Published in final edited form as:
J Probab Stat. 2012 ; 2012(2012): 873570–. doi:10.1155/2012/873570.

Clustering-Based Method for Developing a Genomic Copy Number Alteration Signature for Predicting the Metastatic Potential of Prostate Cancer

Alexander Pearlman[1], **Christopher Campbell**[1], **Eric Brooks**[2], **Alex Genshaft**[2], **Shahin Shajahan**[2], **Michael Ittman**[3], **G. Steven Bova**[4], **Jonathan Melamed**[5], **Ilona Holcomb**[6], **Robert J. Schneider**[7], and **Harry Ostrer**[1]

[1]Department of Pathology, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[2]Human Genetics Program, Department of Pediatrics, NYU Langone Medical Center, New York, NY 10016, USA

[3]Department of Pathology, Baylor College of Medicine, Houston, TX 77030, USA

[4]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

[5]Department of Pathology, NYU Langone Medical Center, New York, NY 10016, USA

[6]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA

[7]NYU Cancer Institute and Department of Microbiology, NYU Langone Medical Center, New York, NY 10016, USA

Abstract

The transition of cancer from a localized tumor to a distant metastasis is not well understood for prostate and many other cancers, partly, because of the scarcity of tumor samples, especially metastases, from cancer patients with long-term clinical follow-up. To overcome this limitation, we developed a semi-supervised clustering method using the tumor genomic DNA copy number alterations to classify each patient into inferred clinical outcome groups of metastatic potential. Our data set was comprised of 294 primary tumors and 49 metastases from 5 independent cohorts of prostate cancer patients. The alterations were modeled based on Darwin's evolutionary selection theory and the genes overlapping these altered genomic regions were used to develop a metastatic potential score for a prostate cancer primary tumor. The function of the proteins encoded by some of the predictor genes promote escape from anoikis, a pathway of apoptosis, deregulated in metastases. We evaluated the metastatic potential score with other clinical predictors available at diagnosis using a Cox proportional hazards model and show our proposed score was the only significant predictor of metastasis free survival. The metastasis gene signature and associated

Correspondence should be addressed to Alexander Pearlman, apearlman@gmail.com.

score could be applied directly to copy number alteration profiles from patient biopsies positive for prostate cancer.

1. Introduction

Prostate cancer is a common public health problem. In 2012, this disease was expected to be diagnosed in an estimated 241,740 men (29% of all male cancers) and to result in 28,170 deaths (9% of male cancer deaths) [1]. If left untreated, around 70% of prostate cancers remain asymptomatic and indolent for decades [2]. If treated with radical prostatectomy or radiation therapy, the risk of metastasis is reduced, but erectile dysfunction, urinary incontinence, and rectal bleeding may occur, affecting the patient's quality of life. Because it is currently difficult to determine accurately which patients will develop metastatic disease, physicians treat patients with mid-to-late stage local disease aggressively, even when such treatment may not be required. Clinical parameters, such as, serum concentration of prostate-specific antigen (PSA), extension beyond surgical margins, invasion of seminal vesicles, extension beyond the capsule, surgical Gleason score, prostate weight, race, and year of surgery, are employed in existing nomograms for prediction of local recurrences after surgery [3], but, many of these parameters are not available at diagnosis and cannot be used for guiding therapeutic decisions. Development of a robust risk model from a biopsy that accurately predicts the potential of a local prostate cancer to metastasize would justify aggressive treatment in high-risk cases and improve the quality of life for men with indolent disease by allowing them to avoid treatment-related side effects. Thus, the goal of this study was to develop a method to identify tumor genomic biomarkers that could be applied to prediction models that help guide clinical treatment decisions.

The method chosen for developing the predictive model was the analysis of genomic DNA copy number alterations (CNAs) in prostate cancers, because these cancers have long been known to harbor multiple genomic imbalances that result from CNAs [4, 5]. High-resolution measurements of CNAs have functional value, in some cases providing evidence for alterations in the quantity of normal, mutant, or hybrid-fusion transcripts and proteins in the cancer cells. The resulting changes in abundance or altered structure of RNA transcripts and proteins (e.g., truncating dominant negative mutations) may impact the fitness of the cell and provide some of the mechanisms necessary for distant site migration, invasion, and growth. From the multiple CNAs identified in tumors, CNA-based gene signatures were developed into a score that suggested the ability to predict metastasis free survival.

2. Methods

2.1. Cohorts and Samples

We studied four publically available prostate cancer cohorts and a fifth cohort reported here: (1) 294 primary tumors and matched normal tissue samples from NYU School of Medicine (NYU $n = 29$), Baylor College of Medicine (Baylor $n = 20$) [6], Memorial Sloan-Kettering Cancer Center (MSK $n = 181$) [7], and Stanford University (SU $n = 64$ (single reference used for each tumor)) [8]. (2) 49 metastatic tumors and matched normal samples from Johns Hopkins School of Medicine (Hopkins $n = 13$) [9] and MSK ($n = 36$) [7]. The 13 patients in

the Hopkins cohort had multiple metastases dissected at autopsy, totaling 55 samples for the study. We also studied a sixth, publically available cohort of 337 cell lines originating from varying tumor cell types (ArrayExpress ID: E-MTAB-38).

2.2. Sample Processing

Genomic DNA (gDNA) from the NYU cohort was extracted from fresh-frozen prostate tumors using a Gentra DNA extraction kit (Qiagen). Purified gDNA was hydrated in reduced TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0). The gDNA concentration was measured using the NanoDrop 2000 spectrophotometer at optical density (OD) wavelength of 260 nm. Protein and organic contaminations were measured at OD 280 nm and 230 nm, respectively. Samples that passed OD quality control thresholds were then run on a 1% agarose gel to assess the integrity of the gDNA. 500 ng of gDNA samples was run on the Affymetrix Human SNP Array 6.0 at the Rockefeller University Genomics Resource Center using standard operating procedures. Samples that were obtained from public sources were processed according to the methods outlined in their respective publications. Affymetrix .cel files were processed using the Birdseed v2 algorithm [10].

2.3. Study Design

The case samples in this study were either metastatic tumors (METS) or primary tumors from men treated with radical prostatectomy that were clinically followed up and reported to develop distant metastases (mPTs). METS and mPTs are clearly discernible phenotypes that can be classified unequivocally as cases. The control samples were defined as primary tumors that had not progressed to form distant metastases following radical prostatectomy either because clinical followup was not available or because the treatment rendered the patient not informative for this outcome. Radical prostatectomy treats both indolent primary tumors (iPTs) that would not metastasize and primary tumors that would otherwise progress to form metastases, if left untreated. Thus, the control primary tumors actually represent a mixture of iPTs and unrealized mPTs. Assuming a randomly sampled cohort, it is expected that about 30% of the control group of primary tumors would be unrealized mPTs [2]. Considering the scarcity of clinically informative mPTs and iPTs for study, our strategy for identifying CNA biomarkers from tumors with inferred metastatic outcomes allowed a greater number of individual genomes to be used. Accordingly, all of the clinically informative mPTs available to us were not used to identify the biomarkers and only tested in a Cox proportional hazard model to assess the clinical usefulness of these predictors. Future tumor cohort study design using the method presented in this paper should consider the prevalence of metastatic progression to assure a large enough representation of both mPTs and iPTs. The natural history of prostate cancer, without medical intervention, (e.g., watchful waiting or active surveillance) is well documented [2]. Assuming a randomly sampled cohort, this information allowed us to estimate the prevalence of mPTs to be 30%.

2.4. Cancer Genomics Copy Number Algorithm

A genomic DNA copy number analysis pipeline (Figure 1) was designed using the R-statistical software [11] (R) to process the raw intensity data through a series of computational steps resulting in ranked lists of genes and associated significance that could be used for functional mining and prediction model development. The R-package will be

provided upon request and raw and processed data can be obtained from Gene Expression Omnibus accession# GSE27105.

2.5. Raw Data Processing

Signal intensity files (.cel) for the Affymetrix SNP Array 6.0 or 500 k mapping arrays were processed using the Affymetrix Power Tools, Birdseed V2 [10], and BRLMM [12] algorithms, respectively, resulting in genotype allele calls and signal intensity measures for each SNP and copy number probe. After the first stage, the genotype calls were prepared for downstream principal component analysis for ethnic identification and quality control testing, especially important when investigating racially driven health disparities (Figure 2). Men of African descent have an increased incidence, earlier onset, and more aggressive form of the disease than those of European origin. Even when adjusted for the increased level of incidence in African Americans, the mortality rate of African American men is more than twice that of Caucasian men [1]. Although not presented in the current work, sophisticated CNA models of metastatic disease may provide a biological explanation for the epidemiological observations of racial health disparity of metastasis.

The probe-summarized intensity signals were log transformed and standardized (mean centered, standard deviation scaled) on an individual array basis and the relative copy number was calculated by subtracting the normal from the tumor intensity for each patient on a probe basis. The resulting copy number profile (CN) represented the amplification and deletion events that accumulated in each cancer sample tested.

Next, the probes were ordered as they appear in the genome and the copy number signal data (CN) was smoothed. The smoothing was conducted using a running median function (runmed in *R*, with endrule parameter equal to “median”). The smoothing function was termed $S(\mathrm{CN})_k$, where k represents the probe width of the smoothing window. The values of k usually range from 5 to 151, depending on the array's probe density and were chosen not to exceed a biologically meaningful span of total genetic distance. Considerations for k should include the average alteration size (estimated empirically from each data set) and distance between probes as determined by the array probe density. As an extreme example, smoothing the entire arm of a chromosome will remove all local variation that exists on that arm. The function $S(\mathrm{CN})_k$ thus yielded n smoothing profiles per sample, with n representing the number of different values used for k. An example of the multiple n values used for chromosome 1 of a particular sample is shown in Figure 3.

2.6. Copy Number Alteration Calling Algorithm

The next part of this stage involved assigning copy number events to each probe. The reason we developed a CNA caller from scratch was because the standard calling algorithms required parameter inputs that were dependent on the signal-to-noise distribution of the copy number measures. Because cancer samples' signal-to-noise are notoriously variable, both on a chromosome basis (within a sample profile) and across samples, this made the standard CNA calling approaches inefficient without significant reconfiguration. Therefore, we developed a method that was dynamic to the signal-to-noise variation observed in cancer genomes. We validate the effectiveness our approach (Figure 4) using a benchmark

simulation data set used to test a variety of algorithms [13]. Given that SNP arrays are not designed to provide quantitative measures of copy number (but do respond linearly to CNAs), we restrict our calls to three categories: amplifications (1), deletions (−1), and neutral events (0). To determine the "center" of the genome so that thresholds can be drawn, we assume that a majority of the intensity values reflect a 2-copy state for the referenced sample, that is, the majority of the referenced tumor sample exists in a 2-copy state (manual calling is used for those samples in which this assumption is not valid). To accomplish this, we sample 10,000 random stretches of probes covering approximately 500 kilobases from the autosomes, calculate the median of each, and use the most frequently occurring value to scale the sample appropriately. Following scaling of the genome, thresholds were drawn based on quantile values and copy number states were assigned to each probe. Since this thresholding scheme was applied to every smoothing, there were n event calls per probe. These calls result in a "ρ" profile, where $T()$ represents the function of trinary binning:

$$\rho_k = T\left(S(CN)_k\right). \quad (2.1)$$

The $n\rho$ calls for each probe were then combined by summation, resulting in a composite profile (ρ') that ranged from $-n$ (signifying that a deletion was called at every smoothing for that probe) to $+n$ (signifying that an amplification was called at every smoothing for that probe):

$$\rho' = \sum_{i=1}^{n} \rho_i. \quad (2.2)$$

One ρ' profile was thus generated per sample, representing a composite of n smoothings, and this metric was used for the rest of the primary analysis. We benchmarked our copy number calling method using a published simulation data set [13] comprised of randomly generated artificial chromosomes. Each chromosome was generated with an aberration flanking the center probe with Gaussian noise $N(0, 0.25^2)$ superimposed. All combinations of signal to noise (SN = 4, 3, 2, and 1) and aberration widths (W = 40, 20, 10, and 5) were produced for a total of 160,000 analysis runs. Receiver-operating characteristics (ROC) were computed from the benchmark simulation dataset [13]; where ROC is defined as a pair, ROC = (TPR, FPR), TPR = (the number of probes within the aberration width that is above a threshold)/(the total number of probes within the aberration width). FPR = (the number of probes outside the aberration width that is above a threshold)/(the total number of probes outside the aberration width . The threshold values are selected to continuously range over the values of the data points, and since ROC is piecewise constant, only changing when a threshold is equal to the value of a data point, we only need to consider values of the data points in their sorted order. The area under the curve (AUC) of each ROC curve was used to gauge performances.

To examine the frequency of amplification and deletions for subgroups of samples or populations and evaluate the sensitivity of our CNA-calling method, we further combined the ρ' data to create ρ'' by summing across the ρ' profiles on a probe basis across multiple samples. Two values of ρ'' were calculated for population or subpopulation. The first value

represented the sum of all positive ρ' values in the population at any probe and was thus called ρ''_{amp}. Likewise, the second value representing the sum of all negative ρ' values in the population at any probe was called ρ''_{del}:

$$\rho''_{\text{amp|del}} = \sum_{i=1}^{n \text{ samples}} \rho'_{[\text{amp|del}]}. \quad (2.3)$$

An example of copy number ρ'' plot (Figure 5) is observed in a select region on chromosome X from metastases of men treated with androgen ablation therapy and primary tumors of iPTs and mPTs from other men not treated. Furthermore, differential analysis of the ρ'' values can be used to identify probes or regions of probes that comprise genes that may contribute to the phenotype being tested (e.g., iPT versus mPT or response to therapy versus no response to therapy).

2.7. Semisupervised Clustering Algorithm

Since sufficient labels were not available to train a model from primary tumors alone, we first created from a cohort of men that developed distant metastases a simplified summary metastasis profile to capture the high-frequency events, that are in part, assumed to correlate to the outcome. This clustering approach is not unsupervised, class-less clustering because we know some information about one of the components which is the summary profile from known metastasis samples. To reflect the frequency of events observed for individual metastasis CNA profiles in the summary metastasis profile, the average number of ρ' events calculated for the group of metastases was used to set a threshold for the number of total ρ' events used to build the summary metastasis profile. The actual probes chosen for the metastasis summary profile were based on their ranked frequency which resulted in a threshold of at least 25% of the samples exhibiting the event. Although not tested here, the theoretical specificity of the summary profile is expected to decrease as the threshold for minimum number of events called decreases, while the sensitivity of the profile decreases as the threshold of minimum number of events called increases. In the case of the MSK cohort, clustering of the 36 metastases ρ' profiles independently yielded two well-separated clusters from which we built two metastasis summary profiles to perform semisupervised clustering with the primary tumors. Alternatively, the 13-patient Hopkins cohort made up of 55 metastases yielded only one homogeneous cluster and associated summary metastasis profile. To overcome the inherent variability with clustering algorithms, we employed a resampling hierarchical clustering method to infer an initial grouping for the unclassified primary tumors. For each iteration, a subset of the individual ρ' profiles from the unknown primary tumors were randomly chosen with replacement and clustered with the summary copy number profile derived from the metastasis samples (one metastasis summary profile from the Hopkins cohort and two from MSK cohort). Therefore, the semisupervised clustering analysis presented here was developed to classify prostate primary tumors into subgroups with different metastatic potential (mPT and iPT) based on their CNA profiles. Distance was calculated using a binary metric, and the samples were joined using hierarchical clustering (complete-linkage method). The cluster tree was divided into two groups at the final join, and the primary tumor samples were scored 1 if they fell in the same cluster as the metastasis profile, and 0 if they were in the other cluster. Using the results

from 20,000 resampling iterations of the clustering, a proximity score was generated for each sample, representing the number of times it fell in a cluster with the metastasis profile. A sample with a high score was considered to be more metastatic (mPT), while lower scoring tumors were more indolent (iPT). The similarity scores distributed throughout the possible range of values (0 to 1), allowing us to form distinct groups of tumors with significant contrast between high- and low-metastatic distance to MSK metastasis signature 1 (Figure 6). The group of samples with scores closer to the center of the distribution were omitted to further define the contrast between high- and low-scoring samples.

2.8. Metastasis Genes Inferred through Evolutionary Selection Modelling

Genomic DNA copy number alterations in local and metastatic prostate tumors are typically numerous, systematic in their genomic placement and varied in size from point mutations to duplications or deletions of entire chromosomes. Given these observations, geneticists have postulated that Darwinian selection may operate on the genomic instability in tumors [14]. High-resolution measurements of CNAs in somatic tumors have informative value, in some cases reflecting the direction in which the biochemistry of the cell controls the quantity of normal, mutant, or hybrid-fusion transcripts and proteins. During this genomic transformation, the resulting modified transcripts and proteins may impact the fitness of the cell. Guided by these principles of evolutionary selection, our analyses sought to identify the CNA landscape that reflects selection mechanisms of metastasis. Genomic selection towards a metastatic cancer phenotype can be both positive and negative and be observed in CNAs exhibiting both amplifications and deletions. For example, genes that promote metastasis and amplified in metastatic tumors would reflect positive selection, while metastasis suppressor genes that are deleted in metastases reflect negative selection. The genes associated with these regions, altered at high frequency in metastatic tumors and enriched in mPTs more so than iPTs, lead to enhanced metastatic potential. We identified specific CNAs that selected positively for metastatic potential, exhibiting amplifications in metastases and mPTs and deletions in iPTs. CNAs identified to exhibit negative selection for metastatic potential were observed to be deleted in metastases and mPTs and amplified in the iPTs. Therefore, we designed models based on Darwin's evolutionary selection theory to score positive and negative selection based on the mPT and iPT classifications derived through semisupervised clustering using the ρ' data. For each probe on the array, we calculated an enrichment score, EN(x), which represented the relative number of amplifications versus deletions, observed in each subgroup (metastasis, mPT and iPT):

$$EN(x) = \frac{(\#\mathrm{Amp} - \#\mathrm{Del})}{\#\mathrm{Samples}}. \quad (2.4)$$

Next, we modeled the relative enrichment by contrasting the metastasis and mPT copy number alterations with those observed in the iPT group:

$$SM = e^{[EN(\mathrm{METS}) + q^* EN(\mathrm{mPT}) - EN(\mathrm{iPT})]}. \quad (2.5)$$

The first two enrichment terms (for metastatic and metastatic-like samples) being summed were designed to assign a higher score when the METS and mPT samples had more amplifications than deletions. Greater amplification enrichment in the METS and mPTs resulted in higher scores. The third term, EN(iPT), was higher when the iPT samples exhibit the opposite effect (enrichment for deletions over amplifications). The middle term, EN(mPT), was multiplied by a data-driven coefficient, *q*, representing the average contribution of mPT on a probe basis (Figure 7).

For example, probes that were amplified in all metastases and mPTs but deleted in all iPTs (positive selection driving the metastasis cells) would yield the highest possible score. Likewise, probes that were deleted in all metastases and mPT samples, but amplified in all iPT samples (negatively select or inhibit the promotion of the metastasis cells), would reach the minimum possible score. Therefore, regions of the genome that enhance and inhibit metastasis formation will be captured by our evolutionary selection model.

Following this probe scoring method we developed a *Z*-score model in order to extend this analysis to the gene level. We assign each probe to a gene, provided it falls within 10,000 bp up- or downstream of the transcription start or stop site. The SM scores for the probes within a gene are averaged and compared to the mean and standard deviation of a background distribution, which was calculated by sampling the top 5th percentile of amplified or deleted probes from all genes on the array with the same number of probes as the gene in question. The result is a *Z*-score for each gene in the genome that is represented on the array.

2.9. Metastatic Potential Score and Survival Analysis

We developed an algorithm based on genomic CNAs to calculate a metastatic potential score (MPS), with a higher score indicating a greater likelihood of metastasis. The MPS score for a new individual patient only depends on the CNA profile of this new patient. It can be calculated without requirement for other samples, since it's simply based on the concordance/discordance relationship to the CNA metastasis gene signature previously identified as selecting for the metastatic phenotype through our selection model. The MPS was calculated using a weighted *Z* score from the top set of CNAs overlapping metastasis genes determined by the significance of their selection model *Z* scores. We used $Z \geq 1.7$ as a cutoff point because for standard normal distribution, the tail of 1.7 is about 5%. The metastatic potential score was defined as the following:

$$\mathrm{MPS}=\sum_{i=1}^{n} Z_i' * \mathrm{Dir}_{\mathrm{sig}}(i) * \mathrm{Dir}_{\mathrm{samp}}(i). \quad (2.6)$$

For each tumor profile, logistic adjusted *Z* scores (Z') from genes ($i \ldots n$) that match the direction of the metastasis gene signature (a vector of −1s and 1s representing whether the gene was deleted or amplified in the signature, resp.) were added, whereas Z' from genes that mismatch the direction of the signature were subtracted. As the direction component of the risk model score (Dir) reflects, if the CNAs of the metastasis signature ($\mathrm{Dir}_{\mathrm{sig}}$) and the unknown sample profile ($\mathrm{Dir}_{\mathrm{samp}}$) are in the same direction, the coefficient will be 1; if they are in opposing directions, the coefficient will be −1; and if $\mathrm{Dir}_{\mathrm{samp}}(i) = 0$, then the entire

term will not count towards the score. For example, if a gene *i*, that is typically amplified in metastases (Dir_{sig} (*i*)) and mPTs, is also amplified in the unknown profile (Dir_{samp} (*i*)) that *Z* score is added, whereas if gene i in the profile is deleted, as expected in iPTs, the *Z* score is subtracted. Neutral genes that are neither amplified nor deleted in the unknown profile are not scored in this model.

Three metastasis signatures, derived from a combination of five cohorts were used to develop the MPS. The first signature was identified using 49 primary tumors of unknown clinical outcome from NYU (*n* = 29) and Baylor (*n* = 20) and a metastasis cohort from Hopkins (*n* = 13). The other two signatures were identified using 75% of the MSK cohort of primary tumors of unknown outcome (*n* = 126) along with a set of metastatic tumors (*n* = 36) from the same MSK cohort. The CNA-based gene signatures from these 2 sets of cohorts were concatenated and derived into the MPS which we assessed in a Cox proportional hazard model with samples set aside for testing purposes only. The test cases were comprised of bona fide mPTs (primary tumors that later developed into distant metastasis), whereas the test controls were derived from a random sample of tumors with unknown outcome not used to build the MPS. All presurgery predictors (PSA, clinical stage, biopsy Gleason) and other demographic variables (age at diagnosis and race) were tested independently and in combination with the MPS in Cox proportional hazards survival analysis with the time variable represented by progression to metastasis.

3. Results

3.1. Prediction Models

Our selection models resulted in three hundred and sixty-eight genes (from 3 metastasis signatures) with a CNA status that was concordant among METS and mPTs and contrasted with iPTs ($Z \geq 1.7$) (Supplemental Table 1, see Supplementary Materials available online at doi:10.1155/2012/873570). With these genes, we developed the MPS and tested the accuracy as an independent predictor of metastasis, with a subset of primary tumors (*n* = 52) not used to develop the signatures (*n* = 13 mPTs and *n* = 39 control primary tumors, Table 1). As a continuous predictor, applying the MPS to a Cox proportional hazards model resulted in a significant association to the endpoint of metastasis-free survival (2.88; 95% CI = 1.15 – 7.2; *P* = 0.02) (Table 2).

Patients diagnosed with prostate cancer have several pretreatment variables, such as, clinical stage (combination of digital rectum exam, PSA, and ultrasound/MRI), biopsy Gleason score and other demographic measures (e.g., age or race) to guide the decision to undergo surgery. These variables have marginal clinical utility and, in our cohorts, none of these clinical variables were statistically significant in univariate or multivariate logistic regression models. In multivariate Cox regression models (Table 2), only the MPS score reached statistical significance, indicating, that the MPS score was the only reproducible predictor of metastasis-free survival.

Notably, the clinical stage was specific when palpable tumor was detected (T2 or greater); however, it lacked sensitivity, because 47% (9/19) of pathological stage-4 cases that evaluated *ex-vivo* were diagnosed as T1C before surgery [7]. Twenty-seven percent (13 out

of 49) of clinical stage T1C tumors that were upstaged following prostatectomy resulted in distant metastasis formation. Therefore, staging at the time of biopsy can seriously underestimate the severity of disease. Similarly, the biopsy Gleason score versus the postsurgery Gleason score was underestimated in 38% of cases and overestimated in 8% [7] (Figure 8).

3.2. Metastatic Potential Score Distributions

Significant differences as measured by Mann-Whitney test of the MPS were observed for the metastasis ($P < 0.001$) and mPT ($P = 0.001$) groups, compared to the control primary tumors (Figure 9). The MPS in the lymph-node-positive primary tumors (derived from the MSK ($n = 9$) and Stanford ($n = 9$) cohorts did not differ significantly from the control tumor group ($P_{\text{MSK}} = 0.34$, $P_{\text{Stanford}} = 0.13$, $P_{\text{Combined}} = 0.08$), which reflected the marginal ability of this clinical parameter to predict distant metastasis in previous reports [15].

Consistent with our assumption that the control cohorts contained a fraction of mPTs, their MPS overlapped the MPS range of the cases. Furthermore, control primary tumors (from MSK cohort) that did not recur biochemically (as measured by PSA) after 80 months of followup, (represented by green Xs in Figure 9) were not significantly correlated with the MPS. To determine whether other cancer types exhibited a similar metastatic landscape of CNAs to that observed in prostate cancer, we calculated the metastatic potential score for 337 cancer cell lines. We observed an overall distribution that overlapped with low-risk prostate primary tumors (Figure 9). However, 22 of the 337 cell lines ranked by MPS were above the 75th percentile of the prostate primary tumors and metastases. These cell lines originated from tumors of the lung ($n = 10$), breast ($n = 3$), colon ($n = 2$), and melanoma ($n = 2$). Other singletons in this group of 22 cell lines originated from thyroid, rectum, pharynx, pancreas, and kidney.

3.3. Biomarker Functional Significance

Another way to validate our algorithms is by data mining the functional attributes of the metastasis genes identified by the selection model. As expected, many of the top-ranking metastasis genes identified have molecular functions related to alteration of nuclear and extracellular matrix structure and metabolic modification that enhance processes characteristic of escape from anoikis (a key metastasis specific process). A heat map of the CNA events of signature genes for all prostate tumors is suggestive of a path toward the different high frequency amplification versus deletion events that contrast the high-risk and low-risk tumors (Figure 10). The mid-risk region with its relative paucity of signature events may represent the starting point of two alternative pathways of subsequent copy number alteration, one leading to metastasis and the other to an indolent state. The locking in of these “antimetastasis” events in indolent tumors may explain why they failed to metastasize despite extended periods of watchful waiting.

One of the top predictor genes, the solute carrier family SLC7A5 gene, deleted on chromosome 16q24.2, encodes a neutral aminoacid transporter protein (LAT1) that has been implicated in multiple cancers (prostate [16], breast [17], ovarian [18], lung [19], and brain [20]) and has been shown to have utility as a diagnostic [21–23] and drug target in cell line

[24–26] and preclinical animal models [27]. The normal function of LAT1 is to regulate cellular aminoacid concentration, L-glutamine (efflux) and L-leucine (influx). Reduced activity of LAT1 results in increased concentrations of L-glutamine which has been shown to constitutively fuel mTOR activity [28]. Seven other solute carrier superfamily members (SLCO5A1, SLC7A2, SLC10A5, SLC26A7, SLC25A37, SLC38A8, and SLC39A14) were predictive of metastatic potential in our models, likely creating a cellular environment conducive to metastasis.

A second subset of signature genes included 6 Cadherin family members encoding calcium dependent cell adhesion glycoproteins (CDH2, CDH8, CDH13, CDH15, CDH17, and PCDH9). Many of the Cadherin family proteins have putative functions associated with metastasis progression [29] and have been included in diagnostic panels [30, 31].

A third subset of 5 genes predicted to contribute to metastatic potential were potassium channels, KCNB2, KCNQ3, KCNAB1, KCTD8, and KCNH4. Notably, 3 other potassium channels reside in the highly amplified region between 8q13 and 8q24 (KCNS2, KCNV1 and KCNK9) that did not rank high in our analysis but may have weak or modifier effects. High levels of cytoplasmic potassium ion concentrations have been shown to inhibit the hallmark mitochondrial apoptotic cascade of membrane disruption and ensuing release of cytochrome C, caspase, and nuclease degradation of cellular components [32]. Furthermore, another study showed that the methylation status of potassium channel, KCNMA1 (10q22.3), was predictive of prostate cancer recurrence [33]. The activity of voltage-gated potassium channels in prostate cancer cell lines, LNCaP (low metastatic potential) and PC3 (high metastatic potential), were observed to be markedly different [34]. The complete set of metastasis signature genes likely represents various subsets of functions. Representation of different gene family members suggests that each tumor may have a unique profile to progress to metastasis, yet different members of a gene family may contribute to a functional redundancy. Notably, the genomic DNA landscape around the androgen receptor locus on chromosome X represents a compelling observation linking CNAs to a functional cause and effect response of androgen ablation therapy (Figure 5).

4. Summary

In this study, we developed a semisupervised clustering algorithm that can infer the classification of a primary tumor based on metastatic risk. This was essential to overcome the limitations inherent to prostate cancer cohorts for collecting long-term clinical outcome data. Our novel approach to modeling the CNA data based on Darwin's evolutionary selection theory allowed us to identify genes associated with the specific metastatic processes of anoikis. Current clinical models for assessing risk are aimed at predicting biochemical recurrence, rather than metastasis, and do not include genomic information. This limitation was underscored in a study with a large cohort of greater than 10,000 men who had undergone radical prostatectomy [35]. Within that cohort, about 20% of men developed biochemical recurrence within 5 years of the procedure, but subsequently only 10% of the men with biochemical recurrence developed distant metastases after 12 years.

This proposed new classification method and selection model allowed us to develop a metastatic potential score that could be used for predicting an individual's metastasis-free survival at the time of diagnosis. With validation in additional cohorts and statistical models with known metastasis outcome, this approach may lead to a significant advancement in determining whether aggressive treatment of prostate cancer is necessary. This predictor might be important for correctly categorizing men at the time of diagnosis and could predict whether surgery, radiation therapy, or watchful waiting was warranted. Because the proposed tool, tumor genomic analysis, is comprehensive for identifying the genetic changes that are associated with the pathogenesis of metastasis, there is a greater likelihood of selecting a sufficient number of markers that are both sensitive and specific predictors. This method could be applied to other cancers (e.g., breast) that exhibit variation in the metastatic potential of the primary tumor and have similar difficulties in collecting tumor samples with long-term clinical outcome data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. Kelly Maxwell for extracting the genomic DNA for the NYU cohort and all of the reviewers for their thoughtful suggestions.

References

[1]. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA—A Cancer Journal for Clinicians. 59(4):225–249. [PubMed: 19474385]

[2]. Klotz L, Zhang L, Lam A, Nam R, Mamedov A, Loblaw A. Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. Journal of Clinical Oncology. 2010; 28(1):126–131. [PubMed: 19917860]

[3]. Ohori M, Kattan M, Scardino PT, Wheeler TM. Radical prostatectomy for carcinoma of the prostate. Modern Pathology. 2004; 17(3):349–359. [PubMed: 14765206]

[4]. Beroukhim R, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905. [PubMed: 20164920]

[5]. Sun J, Liu W, Adams TS, et al. DNA copy number alterations in prostate cancers: a combined analysis of published CGH studies. Prostate. 2007; 67(7):692–700. [PubMed: 17342750]

[6]. Castro P, Creighton CJ, Ozen M, Bcrel D, Mims MP, Ittmann M. Genomic profiling of prostate cancers from African American men. Neoplasia. 2009; 11(3):305–312. [PubMed: 19242612]

[7]. Taylor BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18:11–22. [PubMed: 20579941]

[8]. Lapointe J, Li C, Giacomini CP, et al. Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Research. 2007; 67(18):8504–8510. [PubMed: 17875689]

[9]. Liu W, Laitinen S, Khan S, et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nature Medicine. 2009; 15(5):559–565.

[10]. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics. 2008; 40(10): 1253–1260. [PubMed: 18776909]

[11]. Team, RDC. R Foundation for Statistical Computing. Vienna, Austria: 2009.

[12]. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics. 2006; 22(1):7–12. [PubMed: 16267090]

[13]. Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics. 2005; 21(19):3763–3770. [PubMed: 16081473]

[14]. Cahill DP, Kinzler KW, Vogelstein B, Lengauer C. Genetic instability and darwinian selection in tumours. Trends in Cell Biology. 1999; 9(12):M57–M60. [PubMed: 10611684]

[15]. Boorjian SA, Thompson RH, Siddiqui S, et al. Long-term outcome after radical prostatectomy for patients with lymph node positive prostate cancer in the prostate specific antigen era. Journal of Urology. 2007; 178(3):864–871. [PubMed: 17631342]

[16]. Sakata T, Ferdous G, Tsuruta T, et al. L-type amino-acid transporter 1 as a novel biomarker for high-grade malignancy in prostate cancer. Pathology International. 2009; 59(1):7–18. [PubMed: 19121087]

[17]. Kaira K, Oriuchi N, Imai H, et al. L-type amino acid transporter 1 and CD98 expression in primary and metastatic sites of human neoplasms. Cancer Science. 2008; 99(12):2380–2386. [PubMed: 19018776]

[18]. Kaji M, Kabir-Salmani M, Anzai N, et al. Properties of L-type amino acid transporter 1 in epidermal ovarian cancer. International Journal of Gynecological Cancer. 2010; 20(3):329–336. [PubMed: 20375792]

[19]. Imai H, Kaira K, Oriuchi N, et al. L-type amino acid transporter 1 expression is a prognostic marker in patients with surgically resected stage I non-small cell lung cancer. Histopathology. 2009; 54(7):804–813. [PubMed: 19635099]

[20]. Kobayashi K, Ohnishi A, Promsuk J, et al. Enhanced tumor growth elicited by L-type amino acid transporter 1 in human malignant glioma cells. Neurosurgery. 2008; 62(2):493–503. [PubMed: 18382329]

[21]. Bartlett JMS, Thomas J, Ross DT, et al. Mammostrat as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy. Breast Cancer Research. 2010; 12(4) article no. R47.

[22]. Ring BZ, Seitz RS, Beck RA, et al. A novel five-antibody immunohistochemical test for subclassification of lung carcinoma. Modern Pathology. 2009; 22(8):1032–1043. [PubMed: 19430419]

[23]. Ring BZ, Seitz RS, Beck R, et al. Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer. Journal of Clinical Oncology. 2006; 24(19):3039–3047. [PubMed: 16809728]

[24]. Fan X, Ross DD, Arakawa H, Ganapathy V, Tamai I, Nakanishi T. Impact of system L amino acid transporter 1 LAT1 on proliferation of human ovarian cancer cells: a possible target for combination therapy with anti-proliferative aminopeptidase inhibitors. Biochemical Pharmacology. 2010; 80(6):811–818. [PubMed: 20510678]

[25]. Yamauchi K, Sakurai H, Kimura T, et al. System L amino acid transporter inhibitor enhances anti-tumor activity of cisplatin in a head and neck squamous cell carcinoma cell line. Cancer Letters. 2009; 276(1):95–101. [PubMed: 19058911]

[26]. Kim CS, Cho SH, Chun HS, et al. BCH, an inhibitor of system L amino acid transporters, induces apoptosis in cancer cells. Biological and Pharmaceutical Bulletin. 2008; 31(6):1096–1100. [PubMed: 18520037]

[27]. Oda K, Hosoda N, Endo H, et al. L-Type amino acid transporter 1 inhibitors inhibit tumor cell growth. Cancer Science. 2010; 101(1):173–179. [PubMed: 19900191]

[28]. Nicklin P, Bergman P, Zhang B, et al. Bidirectional transport of amino acids regulates mTOR and autophagy. Cell. 2009; 136(3):521–534. [PubMed: 19203585]

[29]. Yilmaz M, Christofori G. Mechanisms of motility in metastasizing cells. Molecular Cancer Research. 2010; 8(5):629–642. [PubMed: 20460404]

[30]. Celebiler Cavusoglu A, Kilic Y, Saydam S, et al. Predicting invasive phenotype with CDH1, CDH13, CD44, and TIMP3 gene expression in primary breast cancer. Cancer Science. 2009; 100(12):2341–2345. [PubMed: 19799609]

[31]. Lu Y, Lemon W, Liu P-Y, et al. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PLoS Medicine. 2006; 3(12):2229–2243. article e467.

[32]. Ekhterae D, Platoshyn O, Krick S, Yu Y, McDaniel SS, Yuan JXJ. Bcl–2 decreases voltage-gated K^+ channel activity and enhances survival in vascular smooth muscle cells. American Journal of Physiology, Cell Physiology. 2001; 281(1):C157–C165. [PubMed: 11401838]

[33]. Vanaja DK, Ehrich M, Van Den Boom D, et al. Hypermethylation of genes for diagnosis and risk stratification of prostate cancer. Cancer Investigation. 2009; 27(5):549–560. [PubMed: 19229700]

[34]. Laniado ME, Fraser SP, Djamgoz MBA. Voltage-gated K^+ channel activity in human prostate cancer cell lines of markedly different metastatic potential: distinguishing characteristics of PC-3 and LNCaP cells. Prostate. 2001; 46(4):262–274. [PubMed: 11241548]

[35]. Nakagawa T, Kollmeyer TM, Morlan BW, et al. A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. PLoS One. 2008; 3(5) Article ID e2318.

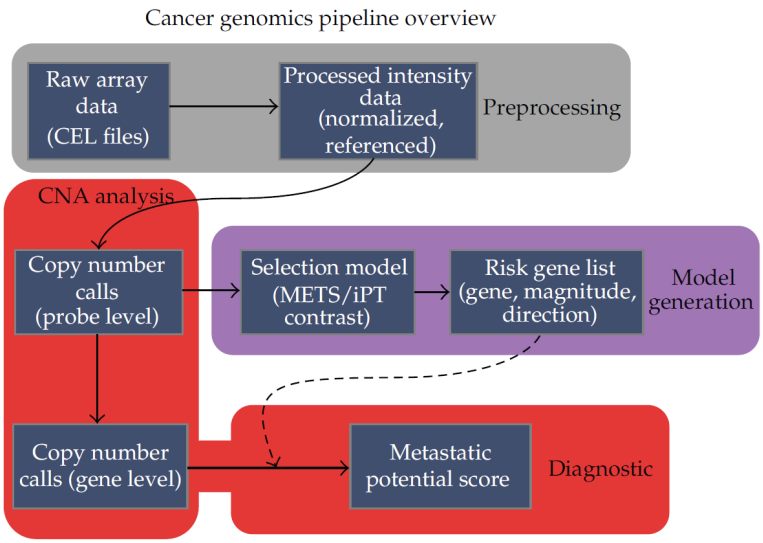

Figure 1.

Array CGH analysis pipeline for processing pixel image data from Affymetrix SNP arrays to produce genotype and signal intensity measures for copy number analysis used for developing bioclinical models and diagnostics.

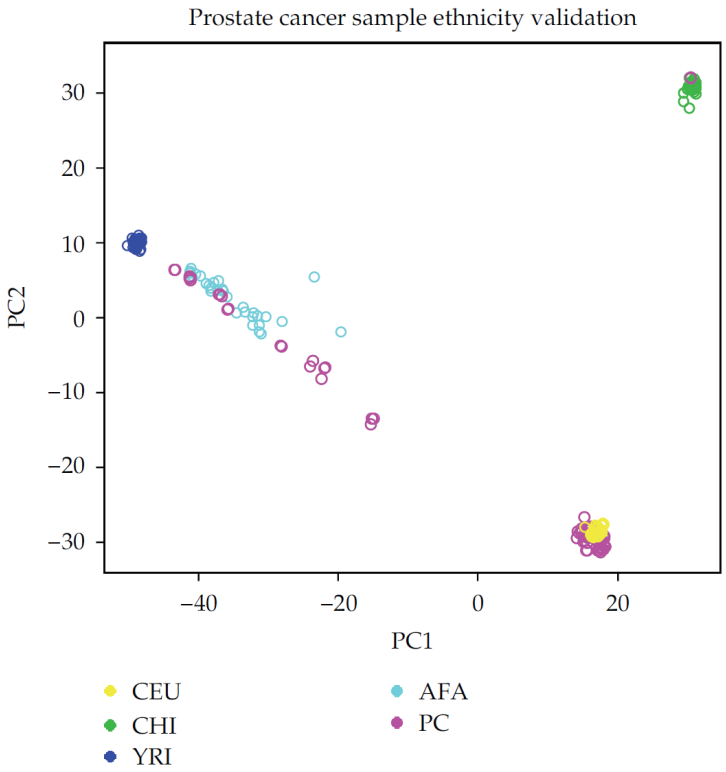

Figure 2.

Principal component analysis identity testing of a variety of normal SNP profiles from the germline DNA of prostate cancer patients (PC) used in the study compared to a set of HapMap normal reference populations from Nigeria (YRI), Europe (CEU), China (CHB), or, of African American (AFA) decent. The *x* and *y* axes represent the 1st and 2nd eigenvectors.

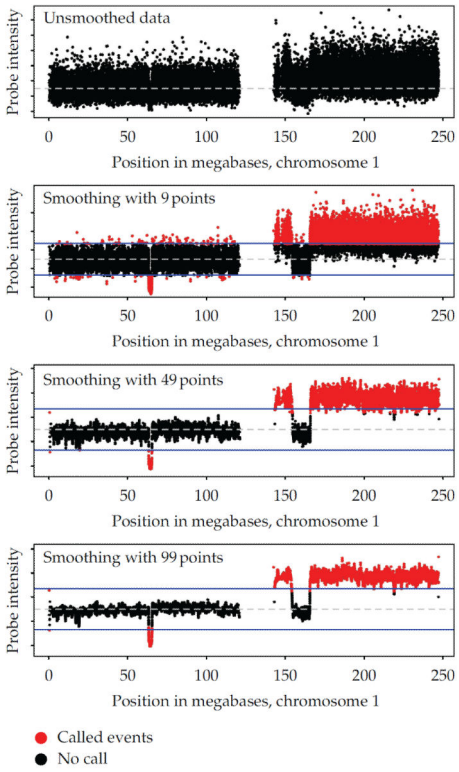

Figure 3.

A representative primary tumor chromosome 1 copy number profile (top panel) and corresponding $S(\mathrm{CN})n$ [k {9, 49, 99}] in the bottom panels. Therefore, $n = 3$ because three different smoothing lengths are used. Black probes represent probes that are not called while red probes are the called events that exceed the amplification and deletion thresholds.

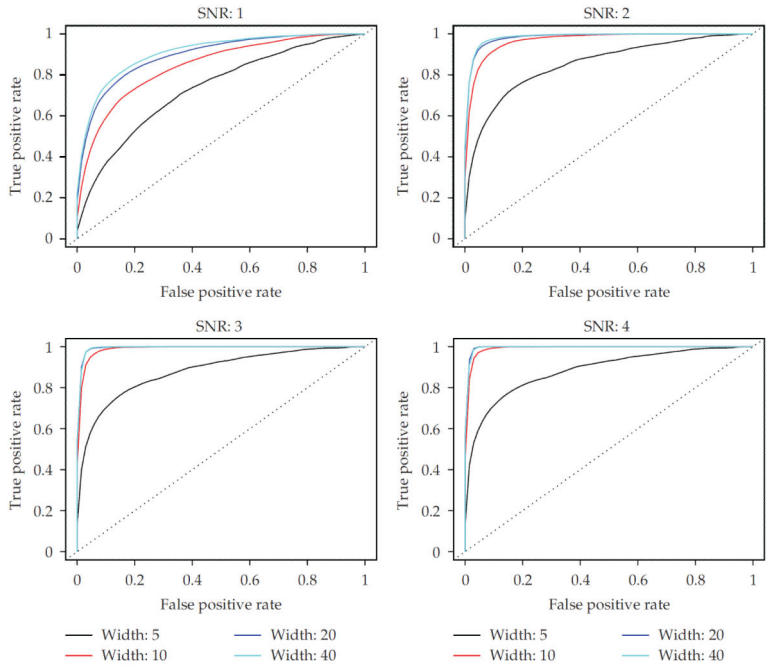

Figure 4.

Receiver-operating characteristic curves showing the performance of our CNA-calling algorithm on the simulated data [13]. Each panel represents a different signal-to-noise ratio and the curves represent varying event widths of the simulated data. The *x*-axis represents the false positive rate, and the *y*-axis represents the true positive rate. Each curve is generated by testing varying thresholds on 100 simulated chromosomes for the condition specified. The curves are combined using vertical averaging. The dashed line represents the random model.

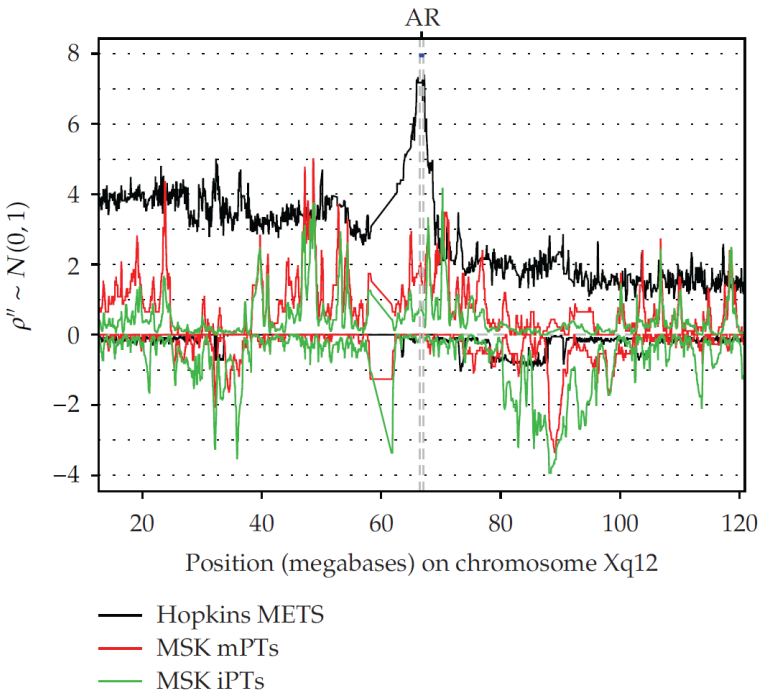

Figure 5.

Copy number profile, ρ'' shows an amplification of a region on chromosome X harboring the androgen receptor (AR) locus. The *x*-axis represents the ordered chromosome position and the *y*-axis represents standardized population frequencies exhibiting amplifications (above 0) and deletions (below 0). The three populations of tumors are represented as red, black, or green lines for mPTs, androgen ablation treated metastases (METS), and iPTs, respectively.

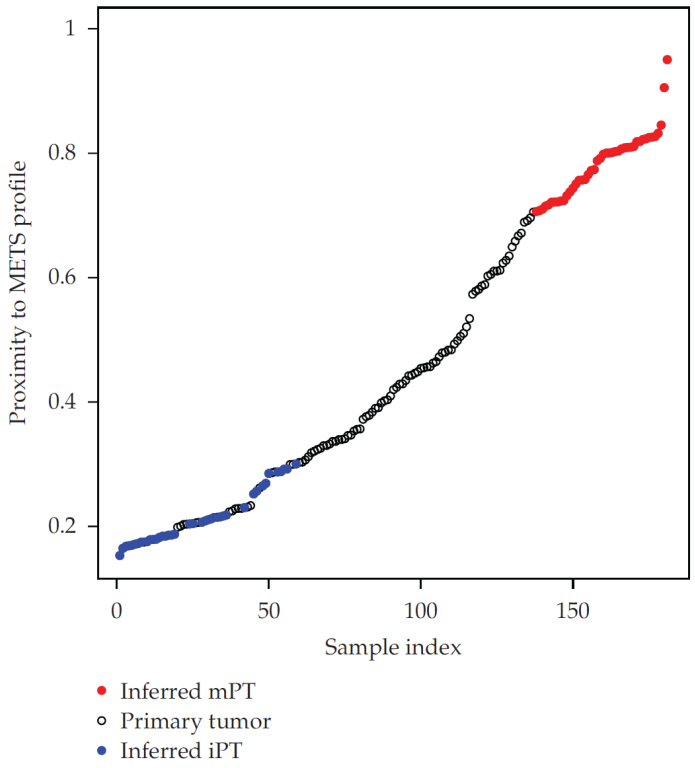

Figure 6.

Plot of ranked proximity score for MSK signature 2. Proximity represents the number of times a particular sample clustered with the MSK metastasis profile 2. The samples with higher scores (red points) are classified as inferred mPTs and the samples with lower scores (blue points) are classified as inferred iPTs. Primary tumors (hollow points) interspersed with the blue iPT tumors were excluded as iPTs for MSK signature 2 because they did not consistently classify as iPTs in the proximity analysis using MSK signature 1.

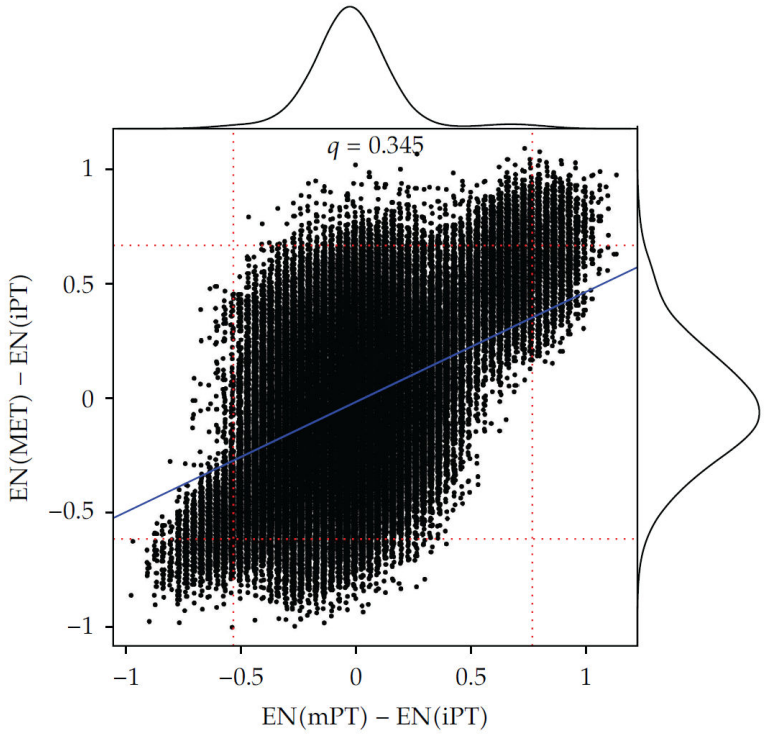

Figure 7.

Scatterplot of the enrichment scores of the METS versus those of mPTs, normalized by the enrichment scores of iPTs. Kernel density estimation curves are shown protruding from the *x* and *y* axes. The horizontal and vertical dashed red lines denote the trim points (quantiles 0.99 and 0.01). A linear regression line, based on the trimmed values, is shown in blue. The value of *q* is the Pearson correlation coefficient for the trimmed *x* and *y* values.

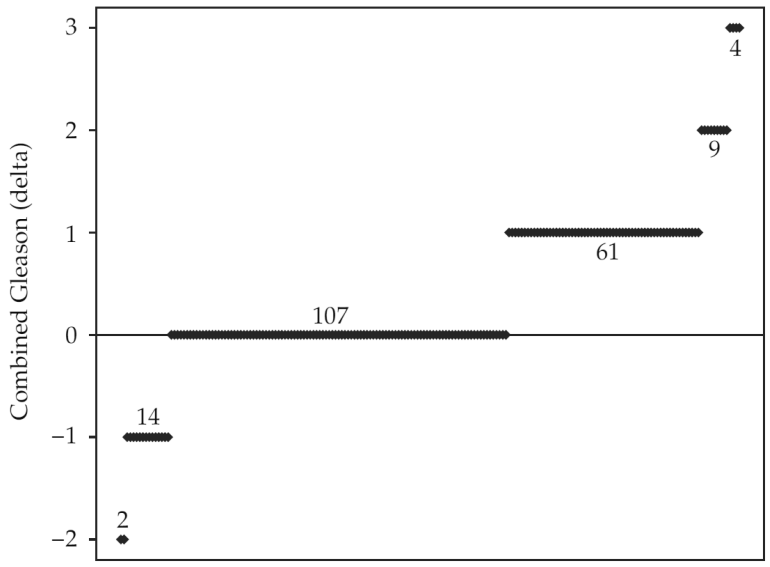

Figure 8.

Biopsy versus pathology Gleason score. The difference between the Gleason score as measured from a biopsy of the tumor relative to the pathological assessment of the score using the radical prostatectomy surgical specimen (*y*-axis). The *x*-axis represents the sample index.

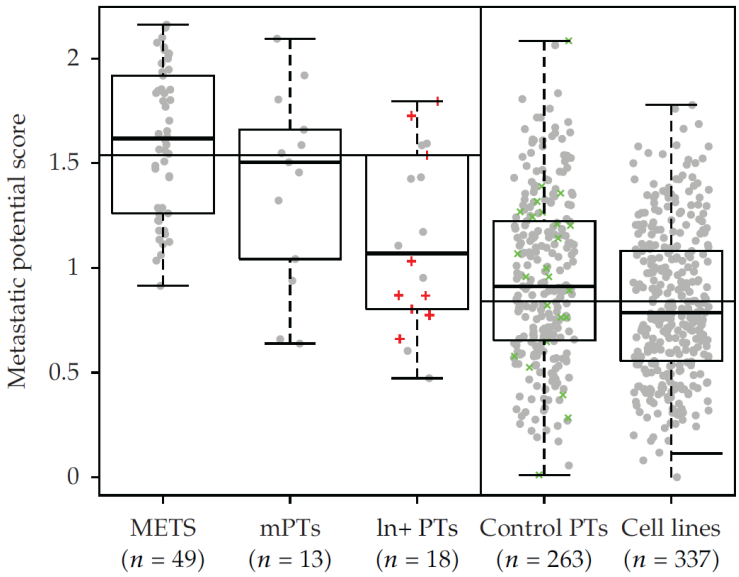

Figure 9.

Boxplot showing the metastatic potential scores for all samples involved in the analysis. All high-risk tumors are shown in the left three boxes (metastases, progressors, and lymph-node-positive samples), while unknown control primary tumors and the publically available cell line data are shown in the right boxes. The red "+" symbols in the lymph-node-positive box represent those samples from the MSK dataset, distinguishing them from the SU cohort lymph-node-positive samples. The green "x" symbols in the control primary tumors plot represent selected low-risk primary tumors (individuals with no biochemical recurrence (PSA) for at least 80 months).

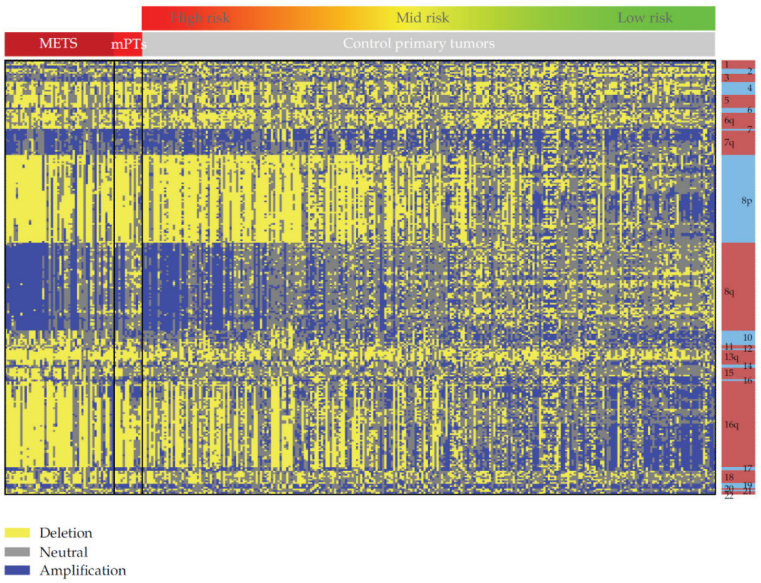

Figure 10.

Heatmap showing copy number amplifications and deletions for tumor samples in the gene signature. The genes are arranged in genomic order; position is indicated by the colored bar on the right. The tumor samples (*x*-axis) are arranged by subtype (metastatic (METS), progressors (mPTs), and control primary tumors) and further sorted by their metastatic potential score. A strong pattern emerges in the metastasis samples on the left and is shared by the progressors and high-risk primary tumors. Further towards the right, the metastatic pattern diminishes and even shows a reversal in copy number pattern in some chromosomal areas.

Table 1

Clinical And histological characteristics of samples used to validate the metastatic potential score model.

	Case	Control
n	13	39
Age		
Mean	59.5	59.1
Median	61	58
Standard deviation	7.1	7.3
Range	46–67	46–73
Race		
Asian	0 (0%)	1 (1.9%)
Black	1 (1.9%)	4 (7.7%)
Unknown	0 (0%)	2 (3.8%)
White Non-Hispanic	12 (23.1%)	32 (61.5%)
Clinical stage		
T1C	4 (7.7%)	23 (44.2%)
T2	5 (9.6%)	16 (30.8%)
T3	4 (7.7%)	0 (0%)
T4	0 (0%)	0 (0%)
Biopsy Gleason score		
5	0 (0%)	0 (0%)
6	4 (7.7%)	26 (50%)
7	7 (13.5%)	10 (19.2%)
8	2 (3.8%)	2 (3.8%)
9	0 (0%)	1 (1.9%)
Prediagnosis biopsy PSA (ng/mL)		
Median	6.9	5.6
<4	2 (3.8%)	6 (11.5%)
4–10	6 (11.5%)	24 (46.2%)
>10	4 (7.7%)	7 (13.5%)
Pretreatment PSA (ng/mL)		
Median	12.8	5.6
<4	2 (3.8%)	7 (13.5%)
4–10	4 (7.7%)	26 (50%)
>10	7 (13.5%)	6 (11.5%)

Table 2

Cox proportional hazards model analysis of the metastatic potential score and clinical predictors.

Component	Hazard ratio	*P*	95% CI
Univariate			
MPS	2.87	0.02	1.2–7.2
Pretreatment PSA	1.00	0.04	1.0–1.1
Clinical stage T2-T3	1.27	0.70	0.4–4.2
Multivariate			
MPS	2.61	0.05	1.0–6.8
Clinical stage T2-T3	0.90	0.87	0.3–3.1
Pretreatment PSA	1.00	0.18	1.0–1.0