sahai E, Segall JE. Classifying collective cancer cell invasion. Nat Cell Biol 2012;14:777–83. doi:10.1038/ncb2548.
[41] Van Neste L, Hendriks RJ, Dijkstra S, Trooskens G, Cornel EB, Jannink SA, et al. Detection of High-grade Prostate Cancer Using a Urinary Molecular Biomarker-Based Risk Score. Eur Urol 2016;70:740–8. doi:10.1016/j.eururo.2016.04.012.

Acknowledgement of Support: This work was funded by the Bob Champion Cancer Trust, The Masonic Charitable Foundation successor to The Grand Charity, The King Family, and The University of East Anglia. We acknowledge support from Movember, from Prostate Cancer UK, Callum Barton, The Big C Cancer Charity, and from The Andy Ripley Memorial Fund. 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. Cancer Research UK Grant 10047 funded the generation of the prostate CancerMap expression microarray dataset. We would like to acknowledge the support of the National Institute for Health Research (NIHR) which funds the Cambridge Bio-medical Research Centre, Cambridge UK. The sponsors did not participate in the design and conduct of the study; data collection, management, analysis, and interpretation; and manuscript preparation, review, and approval

FIGURES AND TABLES



Dataset	Primary	Normal	Type	LPD proc.	Platform	Citation
MSKCC	131	29	FF	8	Affymetrix Exon 1.0 ST	Taylor *et. al.* 2010
CancerMap	137	17	FF	8	Affymetrix Exon 1.0 ST	-
Stephenson	78	11	FF	3	Affymetrix U133A	Stephenson *et. al.* 2005
Klein	182	0	FFPE	5	Affymetrix Exon 1.0 ST	Klein *et. al.* 2015
CamCap	147	73	FF	6	Illumina HT12 v4 BeadChip	Ross-Adams *et. al.* 2015

Figure 1. Latent Process Decomposition (LPD), gene correlations and clinical outcome. a, LPD analysis of Affymetrix expression data from the MSKCC datasets divided the samples into eight processes, each represented here by a bar chart. Samples are represented in all eight processes and height of each bar corresponds to the proportion (p_i) of the signature that can be assigned to each LPD process. Samples are assigned to the LPD group in which they exhibit the highest value of p_i. LPD was performed using the 500 gene probes with the greatest variation in expression between samples in the MSKCC dataset. The process containing DESNT cancers is indicated. **b**, List of datasets used in LPD analysis. The unique number of primary cancer and normal specimens used in LPD are indicated. FF, fresh frozen specimen; FFPE, formalin-fixed paraffin embedded specimen. The CancerMap and CamCap were not independent having 40 cancers in common. Clinical and molecular details for the CancerMap dataset are given in Supplemental Table 4 and Supplemental Data File 2. Clinical details for samples from other datasets used in this study can be found in Supplemental Data File 3. **c**, Correlations of average levels of gene expression between cancers designated as DESNT. All six comparisons for the MSKCC, CancerMap, Stephenson and Klein datasets are shown. The expression levels of each gene have been normalised across all samples to mean 0 and standard deviation 1. **d**, Kaplan-Meier PSA failure plots for the MSKCC, CancerMap and Stephenson datasets.

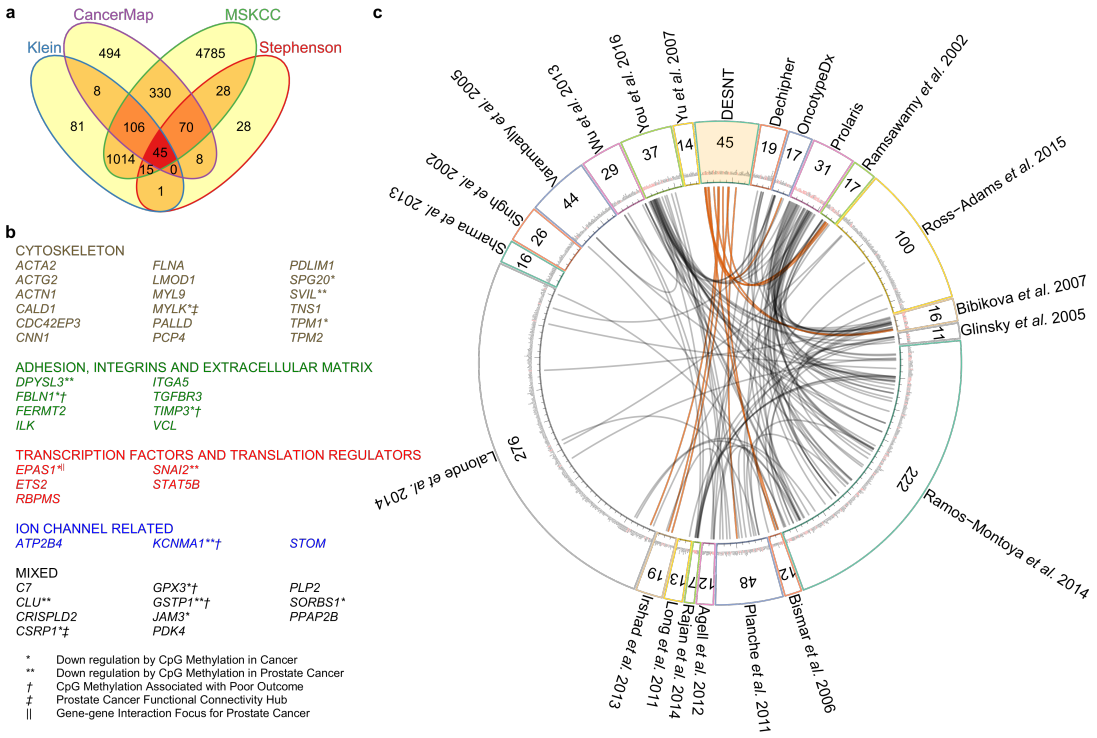


Figure 2. Genes commonly down-regulated in DESNT poor prognosis prostate cancer. a, Number of genes with significantly altered expression in DESNT cancers compared to non-DESNT cancers (P<0.01 after correction for False Discovery Rate). 45 genes had lower expression in DESNT cancers in all four expression microarray datasets, based on a stringency requirement of being down-regulated in at least 80 of 100 independent LPD runs. **b**, List of the 45 genes according to biological grouping. Previous published evidence is represented as superscripts and the supporting references are provided in Supplemental Table 1. Encoded protein functions are shown in Supplemental Table 5. Although some of the 45 genes are preferentially expressed in stromal tissue we found no correlation between stromal content and clinical outcome in both the CancerMap and CamCap patient series, where data on cellular composition were available. When patients were stratified into two groups (above and below median stromal content) Kaplan-Meier plots failed to show outcome difference for both the CancerMap (Log-rank test, p=0.159) and CamCap (p=0.261) patient series. **c.** Relationship between the genes in published poor prognosis signatures for prostate cancer and the DESNT classification for human prostate cancer, represented as a circos plot. Links to the 45 commonly down-regulated genes are shown in brown. References quoted in the circos plot are listed in the Supplemental Information and detailed gene relationships are shown in Supplemental Data File 1.

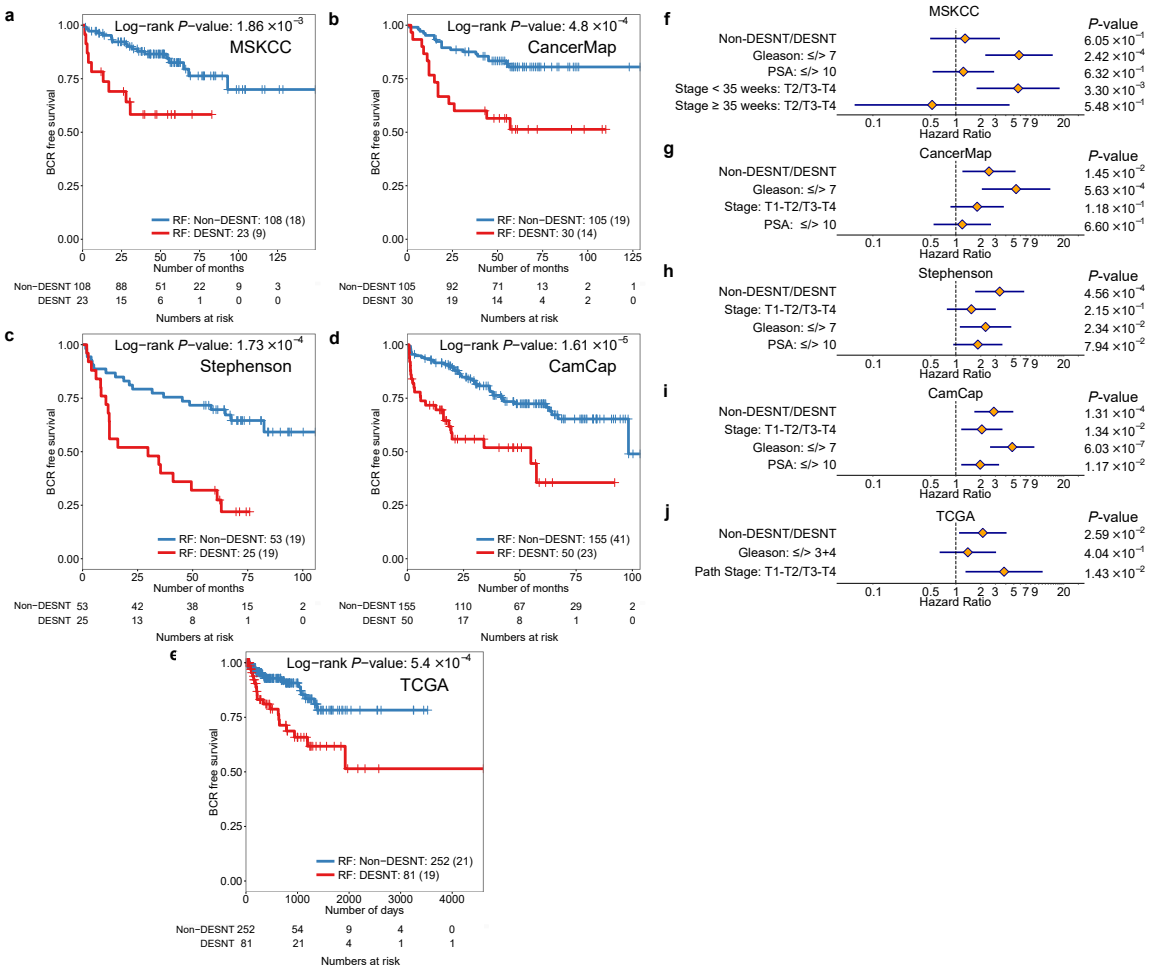


Figure 3. Analysis of outcome for DESNT cancers identified by RF classification. (a-e) Kaplan-Meier PSA failure plots for the MSKCC (**a**), CancerMap (**b**), Stephenson (**c**), CamCap (**d**) and TCGA (**e**) datasets. For each dataset the cancers assigned to DESNT using the 20 gene RF classifier are compared to the remaining cancers. The number of cancers in each group is indicated in the bottom right corner of each plot. The number of cancers with PSA failure is indicated in parentheses. Multivariate analyses were performed as described in the Methods for the MSKCC (**f**), CancerMap (**g**), Stephenson (**h**), CamCap (**i**) and TCGA (**j**) datasets. Pathological Stage covariates for MSKCC and Stephenson datasets did not meet the proportional hazards assumptions of the Cox model and have been modelled as time-dependent variables, as described in the Methods.

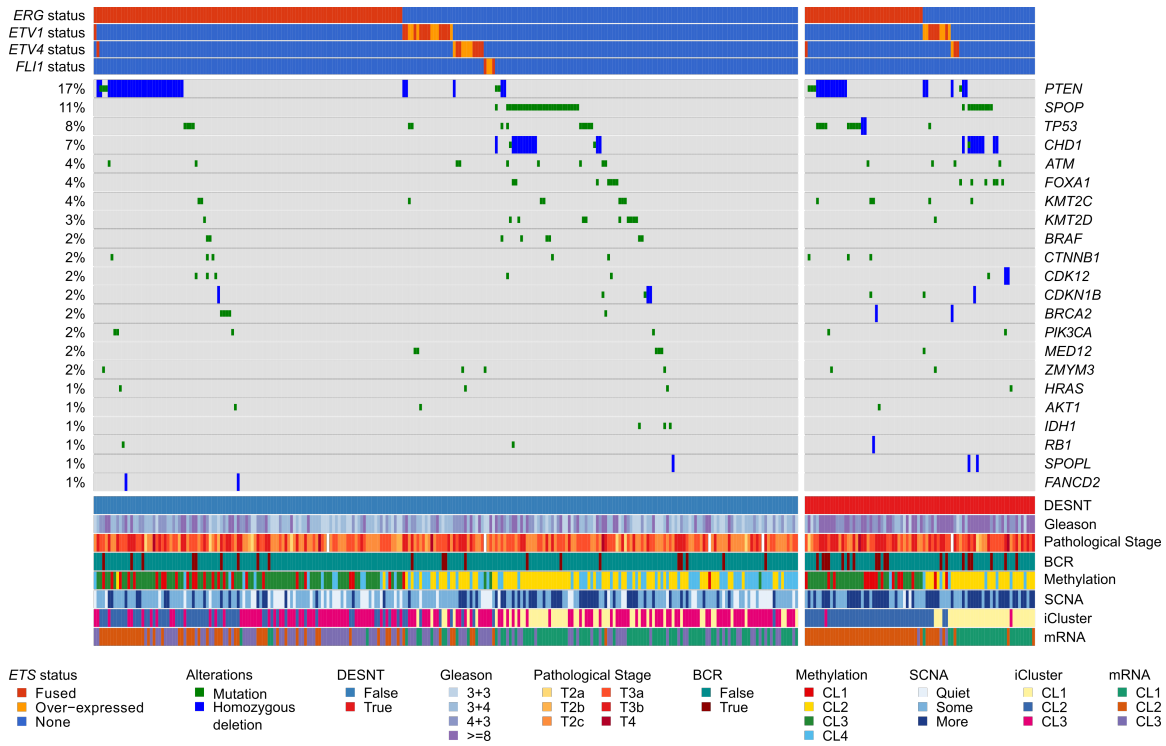


Figure 4. Comparison of RF-DESNT and non-RF-DESNT cancers in The Cancer Genome Atlas dataset. A 20 gene random forest (RF) classifier was used to identify DESNT cancers (designated RF-DESNT cancers). The types of genetic alteration are shown for each gene (mutations, fusions, deletions, and overexpression). Clinical parameters including biochemical recurrence (BCR) are represented at the bottom together with groups for iCluster, methylation, somatic copy number alteration (SVNA) and mRNA[13]. When mutations and homozygous deletions for each gene were combined RF-DESNT cancers contained an excess of genetic alterations in *BRCA2* ($P = 0.021$, χ^2 test) and *TP53* ($P = 0.0038$), but after correcting for multiple testing these differences were not significant ($P > 0.05$).

Table 1: Poor clinical outcome of patients with DESNT cancer

Latent Process Decomposition		
Dataset	**Univariate *p*-value**	**Multivariate *p*-value**
MSKCC	2.65×10^{-5}	3.27×10^{-1}
CancerMap	2.98×10^{-8}	3.66×10^{-3}
Stephenson	4.28×10^{-5}	1.83×10^{-4}
CamCap	1.22×10^{-3}	2.90×10^{-2}
Random Forest		
Dataset	**Univariate *p*-value**	**Multivariate *p*-value**
MSKCC	1.85×10^{-3}	6.05×10^{-1}
CancerMap	4.80×10^{-4}	1.45×10^{-2}
Stephenson	1.75×10^{-4}	4.56×10^{-4}
CamCap	1.61×10^{-5}	1.31×10^{-4}
TCGA	5.41×10^{-4}	2.59×10^{-2}

For each dataset comparisons were made between PSA failures reported for DESNT and non-DESNT cancers. LPD, Latent Process Decomposition; RF, Random Forest. For LPD the log-rank *P*-values represent the modal LPD run selected from the 100 independent LPD runs as described in the Methods. For multivariate analyses Gleason sum, PSA at diagnosis and Pathological Stage are included for all datasets with the exception of the TCGA dataset where only Gleason sum and Clinical Stage data were available. The full analyses are presented in Fig. 3 and Supplemental Fig. 3.