

This Provisional PDF corresponds to the article as it appeared upon acceptance. Fully formatted PDF and full text (HTML) versions will be made available soon.

Molecular sampling of prostate cancer: a dilemma for predicting disease progression

BMC Medical Genomics 2010, **3**:8 doi:10.1186/1755-8794-3-8

Andrea Sboner (andrea.sboner@yale.edu)
Francesca Demichelis (frd2004@med.cornell.edu)
Stefano Calza (calza@med.unibs.it)
Yudi Pawitan (yudi.pawitan@ki.se)
Sunita R Setlur (ssetlur@rics.bwh.harvard.edu)
Yujin Hoshida (hoshida@broad.mit.edu)
Sven Perner (sven.perner@medizin.uni-tuebingen.de)
Hans-Olov Adami (adami@hsph.harvard.edu)
Katja Fall (Katja.Fall@ki.se)
Lorelei A Mucci (lmucci@hsph.harvard.edu)
Philip W Kantoff (Philip_Kantoff@dfci.harvard.edu)
Meir Stampfer (stampfer@hsph.harvard.edu)
Swen-Olof Andersson (swenolof.andersson@orebroll.se)
Eberhard Varenhorst (varenhorst.eberhard@telia.com)
Jan-Erik Johansson (jan-erik.johansson@orebroll.se)
Mark B Gerstein (mark.gerstein@yale.edu)
Todd R Golub (golub@broad.harvard.edu)
Mark A Rubin (asbmgr@gersteinlab.org)
Ove Andren (ove.andren@orebroll.se)

ISSN 1755-8794

Article type Research article

Submission date 5 November 2009

Acceptance date 16 March 2010

Publication date 16 March 2010

Article URL <http://www.biomedcentral.com/1755-8794/3/8>

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

© 2010 Sboner *et al.*, licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

<http://www.biomedcentral.com/info/authors/>

Molecular sampling of prostate cancer: a dilemma for predicting disease progression

Andrea Sboner¹, Francesca Demichelis^{2,3}, Stefano Calza^{4,5}, Yudi Pawitan⁴, Sunita R Setlur⁶, Yujin Hoshida^{7,8}, Sven Perner², Hans-Olov Adami^{4,9}, Katja Fall^{4,9}, Lorelei A Mucci^{9,11,12}, Philip W Kantoff^{8,11}, Meir Stampfer^{9,11,12}, Swen-Olof Andersson¹⁰, Eberhard Varenhorst¹³, Jan-Erik Johansson¹⁰, Mark B Gerstein^{1,14,15}, Todd R Golub^{7,8,16,#}, Mark A Rubin^{2,7,*,#}, Ove Andréén^{10,#}

¹ Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, 06520, USA.

² Department of Pathology and Laboratory Medicine, Weill Cornell Medical Center, New York, New York, USA.

³ Institute for Computational Biomedicine, Weill Cornell Medical Center, New York, New York, USA.

⁴ Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

⁵ Department of Biomedical Sciences and Biotechnologies, University of Brescia, Brescia, Italy.

⁶ Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, 02115, USA.

⁷ The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, 02142, USA.

⁸ The Dana Farber Cancer Institute, Boston, Massachusetts, 02115, USA.

⁹ Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, 02115, USA.

¹⁰ Department of Urology, Örebro University Hospital, Örebro, SE-701 85, Sweden.

¹¹ Harvard Medical School, Boston, Massachusetts 02115, USA.

¹² Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.

¹³ Department of Urology, Linköping University Hospital, Linköping, SE 581 85, Sweden.

¹⁴ Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA

¹⁵ Department of Computer Science, Yale University, New Haven, Connecticut, 06520, USA

¹⁶ The Howard Hughes Medical Institute at The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, 02142, USA.

These authors share senior authorship.

* Correspondence to:

Mark A. Rubin, M.D.

Weill Cornell Medical College

1300 York Avenue Room C 410-A (or box #69)

New York, New York 10021

Phone: (212) 746-6313

Fax: (212) 746-8816

Email: rubinma@med.cornell.edu

Email addresses:

ASB: andrea.sboner@yale.edu

FD: frd2004@med.cornell.edu

SC: calza@med.unibs.it

YP: yudi.pawitan@ki.se

SRS: ssetlur@rics.bwh.harvard.edu

YH: hoshida@broad.mit.edu

SP: sven.perner@medizin.uni-tuebingen.de

HOA: adami@hsph.harvard.edu

KF: Katja.Fall@ki.se

LAM: lmucci@hsph.harvard.edu

PWK: Philip_Kantoff@dfci.harvard.edu

MS: stampfer@hsph.harvard.edu

SOA: swenolof.andersson@orebroll.se

EV: varenhorst.eberhard@telia.com

JEJ: jan-erik.johansson@orebroll.se

MBG: mark.gerstein@yale.edu

TRG: golub@broad.harvard.edu

MAR: rubinma@med.cornell.edu

OA: ove.andren@orebroll.se

Abstract

Background

Current prostate cancer prognostic models are based on pre-treatment prostate specific antigen (PSA) levels, biopsy Gleason score, and clinical staging but in practice are inadequate to accurately predict disease progression. Hence, we sought to develop a molecular panel for prostate cancer progression by reasoning that molecular profiles might further improve current clinical models.

Methods

We analyzed a Swedish Watchful Waiting cohort with up to 30 years of clinical follow up using a novel method for gene expression profiling. This cDNA-mediated annealing, selection, ligation, and extension (DASL) method enabled the use of formalin-fixed paraffin-embedded transurethral resection of prostate (TURP) samples taken at the time of the initial diagnosis. We determined the expression profiles of 6100 genes for 281 men divided in two extreme groups: men who died of prostate cancer and men who survived more than 10 years without metastases (lethals and indolents, respectively). Several statistical and machine learning models using clinical and molecular features were evaluated for their ability to distinguish lethal from indolent cases.

Results

Surprisingly, none of the predictive models using molecular profiles significantly improved over models using clinical variables only. Additional computational analysis confirmed that molecular heterogeneity within both the lethal and indolent classes is widespread in prostate cancer as compared to other types of tumors.

Conclusions

The determination of the molecularly dominant tumor nodule may be limited by sampling at time of initial diagnosis, may not be present at time of initial diagnosis, or may occur as the disease progresses

making the development of molecular biomarkers for prostate cancer progression challenging.

Background

The paramount clinical dilemma in prostate cancer management is how to treat the man with clinically localized disease because the natural history is favorable overall [1] and the benefit from radical treatment modest [2]. Numerous studies have attempted to address this issue but the lack of data with long-term clinical outcomes precludes a definitive assessment. This problem is real and mounting. In 2008, it was estimated that 186,320 new cases of prostate cancer were diagnosed in the United States with the vast majority being clinically localized [3]. The majority of these men are predicted to survive despite prostate cancer for 5 or 10 years regardless of the type of treatment they initially receive [4]. This would suggest that expectant management for localized prostate cancer might be an important modality to deal with this common malignancy. This approach would potentially gain more widespread acceptance if we could sort out those men that were at the greatest risk of disease progression at time of initial diagnosis.

Various approaches using clinical parameters including prostate specific antigen (PSA) levels at time of initial diagnosis have been explored to predict disease progression [5-7]. Although these models work well for men with extreme levels of PSA, the majority of men fall within an intermediate range characterized by a PSA level between 4-10 ng/ml and a Gleason score of 6 or 7. A Gleason score is assigned to a prostate cancer based on its microscopic architectural appearance. It ranges from 2 to 10, with higher values associated with higher tumor grade. The need for additional tests to complement and improve upon these existing approaches would help identify men who must be treated and who can safely be monitored for disease progression.

We reasoned that by performing high-throughput expression profiling of transurethral resection of the

prostate (TURP) samples from a large cohort of men on a Watchful Waiting cohort, we would identify a molecular profile predictive of prostate cancer disease progression. We further reasoned that employing a combination of novel technology and a well-defined clinical cohort should yield a strong lethal prostate cancer signature.

Limitations of prior prostate cancer expression profiling studies have included small sample size, restriction of populations to surgical cohorts, short follow up time, and the use of surrogate endpoints such as PSA biochemical recurrence to define disease progression. To overcome these limitations, we designed a study using prostate cancer samples prospectively registered as part of a Watchful Waiting protocol from two regions in Sweden. Up to 30 years of clinical follow up information was available on these men. All of the cases were detected incidentally in a pre-Prostate Specific Antigen (PSA) screening era.

Methods

Patient population

The present study is nested in a cohort of men with localized prostate cancer diagnosed in the Örebro (1977 to 1994) and South East (1987 to 1999) Health Care Regions of Sweden. Eligible patients were identified through population-based prostate cancer quality databases maintained in these regions (described in Johansson et al., Aus et al., and Andren et al. [1, 8, 9]) and included men who were diagnosed with incidental prostate cancer through (TURP) or adenoma enucleation, i.e. stage T1a-b tumors. In accordance with standard treatment protocols at the time, patients with early stage/localized prostate cancer were followed expectantly (“watchful waiting”). No PSA screening programs were in place at the time.

The study cohort was followed for cancer-specific and all cause mortality until March 1, 2006 through record linkages to the essentially complete Swedish Death Register, which provided date of death or migration. Information on causes of death was obtained through a complete review of medical records by a study end-point committee. Deaths were classified as cancer-specific when prostate cancer was the primary cause of death.

We were able to trace tumor tissue specimens from 92% (1256/1367) of all potentially eligible cases. In order to provide complete and consistent information, available hematoxylin and eosin (H&E) slides from each case were reviewed to identify all tissue specimens with tumor tissue. Slides and corresponding paraffin-embedded formalin-fixed blocks were subsequently retrieved and re-reviewed to confirm cancer status and to assess Gleason score and other notable histopathologic features. The reviewers were blinded with regard to disease outcome. Gleason score was evaluated according to Epstein et al.[10]. All patients gave informed consent for the study.

Study design

Since our overarching aim was to identify signatures predicting a lethal or an indolent course of prostate cancer, we maximized efficiency by devising a study design that included men who either died from prostate cancer during follow up (lethal prostate cancer cases) or who survived at least 10 years after their diagnosis (men with indolent prostate cancer). We thus excluded men with non-informative outcomes, namely those who died from other causes within ten years of their prostate cancer diagnosis or had been followed for less than 10 years with no disease progression (n=595). All men with samples in which high-density tumor regions (defined as more than 90% tumor cells) could be identified were included (n=381). We excluded from the indolent group men who had received any type of androgen deprivation treatment during follow up (n=79), since some of these had potentially lethal disease that was deferred by therapy. Twenty-one men were further excluded due to poor sample quality. In total, 281 men (116 with indolent disease and 165 with lethal prostate cancer) were included in the analyses (see **Figure 1**). The study design was approved by the Ethical Review Boards in Örebro and Linköping. The clinical and pathologic demographics of these of 281 men with prostate cancer are presented in **Additional File 1, Table S1**.

In addition to the standard pathology evaluation we also characterized each case with respect to ERG gene rearrangement, since it appears that this event is an indicator of poor prognosis (**Additional File 1**).

Complementary DNA–Mediated Annealing, Selection, Ligation, and Extension Array Design

An array of 6100 genes (6K DASL) was designed for the discovery of molecular signatures relevant to prostate cancer by using four complementary DNA (cDNA)–mediated annealing, selection, ligation, and extension (DASL) assay panels (DAPs) [11, 12]. Details of this procedure can be found in **Additional File 1** and also at Gene Expression Omnibus (GEO: <http://www.ncbi.nlm.nih.gov/geo/>) with platform accession number: GPL5474. This data set is also available at GEO with accession number: GSE16560.

Supervised classification models: implementation and evaluation.

In order to identify and evaluate a predictive molecular signature, six supervised classification models were implemented: k-Nearest Neighbor (kNN) [13], Nearest Template Prediction (NTP) [14], Diagonal Linear Discriminant Analysis (DLDA)[15], Support Vector Machine (SVM)[16], Neural Network (NN)[13], and Logistic Regression (LR)[17]. Their performances were evaluated and compared through a split-sample validation procedure. Specifically, the entire data set was randomly split into a Learning and a Validation sets, with approximately equal proportion of men with lethal and indolent prostate cancer (**Figure 1**). The Learning set is utilized to create the models and select the best classifier, whose performance is evaluated on the Validation set by means of the Area under the Receiving Operating Curve (AUC). This procedure enables the unbiased estimation of the performance of a classifier since the evaluation is performed on an independent data set [18]. To optimize the classifiers and select the best model, we adopted an iterative cross-validation procedure within the Learning set. The rationale is that results of this procedure enable the identification of the best model

which is then used to build a classifier (using the whole Learning set) that is finally evaluated on the Validation set. Specifically, a stratified 10-fold cross-validation split the Learning set in 10 disjoint partitions, $test_i$ ($i = 1..10$), with approximately equal proportion of lethal and indolent cases each. Given a partition $test_i$, classifiers were created using the cases not in that partition, i.e. $training_i$, and evaluated on $test_i$. This procedure was repeated 10 times and the final results are averaged across the 10 iterations. Moreover, to avoid potential biases in the selection of the 10 partitions, the entire procedure was repeated 100 times resulting in 1000 different partitions. The best model was then identified by comparing the results obtained on the 100 iterations.

Feature Selection: At each iteration of the cross-validation, a feature selection procedure was carried out to identify the subset of genes that are differentially expressed between lethals and indolents. A two-sided t-test was performed for each gene within the $training_i$ partition. Different thresholds on the p-values were used for selection (0.01, 0.001). We ensured that the selection of genes is performed using only the samples used for training, avoiding over-fitting the data. For DLDA and the logistic regression models, a stepwise-like feature selection was implemented. Specifically, genes were sorted according to their t-test p-value and then added to the model one at the time. The best gene set is then selected as the one achieving the best AUC with the fewer number of gene predictors.

Model selection: Each classifier has its own set of parameters that need to be optimized. The identification of the best parameter set for each classification model was performed within the cross-validation procedure.

Homogeneity assessment

The homogeneity analysis provides an indication of how well samples are clustered into separated groups. Homogeneity is based on the computation of *silhouette widths*, which also enables an intuitive

illustration of homogeneity by means of silhouette plots [19]. Briefly, the silhouette width of a sample compares the average distance of that sample from samples of the same group to its average distance from samples of other groups (**Figure 2a**).

Silhouette widths, here called *homogeneity score*, can be represented through silhouette plots (**Figure 2b** – right panel). Moreover, the average homogeneity score within a group provides a means to quantify how the samples in the group are similar to each other with respect to the other group: the higher the average homogeneity score is, the more homogeneous the groups is, and the more dissimilar are the elements of this group to the other group (**Figure 2b**). Details of this analysis are reported in **Additional File 1**.

We explored for biological heterogeneity (and its converse, homogeneity) in this prostate cancer data set and compared our findings with other tumor tissues. We defined heterogeneity in terms of the molecular signature by evaluating the “distance” between patients belonging to the same group, e.g. lethals, to that of patients belonging to different groups, e.g. indolents. Clearly, in homogeneous tissues, biopsy sampling is not an issue and patients belonging to the same group should be molecularly “closer” to each other than to those belonging to different groups. On the other hand, heterogeneous tissues should not show a clear separation as the molecular profiles of samples in both groups intermingle (**Figure 2b** – left panel).

We performed the homogeneity analysis on the prostate data set considering the two groups of lethal and indolent patients. Furthermore, we compared these results with 5 well-known publicly available data sets, with different levels of heterogeneity (see **Additional File 1** and **Additional File 1, Table S2**).

Results

Association with clinical variables

We first examined associations between clinical variables and outcome (see **Additional File 1, Table S1**). Gleason score, divided into 3 groups: 4-6; 7; and 8-10, showed the strongest association with outcome (Cramer's V: 0.45 and Fisher's exact test p-value = 6×10^{-14}). In this cohort, men with *ERG* rearranged prostate cancer were significantly more likely to be in the lethal class than the indolent class with an odds ratio of 7.2 (95% CI=[2.8,19.0]; Fisher's exact test p-value = 2.3×10^{-6}) (**Figure 3**).

Supervised analysis results

The results on the Learning set showed that no classification model clearly outperformed the others in predicting lethal cases (**Additional File 1, Table S3**). Indeed, most of them had similar performance. Therefore, to simply illustrate and summarize these findings, we report here the complete results of the logistic regression models (**Figure 3a**).

The molecular classifier alone achieved an AUC of 0.71 (95% CI = [0.67,0.75]) including 18 genes. Surprisingly, however, it did not perform better than models using only clinical features (AUC = 0.76; 95% CI = [0.67,0.84]) for the model with Gleason score). Moreover, when the model combines molecular and clinical features, no improvement over the clinical model was observed (AUC = 0.75; 95% CI = [0.71,0.79]) for the classifier comprising Age, Gleason score and 12 genes.

Gleason score was the most important clinical parameter as all the top models included Gleason score

in their classifiers. Although it is well known that inter-observer variability may affect this subjective parameter [20-22], the results demonstrate that it is a strong outcome predictor. Although differences among the top models were marginal, the best classifier of lethal prostate cancer included Gleason score and ERG rearrangement status (AUC = 0.79; 95% CI = [0.71,0.87]).

Lack of a significant improvement in prediction using the molecular profile suggested several possibilities. First, perhaps our definition of lethal and indolent prostate cancer does not capture the biological progression of the tumor. In order to assess how our definition of “extreme” cases affects the results, we ran several experiments by modifying the definition of lethals and indolents. **Additional File 1, Table S7** reports the results for DLDA. Similar results are obtained with the other classification models. When the definition of lethal or indolent is very stringent we can achieve some improvement. However, this is obtained at the expense of the number of cases that are classified. Moreover, with very stringent thresholds, we enriched for high and/or low Gleason scores in the two groups. Hence, although a better classification performance can be achieved, it is likely that no additional information about the more critical cases (Gleason score 7) can be obtained. Second, we reasoned that stroma-contaminated samples may have prevented us to discover a molecular signature of aggressive prostate cancer. Therefore, in order to seek for stroma-contaminated samples, we employed a molecular profile developed by Tomlins et al. [23] where they applied laser capture micro-dissection (LCM) to prostate tissues (see **Additional File 1** for details). We identified in our data set a cluster of samples exhibiting stroma-like profile based on a set of 47 top ranked common genes (see **Additional File 2, Figure S3**). These samples (n=17) were then excluded from the Learning set and the remaining samples were used as a new Learning set. The same iterative cross-validation procedure was employed for a SVM classifier (polynomial degree=1; cost=0.1; p-value=0.01) which achieved an AUC of 0.77 (95%C.I. [0.73-0.81]). We believe that this result, which is comparable to the one using the full set (see

Additional File 1, Table S3), is not sufficient to argue that stroma-contaminated tissue have prevented us to develop an accurate prediction model. Furthermore, we considered that, perhaps, the genes assayed on the this DASL array platform might not include the actual genes driving tumor progression. However, the 6K DASL gene set was developed specifically for this project. We selected genes showing the maximum variation in expression in 24 expression profiling studies from 15 different tumor types or because they were transcriptionally deregulated in previous prostate cancer studies. These genes cover most of the known pathways. Moreover, we demonstrated that this same platform and a slightly larger cohort can reliably identify a molecular signature for *ERG* rearrangement status [24]. Nevertheless, we performed an additional analysis by evaluating the consistency of the Gleason-score correlated genes (see **Additional File 1**) confirming its reliability. We thus favored that inter-tumor heterogeneity was the main reason and thus explored the potential impact of tissue heterogeneity by performing a homogeneity analysis.

Homogeneity analysis results

For prostate cancer, we computed homogeneity scores of the samples using a subset of the genes assessed on the array. We selected the genes that best distinguish the two groups, namely lethal and indolent prostate cancer, on the entire cohort of 281 patients, intentionally over-fitting the data to obtain the best molecular descriptors of the two groups. Specifically, genes were selected by a two-tailed t-test p-values after correcting for multiple hypothesis testing ($q\text{-value} < 0.05$), yielding 118 genes (see **Additional File 1, Table S5**)[25].

We performed the same analysis for other tumor data sets and compared the results with our data set. For illustration purposes, **Figure 4a** shows the silhouette plot for our prostate cancer data set compared

with the Burkitt's lymphoma data set [26]; whereas **Figure 4b** reports the results for all data sets. We compared prostate cancer with Burkitt's lymphoma because both harbor a recurrent translocation that leads to the over expression of two known oncogenes: c-MYC for Burkitt's lymphoma and ERG for prostate cancer.

The results support the heterogeneity hypothesis for prostate cancer. The average homogeneity score of the lethal group is lower than zero, meaning that on average, samples in the lethal group are more similar to samples in the indolent group. On the other hand, indolent cases seem to be slightly more homogeneous than lethal, as expected, although the average homogeneity score is rather low.

Conversely, the homogeneity scores on Burkitt's lymphoma data set are quite striking when compared with prostate cancer. Burkitt's lymphoma is a molecularly defined disease, with marked differences with respect to the broader class of lymphoma. Dave et al. identified a signature comprising 228 genes which is able to discriminate between Diffuse Large B-Cell Lymphoma (DLBCL) and Burkitt's lymphoma. This signature resulted in an average homogeneity score of 0.71, suggesting a *strong structure* of Burkitt's lymphoma. This is in contrast with the DLBCL group, which is more heterogeneous and consists of multiple sub-classes. The homogeneity analysis confirms this notion yielding an average homogeneity score of 0.34, interpreted as a *weak structure* (see **Additional File 1** for additional detail).

Among the other studies, AML and ALL show the highest degree of homogeneity with both classes scoring higher than 0.6, whereas breast and lung cancer are confirmed to be heterogeneous (**Figure 4b**). Similarly to prostate cancer, we selected the most informative genes separating the groups for each study. Specifically, the most informative genes of Sørli et al. [27] were selected by computing a

Wilcoxon test between ER+ and ER- samples and using a p-value cut-off of 0.01. Battacharjee et al.[28] identified 675 genes whose differential expression levels were the most highly reproducible. For the leukemia data set, we selected to top 50 genes according to the correlation-based score proposed by Golub et al.[29](see **Additional File 1** for more detail).

Homogeneity of ERG rearranged subclass

We recently reported a molecular signature including 87 genes characteristic of *ERG* rearranged cases in the same cohort of patients [24], which was also validated on a U.S. based cohort. The homogeneity analysis using this gene signature supports the hypothesis that *ERG* rearranged cases represent a distinct subclass, although we cannot extend this result for the entire population of *ERG* rearranged prostate cancers. Indeed, these cases show a homogeneity score of 0.39 (**Additional File 2, Figure S2**).

Discussion

Current prognostic models of prostate cancer, including PSA, Gleason score and clinical stage fail to accurately predict disease progression, especially for men with intermediate disease. Two large randomized trials evaluating the effect of PSA screening on prostate-cancer mortality, namely the Prostate, Lung, Colorectal, and Ovarian (PLCO) and the European Randomized Study of Screening for Prostate Cancer (ERSPC), showed that during the first decade of follow-up, PSA screening has at best a modest effect (20% relative reduction of PCA specific death in the ERSPC) on PCA mortality, with substantial risks of negative biopsy, over diagnosis and over-treatment [30, 31]. The need to better identify patients with a more aggressive disease is thus an open challenge given the clinical observation that prostate cancer is a heterogeneous disease. This observation is based on the experience of clinicians who witness men with localized disease that should fair well but on occasion do not and less commonly men with apparently aggressive disease who do well. How can we account for this clinical heterogeneity? We anticipated that a well-designed molecular study interrogating thousands of genes implicated in cancer and specifically prostate cancer would help us determine a molecular signature for lethal and indolent disease. Perhaps what is clinically referred to as “heterogeneity” really represents our inability through Gleason grading or other clinical attributes to untangle the key elements that would, if known, help us predict which men will succumb to disease progression. The findings of the current study and other recent studies described below point to a more concerning reality about what accounts for heterogeneity.

This study found that molecular predictors can distinguish aggressive from indolent prostate cancer similarly to models generated from Gleason score and other clinical parameters. However, by combining clinical and molecular data, we were not able to improve on known predictors. The

explanation is manifold. First, we must consider the important limitation of prostate cancer sampling. We know that a prostate gland harbors often up to 5 geographically distinct tumor nodules [32-36] and these nodules are often clonally distinct. If we consider the homogeneity of *ERG* rearrangement in circulating tumor cells (CTCs) [37] and that the ETS gene rearrangements occur early in the development of prostate cancer, as they are often seen in high-grade PIN [38], and, when observed, are present in all tumor cell within a nodule; then, we can consider this a possible marker of tumor clonality. Observations from three independent groups demonstrate that up to 50% of prostate cancers with multiple nodules have clonally distinct lesions [39-41]. This would strongly support why sampling of the “right” cancerous nodules is so critical in prostate cancer. A prostate needle biopsy or TURP sample may or may not capture the driving lesion leaving an important clone undetected. This inability to identify the molecularly dominant nodule (intra-tumor heterogeneity) would then help explain the “heterogeneity” observed in the clinical assessment at time of diagnosis with outcome.

However, if intra-tumor heterogeneity were the main explanation for our results, and inter-tumor heterogeneity, i.e. the presence of many alternative pathways which lead to lethality in prostate cancer, only marginal, then all cancer foci across individuals should share a similar molecular profile. How does sampling play a role in this scenario? The set of indolent prostate cancer samples is not affected by sampling, the set of lethal prostate cancer samples is affected, in that the lethal focus is 'sub-sampled'. Let assume that this causes a 50% dilution of the lethal molecular signal. Due to our study design and combinations of supervised and unsupervised analysis approaches, we should still have been able to detect the presence of a strong and consistent lethal signal, even if this was for a subset of the lethal prostate cancer population. Hence, we believe that our results are best explained by high degree of heterogeneity between lethal prostate cancers.

However, another possible alternative explanation for clinical heterogeneity might be that the lethal signature develops with the accumulation of molecular lesions over time and therefore may not be present at time of initial diagnosis in contrast to the homogeneity of *ERG* rearrangement in CTCs [37]. This would not be mutually exclusive from inter-tumor heterogeneity but could compound the problem. Finally, the molecular signature may be embedded in the adjacent non-cancerous stromal tissue as recently observed in hepatocellular cancer [42] or perhaps due to a host immune response to the tumor that might not be measurable by examination of the tumor sample. Regardless of what the mechanism or combination of mechanisms is, we are still faced with an inability to consistently detect the lethal molecular signature as observed in the current study.

Our study results are in fact consistent with other emerging data from U.S. cohorts using similar and different molecular platforms. Nakagawa et al. recently attempted to develop a biomarker panel to predict which men with rising PSA following surgery would progress with clinically significant disease [43]. They employed a case-control design where cases were defined as men with rising PSA who progressed within 5 years after initial surgery. Controls were men with rising PSA but no sign of clinical disease progression within the first 5 years following surgery. A total of 213 cases and 213 controls were used for this study and, similar to the current study, the cases and controls were divided into training and validation set. Although the results on the training set seemed promising (see **Additional File 1**), the validation phase showed mis-classifications in both directions and none of the models with molecular and clinical parameters performed better than an AUC of 0.75 [43].

Another recent study is significant because a two-phase biomarker development approach was used to classify long-term disease progression or death due to prostate cancer. Cheville et al. reported on a molecular classifier developed using a profile developed from tumor samples isolated by laser capture

micro-dissection [44]. They used quantitative RT-PCR to measure gene expression and cancer specific death following surgery or development of metastatic disease as the clinical endpoint. They used a 2-phase design with a training set of 157 high-risk patients and a validation set of 57 high-risk patients. Their results demonstrated that a model including topoisomerase-2a, cadherin-10, ETS genes involved in gene fusion (i.e., *ERG*, *ETV1*, and *ETV4*), and aneuploidy status had an AUC of 0.81 and 0.79 for training and validation sets, respectively.

Based on the published series (Nakagawa et al., Glinsky et al., Lapointe et al., Singh et al., Yu et al., Cheville et al.) and the current study, it is therefore impressive that all of these reports using different platforms and patient populations achieve similar results [43-49].

Although other explanations may be possible, we favor that inter-tumor heterogeneity plays a more critical role. The strongest evidence from the current study has to do with the association of *ERG* rearrangement status and lethality (see also Attard et al. [37]).

The association between *ERG* rearranged cases and the lethal phenotype suggests that *ETS* rearrangements describe a particularly aggressive subclass of prostate cancer. In the current study 41 of 46 *ERG* rearranged prostate cancers were lethal; the unadjusted odds ratio for lethal disease associated with *ERG* rearrangement status was 7.2 (95% CI 2.8-19.0). This confirms and extends observations from 111 men in the expectant management cohort from Örebro where men with *ERG* rearranged prostate cancer were significantly more likely to have lethal disease than men with fusion negative tumors (cumulative incidence ratio = 2.7, p-value < 0.01, 95% CI=[1.3,5.8]) [50]. From the United Kingdom, Attard et al. reported associations between *TMPRSS2-ERG* fusion with interstitial deletion and cause specific survival taking into account age, Gleason score, and pre-treatment PSA in a cohort

of 445 men conservatively treated for prostate cancer [51]. Interestingly, aneuploidy in combination with *TMPRSS2-ERG* fusion was associated with the worst clinical outcome (hazard ratio=6.10, 95% CI=[3.33,11.15], p-value<0.001, 25% survival at 8 years). The relatively low frequency of *ERG* rearrangement in this cohort may represent the admixture of peripheral zone tumors with a presumed *ERG* rearrangement frequency of 45% [52] and transition zone tumors with a significantly lower *ERG* rearrangement frequency [53].

Conclusions

In summary, this study attempted to identify a molecular signature for lethal prostate cancer. Molecular profiles developed in this study performed similar to clinical models and no model was identified that improved on the clinical models by including the profiling data. One significant result is the association of *ERG* rearrangement with lethality (OR=7.2 95% CI=[2.3,19.0], Fisher's exact test p-value= 2.3×10^{-6}). Although other explanations may be plausible, we believe that prostate cancer tumor heterogeneity is highly likely to be a major limitation in the development of a lethal prostate cancer signature.

This study underlines the importance of developing a better strategy to best capture the molecular complexity of prostate cancer. One possibility could be using circulating tumor cells, known as liquid biopsies, to reduce the confounding effect of sampling multiple tumor nodules in a prostate gland and improve the current biopsy strategy [54, 55]. After which, we might be able to focus on characterizing the multiple lethal signatures that may exist.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ASB carried out the supervised analysis related to SVM, NN and logistic regression; conceived the homogeneity analysis and implemented the silhouette widths and plots; and drafted the manuscript. FD participated in the design of the supervised analysis and carried out the SVM and NN analysis. SC and YP participated in the design of the supervised analysis and carried out the DLDA and logistic regression. YP also coordinated the supervised analysis. SRS characterized the *ERG* rearrangement status of the samples and helped to draft the manuscript. YH participated in the generation of the expression measurements and carried out the supervised analysis with NTP and k-NN. SP evaluated the pathology specimen of all the samples, participated in the characterization of *ERG* rearrangements and help to draft the manuscript. HOA participated in the design of the study and its coordination. KF helped to organize the selection of the patients and to draft the manuscript. LAM participated in the statistical analysis of the results. PWK participated in the design and coordination of the study. MS participated in the design of the study and helped the statistical analysis of the results. SOA, EV, JEJ participated in the design of the study, collected and curated the samples. MBG participated in the supervised and the homogeneity analysis and helped to draft the manuscript. TRG participated in the design of the study and helped coordinating the gene expression measurements. MAR participated in the study design and in its coordination, evaluated the pathology specimen, helped to draft the manuscript. OA participated in the study design and its coordination.

All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge the National Cancer Institute (NCI) grant P50 90381 support for the Dana Farber/Harvard Cancer Center Prostate S.P.O.R.E. and the National Institute of Health (NIH) grant RR19895 for the Yale University Biomedical High Performance Computing Center.

References

1. Johansson J, Andrén O, Andersson S, Dickman PW, Holmberg L, Magnuson A, Adami H: **Natural history of early, localized prostate cancer.** *JAMA* 2004, **291**:2713-9.
2. Bill-Axelsson A, Holmberg L, Filen F, Ruutu M, Garmo H, Busch C, Nordling S, Haggman M, Andersson S, Bratell S, Spangberg A, Palmgren J, Adami H, Johansson J, for the Scandinavian Prostate Cancer Group Study Number 4: **Radical Prostatectomy Versus Watchful Waiting in Localized Prostate Cancer: the Scandinavian Prostate Cancer Group-4 Randomized Trial.** *J. Natl. Cancer Inst.* 2008, **100**:1144-1154.
3. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ: **Cancer Statistics, 2008.** *CA Cancer J Clin* 2008, **58**:71-96.
4. Bill-Axelsson A, Holmberg L, Ruutu M, Häggman M, Andersson S, Bratell S, Spångberg A, Busch C, Nordling S, Garmo H, Palmgren J, Adami H, Norlén BJ, Johansson J: **Radical prostatectomy versus watchful waiting in early prostate cancer.** *N Engl J Med* 2005, **352**:1977-84.
5. Kattan M, Eastham J, Stapleton A, Wheeler T, Scardino P: **A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer.** *J. Natl. Cancer Inst.* 1998, **90**:766-771.
6. Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD: **Contemporary update of**

prostate cancer staging nomograms (Partin Tables) for the new millennium. *Urology* 2001, **58**:843-848.

7. Kattan MW, Cuzick J, Fisher G, Berney DM, Oliver T, Foster CS, Møller H, Reuter V, Fearn P, Eastham J, Scardino PT, Group ATTP: **Nomogram incorporating PSA level to predict cancer-specific survival for men with clinically localized prostate cancer managed without curative intent.** *Cancer* 2008, **112**:69-74.

8. Aus G, Robinson D, Rosell J, Sandblom G, Varenhorst E: **Survival in prostate carcinoma--outcomes from a prospective, population-based cohort of 8887 men with up to 15 years of follow-up: results from three countries in the population-based National Prostate Cancer Registry of Sweden.** *Cancer* 2005, **103**:943-51.

9. Andren O, Fall K, Franzen L, Andersson S, Johansson J, Rubin MA: **How Well Does the Gleason Score Predict Prostate Cancer Death? A 20-Year Followup of a Population Based Cohort in Sweden.** *The Journal of Urology* 2006, **175**:1337-1340.

10. Epstein JI, Srigley J, Grignon D, Humphrey P: **Recommendations for the reporting of prostate carcinoma: Association of Directors of Anatomic and Surgical Pathology.** *Am J Clin Pathol* 2008, **129**:24-30.

11. Fan J, Yeakley JM, Bibikova M, Chudin E, Wickham E, Chen J, Doucet D, Rigault P, Zhang B, Shen R, McBride C, Li H, Fu X, Oliphant A, Barker DL, Chee MS: **A Versatile Assay for High-Throughput Gene Expression Profiling on Universal Array Matrices.** *Genome Res.* 2004, **14**:878-

885.

12. Bibikova M, Talantov D, Chudin E, Yeakley JM, Chen J, Doucet D, Wickham E, Atkins D, Barker D, Chee M, Wang Y, Fan J: **Quantitative Gene Expression Profiling in Formalin-Fixed, Paraffin-Embedded Tissues Using Universal Bead Arrays.** *Am J Pathol* 2004, **165**:1799-1807.
13. Duda RO, Hart PE, Stork DG: *Pattern Classification*. 2nd edition. New York, NY: John Wiley and Sons; 2001.
14. Xu L, Shen SS, Hoshida Y, Subramanian A, Ross K, Brunet J, Wagner SN, Ramaswamy S, Mesirov JP, Hynes RO: **Gene Expression Changes in an Animal Melanoma Model Correlate with Aggressiveness of Human Melanoma Metastases.** *Mol Cancer Res* 2008, **6**:760-769.
15. Dudoit S, Fridlyand J, Speed TP: **Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.** *Journal of the American Statistical Association* 2002, **97**:77-87.
16. Vapnik VN: *Statistical Learning Theory*. New York, NY: Wiley-Interscience; 1998.
17. Agresti A: *An Introduction to Categorical Data Analysis*. 2nd edition. Hoboken, New Jersey: Wiley-Interscience; 2007.
18. Varma S, Simon R: **Bias in error estimation when using cross-validation for model selection.** *BMC Bioinformatics* 2006, **7**:91.

19. Rousseeuw PJ: **Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.** *Journal of Computational and Applied Mathematics* 1987, **20**:53-65.
20. De la Taille A, Viellefond A, Berger N, Boucher E, De Fromont M, Fondimare A, Molinié V, Piron D, Sibony M, Staroz F, Triller M, Peltier E, Thiounn N, Rubin MA: **Evaluation of the interobserver reproducibility of Gleason grading of prostatic adenocarcinoma using tissue microarrays.** *Hum Pathol* 2003, **34**:444-9.
21. Evans AJ, Henry PC, Van der Kwast TH, Tkachuk DC, Watson K, Lockwood GA, Fleshner NE, Cheung C, Belanger EC, Amin MB, Boccon-Gibod L, Bostwick DG, Egevad L, Epstein JI, Grignon DJ, Jones EC, Montironi R, Moussa M, Sweet JM, Trpkov K, Wheeler TM, Srigley JR: **Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens.** *Am J Surg Pathol* 2008, **32**:1503-12.
22. Burchardt M, Engers R, Müller M, Burchardt T, Willers R, Epstein JI, Ackermann R, Gabbert HE, de la Taille A, Rubin MA: **Interobserver reproducibility of Gleason grading: evaluation using prostate cancer tissue microarrays.** *J Cancer Res Clin Oncol* 2008, **134**:1071-8.
23. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM: **Integrative molecular concept modeling of prostate cancer progression.** *Nat Genet* 2007, **39**:41-51.
24. Setlur SR, Mertz KD, Hoshida Y, Demichelis F, Lupien M, Perner S, Sboner A, Pawitan Y, Andren

O, Johnson LA, Tang J, Adami H, Calza S, Chinnaiyan AM, Rhodes D, Tomlins S, Fall K, Mucci LA, Kantoff PW, Stampfer MJ, Andersson S, Varenhorst E, Johansson J, Brown M, Golub TR, Rubin MA: **Estrogen-Dependent Signaling in a Molecularly Distinct Subclass of Aggressive Prostate Cancer.** *J. Natl. Cancer Inst.* 2008, **100**:815-825.

25. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289-300.

26. Dave SS, Fu K, Wright GW, Lam LT, Kluin P, Boerma E, Greiner TC, Weisenburger DD, Rosenwald A, Ott G, Muller-Hermelink H, Gascoyne RD, Delabie J, Rimsza LM, Braziel RM, Grogan TM, Campo E, Jaffe ES, Dave BJ, Sanger W, Bast M, Vose JM, Armitage JO, Connors JM, Smeland EB, Kvaloy S, Holte H, Fisher RI, Miller TP, Montserrat E, Wilson WH, Bahl M, Zhao H, Yang L, Powell J, Simon R, Chan WC, Staudt LM, the Lymphoma/Leukemia Molecular Profiling Project: **Molecular Diagnosis of Burkitt's Lymphoma.** *N Engl J Med* 2006, **354**:2431-2442.

27. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale A: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proceedings of the National Academy of Sciences* 2001, **98**:10869-10874.

28. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals**

distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U.S.A* 2001, **98:13790-13795.**

29. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.** *Science* 1999, **286**:531-537.

30. Andriole GL, Grubb RL, Buys SS, Chia D, Church TR, Fouad MN, Gelmann EP, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, Crawford ED, O'Brien B, Clapp JD, Rathmell JM, Riley TL, Hayes RB, Kramer BS, Izmirlian G, Miller AB, Pinsky PF, Prorok PC, Gohagan JK, Berg CD, the PLCO Project Team: **Mortality Results from a Randomized Prostate-Cancer Screening Trial.** *N Engl J Med* 2009: **360**(13):1310-9.

31. Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, Kwiatkowski M, Lujan M, Lilja H, Zappa M, Denis LJ, Recker F, Berenguer A, Maattanen L, Bangma CH, Aus G, Villers A, Rebillard X, van der Kwast T, Blijenberg BG, Moss SM, de Koning HJ, Auvinen A, the ERSPC Investigators: **Screening and Prostate-Cancer Mortality in a Randomized European Study.** *N Engl J Med* 2009: **360**(13):1320-8.

32. Greene DR, Wheeler TM, Egawa S, Dunn JK, Scardino PT: **A comparison of the morphological features of cancer arising in the transition zone and in the peripheral zone of the prostate.** *J Urol* 1991, **146**:1069-76.

33. Sakr WA, Macoska JA, Benson P, Grignon DJ, Wolman SR, Pontes JE, Crissman JD: **Allelic loss in locally metastatic, multisampled prostate cancer.** *Cancer Res* 1994, **54**:3273-7.

34. Qian J, Bostwick DG, Takahashi S, Borell TJ, Herath JF, Lieber MM, Jenkins RB: **Chromosomal anomalies in prostatic intraepithelial neoplasia and carcinoma detected by fluorescence in situ hybridization.** *Cancer Res* 1995, **55**:5408-14.
35. Cheng L, Song SY, Pretlow TG, Abdul-Karim FW, Kung HJ, Dawson DV, Park WS, Moon YW, Tsai ML, Linehan WM, Emmert-Buck MR, Liotta LA, Zhuang Z: **Evidence of independent origin of multiple tumors from patients with prostate cancer.** *J Natl Cancer Inst* 1998, **90**:233-7.
36. Arora R, Koch MO, Eble JN, Ulbright TM, Li L, Cheng L: **Heterogeneity of Gleason grade in multifocal adenocarcinoma of the prostate.** *Cancer* 2004, **100**:2362-6.
37. Attard G, Swennenhuis JF, Olmos D, Reid AH, Vickers E, A'Hern R, Levink R, Coumans F, Moreira J, Riisnaes R, Oommen NB, Hawche G, Jameson C, Thompson E, Sipkema R, Carden CP, Parker C, Dearnaley D, Kaye SB, Cooper CS, Molina A, Cox ME, Terstappen LW, de Bono JS: **Characterization of ERG, AR and PTEN Gene Status in Circulating Tumor Cells from Patients with Castration-Resistant Prostate Cancer.** *Cancer Res* 2009, **69**:2912-2918.
38. Mosquera J, Perner S, Genega EM, Sanda M, Hofer MD, Mertz KD, Paris PL, Simko J, Bismar TA, Ayala G, Shah RB, Loda M, Rubin MA: **Characterization of TMPRSS2-ERG fusion high-grade prostatic intraepithelial neoplasia and potential clinical implications.** *Clin Cancer Res* 2008, **14**:3380-5.
39. Barry M, Perner S, Demichelis F, Rubin MA: **TMPRSS2-ERG fusion heterogeneity in multifocal**

prostate cancer: clinical and biologic implications. *Urology* 2007, **70**:630-3.

40. Mehra R, Han B, Tomlins SA, Wang L, Menon A, Wasco MJ, Shen R, Montie JE, Chinnaiyan AM, Shah RB: **Heterogeneity of TMPRSS2 Gene Rearrangements in Multifocal Prostate Adenocarcinoma: Molecular Evidence for an Independent Group of Diseases.** *Cancer Res* 2007, **67**:7991-7995.

41. Clark J, Attard G, Jhavar S, Flohr P, Reid A, De-Bono J, Eeles R, Scardino P, Cuzick J, Fisher G, Parker MD, Foster CS, Berney D, Kovacs G, Cooper CS: **Complex patterns of ETS gene alteration arise during cancer development in the human prostate.** *Oncogene* 2008, **27**:1993-2003.

42. Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, Gupta S, Moore J, Wrobel MJ, Lerner J, Reich M, Chan JA, Glickman JN, Ikeda K, Hashimoto M, Watanabe G, Daidone MG, Roayaie S, Schwartz M, Thung S, Salvesen HB, Gabriel S, Mazzaferro V, Bruix J, Friedman SL, Kumada H, Llovet JM, Golub TR: **Gene Expression in Fixed Tissues and Outcome in Hepatocellular Carcinoma.** *N Engl J Med* 2008, **359**:1995-2004.

43. Nakagawa T, Kollmeyer TM, Morlan BW, Anderson SK, Bergstralh EJ, Davis BJ, Asmann YW, Klee GG, Ballman KV, Jenkins RB: **A Tissue Biomarker Panel Predicting Systemic Progression after PSA Recurrence Post-Definitive Prostate Cancer Therapy.** *PLoS ONE* 2008, **3**:e2318.

44. Cheville JC, Karnes RJ, Therneau TM, Kosari F, Munz J, Tillmans L, Basal E, Rangel LJ, Bergstralh E, Kovtun IV, Savci-Heijink C, Klee EW, Vasmatazis G: **Gene Panel Model Predictive of Outcome in Men at High-Risk of Systemic Progression and Death From Prostate Cancer After**

Radical Retropubic Prostatectomy. *J Clin Oncol* 2008, **26**:3930-3936.

45. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** *Cancer Cell* 2002, **1**:203-209.

46. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, McDonald C, Thomas R, Dhir R, Finkelstein S, Michalopoulos G, Becich M, Luo J: **Gene Expression Alterations in Prostate Cancer Predicting Tumor Aggression and Preceding Development of Malignancy.** *J Clin Oncol* 2004, **22**:2790-2799.

47. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer.** *Proc Natl Acad Sci USA* 2004, **101**:811-816.

48. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL: **Gene expression profiling predicts clinical outcome of prostate cancer.** *J. Clin. Invest.* 2004, **113**:913-923.

49. Glinsky GV, Berezovska O, Glinskii AB: **Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer.** *J Clin Invest.* 2005, **115**:1503–1521.

50. Demichelis F, Fall K, Perner S, Andren O, Schmidt F, Setlur SR, Hoshida Y, Mosquera J, Pawitan Y, Lee C, Adami H, Mucci LA, Kantoff PW, Andersson S, Chinnaiyan AM, Johansson J, Rubin MA:

TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort.

Oncogene 2007, **26**:4596-4599.

51. Attard G, Clark J, Ambroisine L, Fisher G, Kovacs G, Flohr P, Berney D, Foster CS, Fletcher A, Gerald WL, Moller H, Reuter V, De Bono JS, Scardino P, Cuzick J, Cooper CS: **Duplication of the fusion of TMPRSS2 to ERG sequences identifies fatal human prostate cancer.** *Oncogene* 2008, **27**:253-263.

52. Mosquera J, Mehra R, Regan MM, Perner S, Genega EM, Bueti G, Shah RB, Gaston S, Tomlins SA, Wei JT, Kearney MC, Johnson LA, Tang JM, Chinnaiyan AM, Rubin MA, Sanda MG: **Prevalence of TMPRSS2-ERG fusion prostate cancer among men undergoing prostate biopsy in the United States.** *Clin. Cancer Res* 2009, **15**:4706-4711.

53. Guo CC, Zuo G, Cao D, Troncoso P, Czerniak BA: **Prostate cancer of transition zone origin lacks TMPRSS2-ERG gene fusion.** *Mod Pathol* 2009, **22**:866-871.

54. Nagrath S, Sequist LV, Maheswaran S, Bell DW, Irimia D, Ulkus L, Smith MR, Kwak EL, Digumarthy S, Muzikansky A, Ryan P, Balis UJ, Tompkins RG, Haber DA, Toner M: **Isolation of rare circulating tumour cells in cancer patients by microchip technology.** *Nature* 2007, **450**:1235-1239.

55. Maheswaran S, Sequist LV, Nagrath S, Ulkus L, Brannigan B, Collura CV, Inserra E, Diederichs S, Iafrate AJ, Bell DW, Digumarthy S, Muzikansky A, Irimia D, Settleman J, Tompkins RG, Lynch TJ, Toner M, Haber DA: **Detection of Mutations in EGFR in Circulating Lung-Cancer Cells.** *N Engl J Med* 2008, **359**:366-377.

Figure legends

Figure 1: Study design. From 1256 men of a Watchful Waiting Cohort, we selected the “Extreme” cases: those who died of prostate cancer or men who lived more than 10 years without signs of progression. We also filtered out some patients based on tumor tissue availability, sample quality or because they were treated. Finally, we randomly divided the patients in a Learning and Validation sets, ensuring that similar proportions of lethals and indolents are present in the two groups.

Figure 2: Schematic of silhouette widths, i.e. homogeneity scores, and silhouette plots.

A. (left) Given an element in a group (the orange cross surrounded by a diamond) the distances from elements in the same group (magenta lines) and from those in the other group (green lines) are computed. The homogeneity score can be viewed as the difference between the averages of the inter-group distance (green) and the intra-group distance (magenta). (right) The homogeneity score of each sample is plotted on a horizontal bar, after sorting the samples within each group. The average of the homogeneity scores is computed for each group yielding an estimation of the homogeneity of the cluster. **B.** Four different categories of homogeneity (left) and the corresponding silhouette plots (right) are depicted. Specifically: Scenario 1. two homogeneous and well-separated groups; Scenario 2. one homogeneous and one heterogeneous group, well-separated; Scenario 3. one homogeneous and one heterogeneous group, overlapping; Scenario 4. two heterogeneous overlapping groups. The empirical interpretation of the average homogeneity score for a group is shown at the bottom.

Figure 3: Supervised analysis. A. Results of logistic regression on the Validation dataset. On top are reported the AUCs of the models, whereas on the bottom the parameters that are used in the corresponding model are shown. A colored square means that the parameter was used in the model, whereas a white square means that the parameter was not used. The last row reports the number of genes that were used by the model, if any. Models including clinical and molecular parameters are reported only if they improved on the corresponding models using clinical parameters only. Models are sorted from left to right according to their AUC. We estimated the Confidence intervals (CIs) for models including genes using the sampling distribution of AUCs generated by the iterative cross-validation procedure on the Learning set. For the other models, a bootstrap estimation of CIs was computed on the Validation set. The genes that are involved in the models are reported in Additional file 1, Table S4. **B.** Contingency table showing ERG rearrangement status association with clinical outcome. In parenthesis the expected numbers of cases if no association is assumed.

Figure 4. Homogeneity analysis. A. Silhouette plot for Burkitt's lymphoma (left) and prostate cancer (right). The numbers report the average homogeneity score for each group. **B.** Average homogeneity score for different cancer data sets.

Additional files

Additional file 1:

Title: Supplementary material

Description: This file contains additional information regarding the experimental protocols, the supervised data analysis and the homogeneity data analysis, as well as additional results to further support the main conclusions.

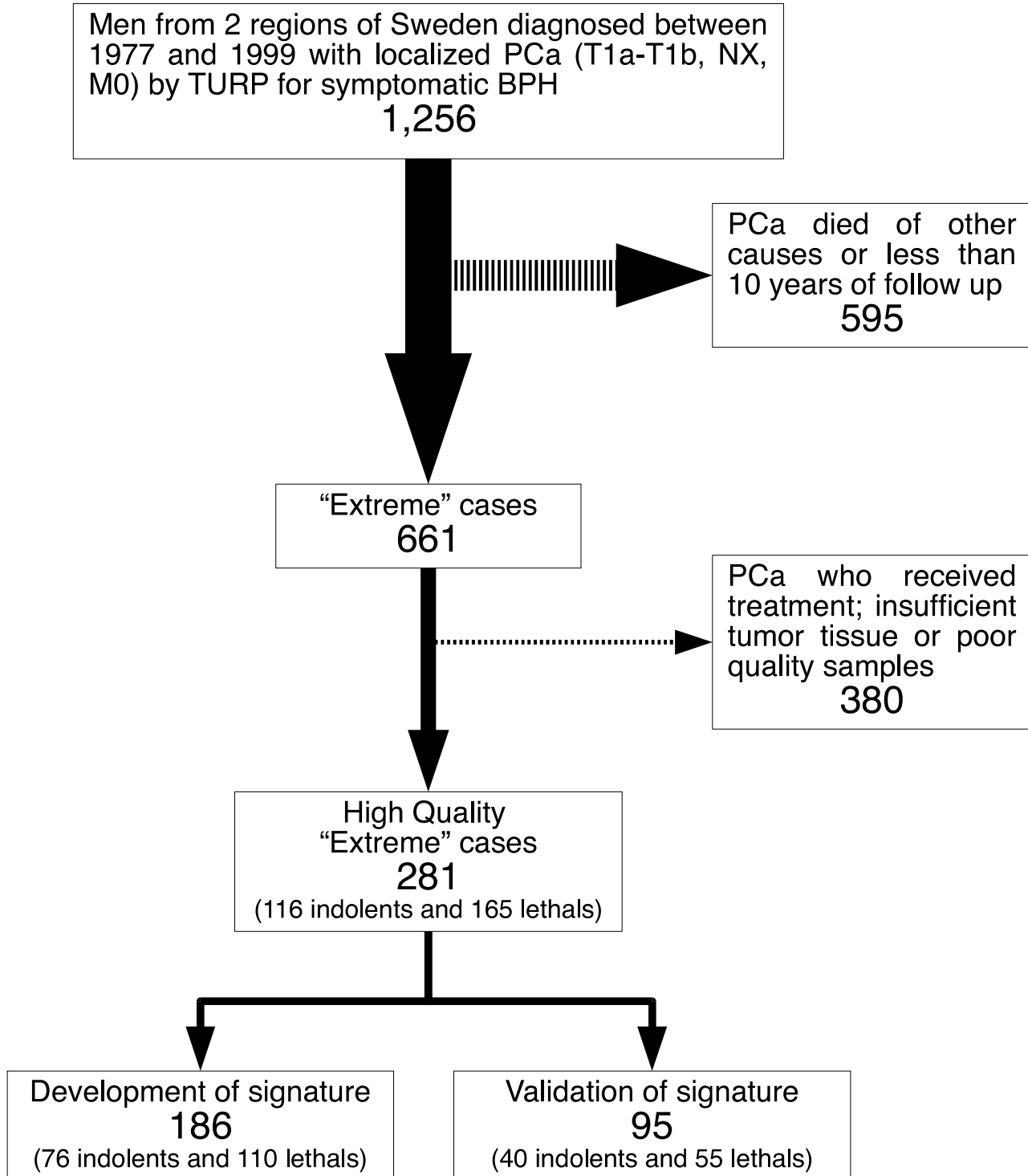
Additional file 2:

Title: Supplementary figures.

Description: This file contains supplementary figures.

“Extreme” cases study design

Swedish Watchful Waiting Cohort



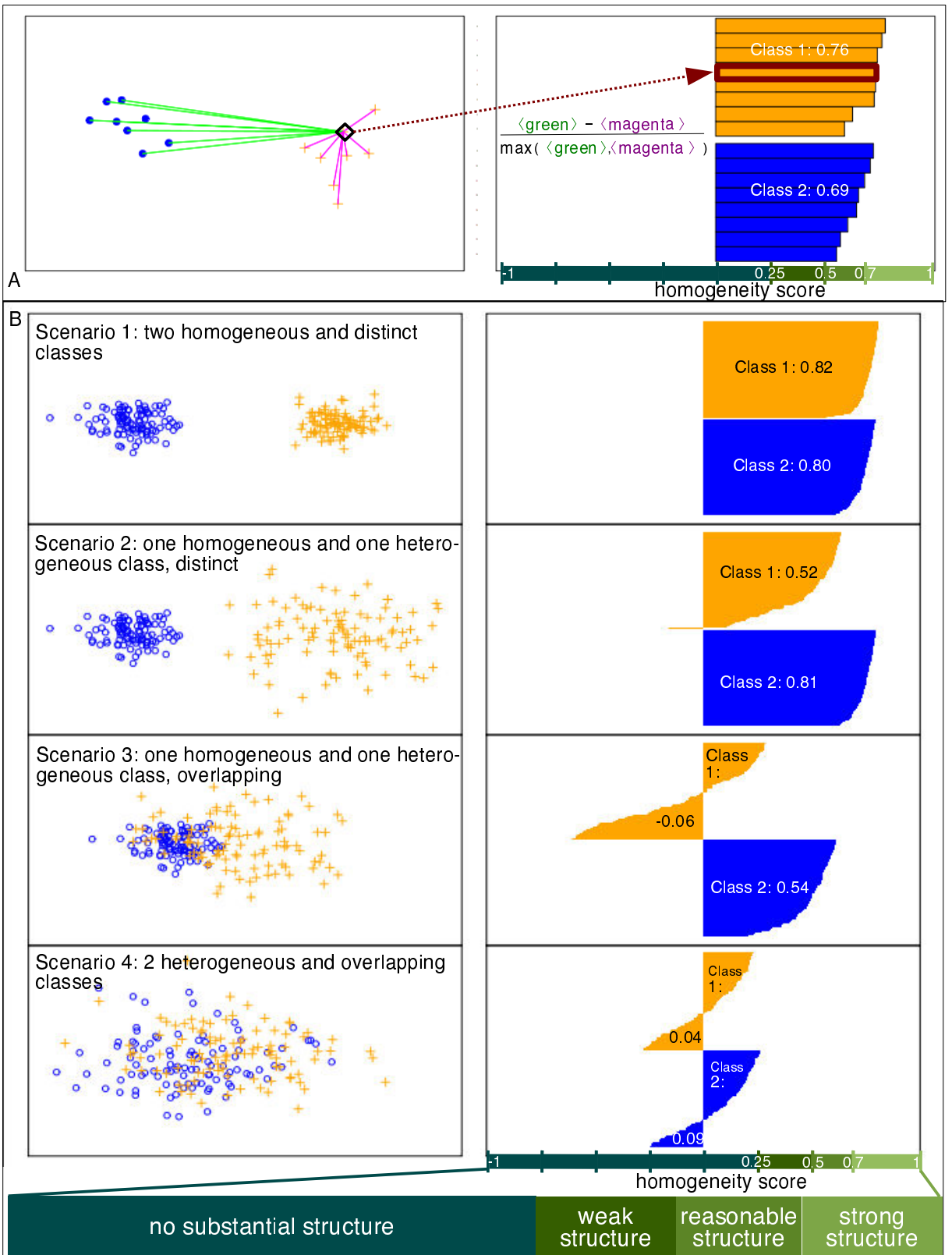
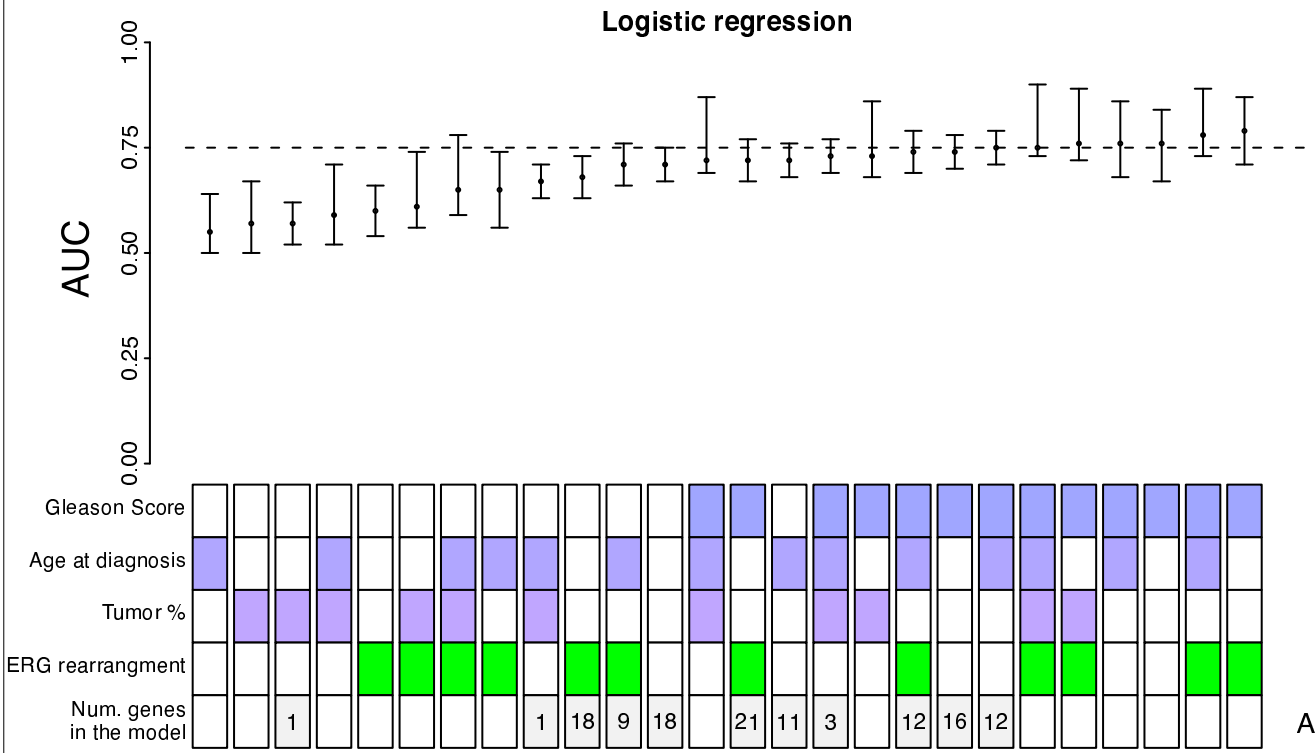


Figure 2



		ERG rearrangement status		
		<i>Negative</i>	<i>Positive</i>	<i>Total</i>
Extreme group	<i>Indolent</i>	106 (92)	5 (19)	111
	<i>Lethal</i>	120 (134)	41 (27)	161
	<i>Total</i>	226	46	272

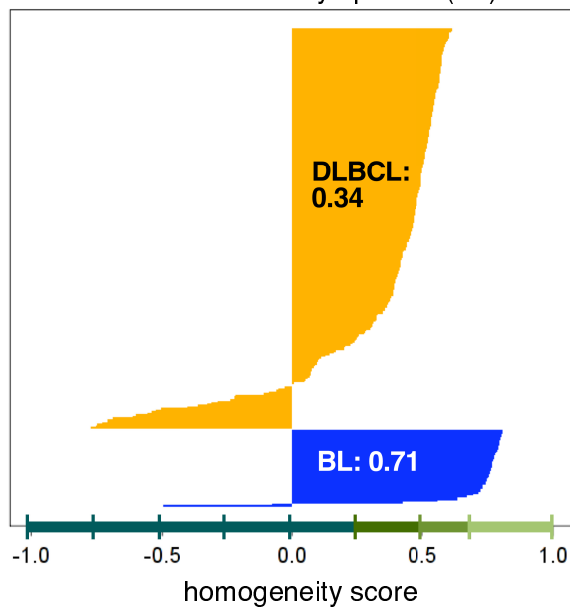
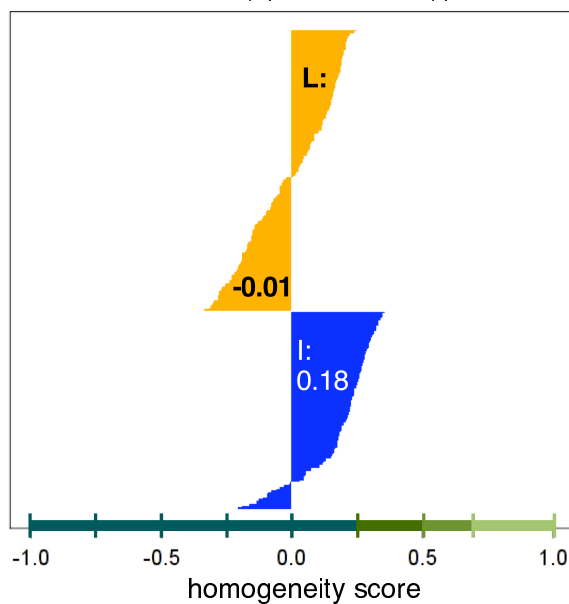
OR = 7.2 95% CI [2.3,19.0]
Fisher's exact test p-value = 2.3×10^{-6}

B

Figure 3

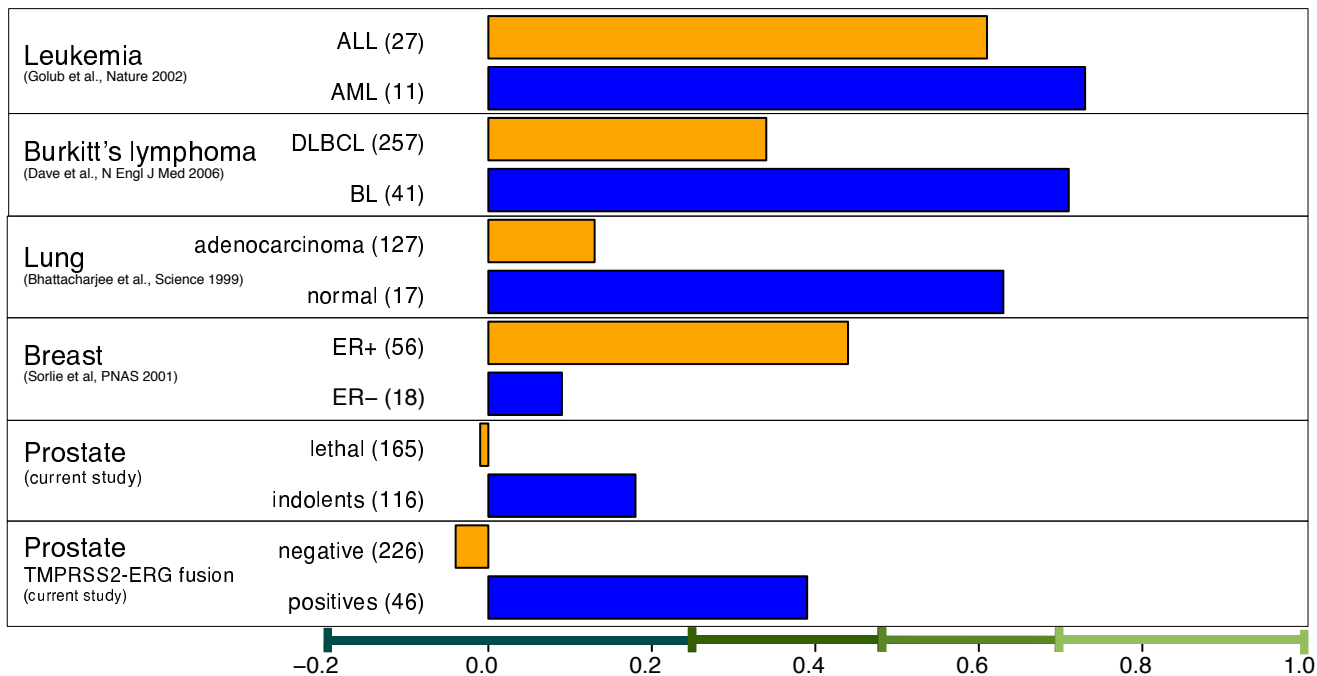
A

Silhouette Plots

Diffuse Large B-Cell Lymphoma (DLBCL)
vs Burkitt's Lymphoma (BL)Prostate cancer:
Lethal (L) vs Indolent (I)

B

Average Homogeneity Score



Additional files provided with this submission:

Additional file 1: Suppl_Sboner_PCa_heterogeneity.doc, 234K

<http://www.biomedcentral.com/imedia/1205665330363679/supp1.doc>

Additional file 2: Suppl_Sboner_PCa_heterogeneity_figureS1-S4.pdf, 299K

<http://www.biomedcentral.com/imedia/9133988933490561/supp2.pdf>