Prostate Cancer Health Disparity in African American and Caucasian American Men Characterized through the Landscape of Genomic Instability

By

Alexander Pearlman

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Program in Computational Biology

New York University

September, 2009

_____

Harry Ostrer, M.D.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Dedicated to those afflicted by cancer and their loved ones.

Abstract

In 2008, prostate cancer was the most prevalent cancer in men in the United States with 186,320 estimated new cases and 28,660 deaths. When detected early, prostate cancer can be curable, but procedures that offer the best prognosis such as radical prostatectomy - the complete removal of the prostate gland - often result in severe side. Clinical measures of metastatic potential of a localized prostate tumor often result in overly aggressive treatment. The goals of this research were to: 1) provide clinicians diagnostic markers with strong predictive power, 2) to reveal the mechanisms behind metastatic potential, and 3) identify targets for preventative and disease treatment. An integrated analysis combining data from genomic DNA copy number, gene expression and genome wide association was performed. We utilized the principles of evolutionary selection to build a model comparing the primary cancers of African American men and Caucasian American men with those of metastases which allowed us to delineate genes that select for metastatic potential versus those that oppose it. Results from this study suggest that a racial disparity exists, reflected in the somatic tumor genomes of African American and Caucasian American men. This genomic racial disparity involves putative prostate cancer candidate genes such as PTEN/PREX2a, AR and ERCC1 along with several novel candidates such as VASP and NME4. A comprehensive analysis in the context of putative protein interactions and gene sets revealed an enrichment for gene ontologies involved in cell adhesion, proliferation and cytoskeleton formation. Overall, the results imply that racial disparity for metastatic disease is driven by

a variety of genes, each with variable selective influence over the function of three interconnected pathways controlling cellular structure and growth.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FILES

Supplement 1: Complete table of gene set enrichment results. (available online)

Supplement 2: Complete table of copy number, gene expression and GWAS results. (available online)

Chapter 1

1.1 Prostate cancer epidemiology

Prostate cancer develops as a result of a complex series of cellular, physiological, behavioral, environmental and socio-economic inputs. These inputs interact with germline genetic and somatic risk factors to produce various clinical outcomes. New cases for 2008 were estimated at 186,320 (25% of all male cancers) and new deaths were estimated at 28,660 (10% of all male cancer deaths). The probability of developing prostate cancer increases with age — 1 in 39 (2.5%) men will develop the disease by age 59, while 1 in 6 men (16%) will develop invasive prostate cancer by age 85.[1]

Studies of twins that typically have been used to determine the relative contributions of genetic and environmental influences support a major role of germline genetic susceptibility. The National Academy of Sciences Twin Cohort, comprised of 38,848 male veteran twins of whom 1,009 had prostate cancer, has revealed a monozygotic concordance rate for prostate cancer of 27.1%, compared with a rate of 7.1% for dizygotic twins. This provides strong evidence of the influence of genetic susceptibility to prostate cancer. The heritability of disease risk was estimated to be 57%.[2] Similar results were observed in a study of the twin registries of Denmark, Sweden and Finland, where the proportion of variance attributed to heritable factors was calculated to be 42%.[3]

Family history is a major risk factor in the development of prostate cancer, and men with prostate cancer have a positive family history in 5-15% of cases.[4-6] A meta analysis of 32 population-based studies demonstrated that family members of a prostate cancer patient experienced a 2.46-fold (95%:

2.14-2.82) increase in risk of prostate cancer.[7] The risk of disease is increased over 4-fold if 2 or more first-degree relatives are affected. The nature of this familial clustering is such that the risk ratio rises with the following: decreasing age at diagnosis of the patient and his family members; increased genetic relatedness of the affected relative(s); and, as noted, increased number of individuals affected within the family. This familial risk of prostate cancer is independent of ethnicity and has been observed in varied populations.

Early investigations of the genetic etiology of prostate cancer through familial clustering estimated an autosomal-dominant mode of inheritance.[4] To date, however, no gene has been identified that conforms to this model. Inheritance of mutations in susceptibility genes may be a major determinant of prostate cancer risk and outcome. Identification of prostate cancer-related markers or genes may lead to pre-symptomatic germline genetic testing for risk modification.

The biology of cancer susceptibility, tumor suppressor genes and loss of heterozygosity was popularized in the early 1970s by Alfred Knudson through studying the records of retinoblastoma patients.[8] Retinoblastoma is a malignant tumor believed to derive from the photoreceptors of the retina. Knudson's theory was based on the observation that individuals who developed bilateral or multiple tumors had a mean age of onset two-fold earlier than those who presented with unilateral or single tumors. Additionally, only the bilateral cases had affected family members. Knudson reasoned that an early germline mutation, obtained through hereditary transmission, would result in an increased probability of obtaining early and multi-focal tumors as a consequence of secondary somatic mutations later in life. The Knudson "two-hit" model has

withstood the test of time and is currently used to drive clinical treatment plans. Individuals without family history and with a unilateral presentation are often treated immediately with enucleation of the affected eye, whereas infants or toddlers presenting multiple bilateral tumors in one eye are treated conservatively with local chemotherapy, laser therapy, freezing or radiation discs in anticipation of new tumors forming in the second eye.

Differences in the incidence and natural history of prostate cancer among ethnic groups may also be indicative of variable germline genetic risks. A well-known health disparity exists between the prostate cancer cases of Caucasian and African-American men. Men of African descent have an increased incidence, earlier onset and more aggressive form of disease than men of predominantly European origin. Even when adjusted for the increased level of incidence in African Americans, mortality rates are still higher. Incidence and death rates among African American men are more than twice those of Caucasian men.[1] Although attempts to reconcile whether socioeconomic variables contribute to this effect have resulted in some indicators implicating life-style, education, and insurance status with mortality, incidence did not show any clear correlation with socioeconomic status.[1,9,10] Blood steroid levels in African or Caucasian men, such as testosterone and estradiol, have shown no significant association with prostate cancer risk; however, the ratio of testosterone to estradiol showed significant associations with cancer risk in Caucasian men.[11-13]

1.2 Prostate gland and prostate tumor biology

The prostate gland is an exocrine gland which produces fluid that forms part of the semen, which is stored with sperm in the seminal vesicles. Ejaculation causes muscular contractions that secrete the semen through the urethra, where it is expelled from the body through the penis. In addition to the prostate's role in producing ejaculate, it also contributes to controlling the flow of urine. The prostate wraps around the urethra as it passes from the bladder to the penis. Prostate tissue is made up of basal and luminal epithelial cells along with variety of stromal cells – neuroendocrine, vascular and hematological. Enlargement of the prostate can lead to painful urinary retention requiring medical care.

Examples of non-cancerous growth are benign prostate hyperplasia and prostatitis. These are thought to be caused, respectively, by increased sensitivity to testosterone and viral or bacterial infection. If, however, a physician notices several risk factors, described later, he or she will may suspect cancerous growth and encourage a biopsy. This will then be inspected by a pathologist.

Transformation of epithelial cells occurs within the lumen after the emergence of the prostate intraepithelial neoplasia cell (PIN), a precursor cancer cell that typically grows for ten years before it is observed to be an invasive carcinoma (figure 1b).[14] The assignment of cancer precursor to PIN lesions is a result of their proximity to the peripheral zone shared by invasive carcinoma and the early observations of multi-focal allelic imbalance, thought to be a hallmark of carcinomas.[15] Multi-focality was rescinded in a recent high-resolution allelic

imbalance analysis of a series of multiple prostate cancer metastases dissected during autopsy, arguing for a monoclonal origin.[16]

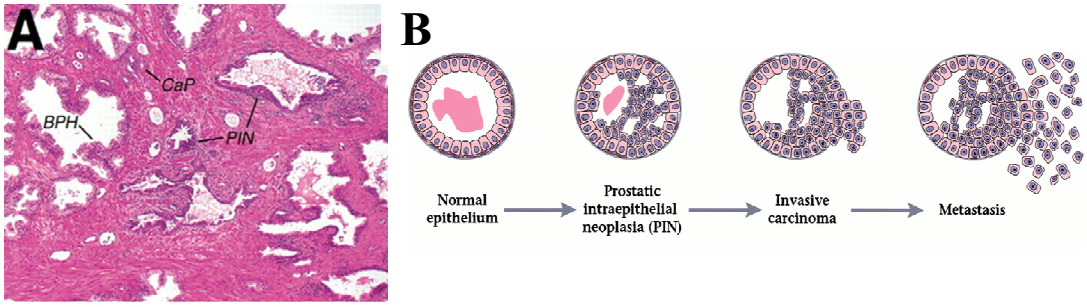

Figure 1: Prostate tissue histology and tumor genesis. (Adapted with permission from Shen *et al.* [15]) A) Prostate tissue illustrating normal, benign prostate hyperplasia (BPH), prostatic intraepithelial neoplasia, primary carcinoma (CaP). B) Transitions from normal epithelium to metastasis.

1.3 Prostate tumor molecular biology

Along with germline mutations, somatic genomic alterations contribute to the development of prostate cancer. These alterations include allelic loss, gene amplification and fusion/rearrangements. A comprehensive review of 41 studies comprised of 872 advanced (n=255) and primary (n=659) tumors has implicated several genomic loci and corresponding candidate genes as part of recurrently imbalanced regions of the genome. High frequency amplified chromosomal regions include 8p21.3, 13q21.3 and 16q22.1. Deleted regions occur at 8q22.2, 17q25.2 and 7q21.11 .[17] These loci average 39 million bases (MB) and are present in patients at frequencies ranging from 6.66% to 34.09%. They encode a total of 4,859 proteins, averaging 250 per locus. Genes that are identified within the smallest region of overlapping deletions may represent

tumor suppressor genes (TSGs).[18,19] These genes may become targets of germline testing for prostate cancer susceptibility and potential therapeutic targets, and they may result in functional deregulation reflected at the transcript or messenger RNA levels. The most notable of these relationships is observed with the TMRPSS2-ERG fusion/rearrangement located on chromosome 21q22l.3.[20-22]

1.4 Diagnostics

Early stage diagnosis is instrumental in the effective treatment of prostate cancer. Current diagnostic methods assaying blood or tissue biopsy screening are controversial and of low information value, especially when predicting metastatic disease after a diagnosis of local carcinoma. Aside from a non-specific and highly insensitive physical examination of the affected area (digital rectum exam [DRE]), three techniques are commonly used to test for the presence of carcinoma: prostate specific antigen testing, Gleason scoring based on prostate biopsy and prostate cancer staging.

Prostate Specific Antigen, as determined through a PSA test, is a neutral serine protease produced by columnar epithelial cells lining the ducts of the prostate and periurethral glands.[23,24] PSA leaks from the prostatic ducts into the blood stream, where it is easily detected via immunoassay.[25] PSA levels increase with age and have been shown to be higher in African-American men than in Caucasian Americans.[26] PSA screening, however, has a high false positive *and* false negative rate and as single marker was recently shown to have no significant impact on mortality over non-screening.[27] Other tests, such

as PSA-velocity, which measures the rate of change of serum levels in men with relatively low PSA levels (2.0-4.0 ng/ml),[28] have gained some popularity, although a recent review also argues against the clinical utility of PSA-velocity.[29]

The second commonly used technique is prostate biopsy. Prostate biopsies are performed transrectally by urologists, with up to 12 independent specimens being obtained. Should a tumorous lesion be seen, it is scored using the Gleason score grading system. This is based on the histologic pattern of arrangement of carcinoma cells in hematoxylin- and eosin- (H&E) stained sections. These stains target, respectively, nucleic acid and extracellular proteins in tissue slices obtained from biopsy or prostatectomy. Five basic grade patterns (Figure 2) are used to generate a histologic score, which can range from 2 to 10.[30] This is a highly subjective system that as a single predictor has poor specificity and sensitivity for predicting risk of metastatic disease.[31]

Data from prostate cancer biopsies and other biological sources are assessed by the tumor staging system developed by the American Joint Committee on Cancer (AJCC), the third common technique. This classifies the primary tumor into a range of categories reflecting low level tumor cell detection (T1a) through carcinomas that invade nearby tissues (T4). T4, however, does not necessarily represent metastatic growth.

Once diagnosis of localized prostate cancer is made, predictors of mortality are critical for clinical decision making. Strategies such as Cancer of the Prostate Risk Assessment (CAPRA) score have purported to predict metastasis and mortality with good accuracy. The CAPRA score utilizes age at diagnosis, PSA, Gleason score, percentage of positive biopsy cores, AJCC stage and comorbidity.[32] In a recent retrospective study of 10,000 men undergoing

various forms of treatment for local prostate cancer, a CAPRA score (range 0-10) of 0 through 2 resulted in 10 year survival of 97.1%, whereas a score of 6 through 10 resulted in a 79.1%.[33] Although seemingly impressive, this result indicates that around 80% of individuals in even the most severe CAPRA subgroup did not present with metastatic disease. Other shortcomings of this analysis include: 1) A 10-year survival may be a premature time frame, representing a non-informative period for patients with a CAPRA score from 0 to 2, since early-stage carcinoma typically progresses slowly and takes over a decade to materialize into metastatic disease;[14] 2) Greater than 50% of the individuals sampled had radical prostatectomies, and positive outcomes from these patients are not necessarily informative if metastases had not yet formed; and 3) clinical regimens, such as radiation or androgen deprivation, were variable over the length of the study.

Overall, CAPRA and its component measures may have significant clinical utility for diagnosing carcinoma; however, the score lacks the ability to gauge metastatic potential, which is critical for clinical decision-making about the level of treatment.

Lastly, a potential non-invasive diagnostic that has not yet been used in the mainstream medical practice is PCA3, an androgen-responsive gene specific to prostate cancer cells. This gene is measured in urine or blood. This diagnostic has been shown, in combination with PSA, to slightly improve the predictive potential of diagnosing primary carcinoma.[34-36]

Although, early diagnosis coupled with radical prostatectomy is the currently accepted failsafe scenario for preserving longevity, this procedure can diminish a patient's quality of life. All of the diagnostic methods in this section

have been discussed in terms of relative power to diagnose *primary* carcinoma. As we have shown, none of these, however, is a tool for predicting metastatic potential of a primary prostate cancer. Considering a model in which primary cancer and metastasis are related but separate diseases, separate diagnostics must be developed to determine the probability for metastatic disease to occur.

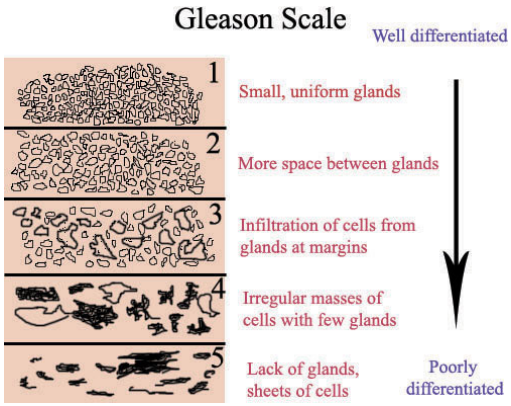

Figure 2: Gleason score of various histological grades (1-5) of prostate tumor, observed under a light microscope and summarizing a diagnosis by adding scores from two regions of the same tumor nodule.

1.5 Treatment of primary prostate cancer and metastatic disease

Once a positive diagnosis is made for localized prostate cancer, a range of treatment options are available based on the individual physician's guidelines. Treatments may include watchful waiting, primary androgen deprivation, external beam radiotherapy, brachytherapy (radiation pellets), cryotherapy or radical prostatectomy. Advanced stage metastatic disease first treated with androgen ablation therapy sends patients into remission for an average of twenty months, after which hormone refractory disease emerges where androgen-independent clones develop. At this late stage, chemotherapy and radiation are applied to help manage pain typically associated with bone metastasis.[37] A single clinical study has shown a reversion to androgen sensitivity when chemotherapy (chlorambucil and lomustine) was introduced to

patients with hormone refractory disease.[38] Furthermore, recent *in vitro* studies reveal synergistic of growth when inhibition is applied to the Hedgehog and ErbB signaling pathways in circulating tumor cells of hormone refractory patients and the androgen insensitive cell line LNCaP. The pathways were inhibited by cyclopamine, gefitinib and lapatinib.[39] This study was undertaken to identify other targets for synergistic therapies and molecular predictors of metastatic potential in primary tumors.

1.6 Data used in this study

Three main data types were used in this study: genomic copy number, gene expression and genome wide association study data. The human genome sequence is a structurally dynamic, redundant and selectively evolving system. As the structural state of the individual genome varies, so does the phenotypic state of the cell as it develops and interacts with the environment. A range of sequence modifications from a change in a single nucleotide to the duplication or deletion of entire chromosomes may result in traits that are benign, advantageous, or detrimental to the vitality of the organism. Some sequence transformations are random, whereas others are deterministic; these operate at a range of frequencies. Advances in technologies that measure sequence copy number on a genomic scale provide a high resolution and global perspective of the location, prevalence and systematic nature of these variations.[40-42]

Given that prostate cancer has a clear genetic component[7] and has historically been characterized as a disease of genomic instability,[17] the screening of both normal and tumor prostate DNA and RNA will enable the

identification of somatic cancer-causing events that accumulate through the various stages of carcinogenesis and metastasis. The research presented in this dissertation utilized data measuring tumor DNA copy number, tumor gene expression and germline genome wide association to elucidate the markers and their associated genes that lead to the inheritance and progression of prostate cancer (Figure 3).

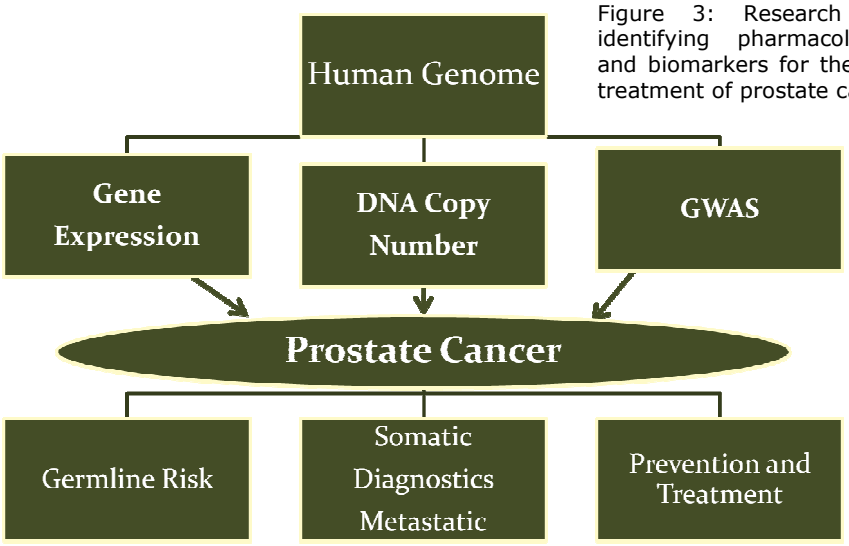

Figure 3: Research strategy for identifying pharmacological targets and biomarkers for the diagnosis and treatment of prostate cancer.

Copy number variations (CNV), recurrent amplifications and deletions, are prevalent in the normal human genome. Array comparative genomic hybridization (aCGH) is a common method for measuring DNA copy number. The Database of Genomic Variants,[40] a catalogue of copy number variations, currently cites 31 publications representing 6551 unique CNVs, spanning 850 million bases (MB) or ~28% of the complete genome sequence. These regions of variation, a proportion of which are ethnic-specific, have not yet been associated with risk of disease; however, because a significant proportion of

CNVs occur in known functional areas of the genome, it is plausible that an individual's background CNV profile may confer risk for or protection from disease. Recurrent sequence deletions and amplifications were shown to be enriched within regions of segmental duplication.[40,41,43-47] These have been implicated in promoting a variety of inherited and sporadic genomic disorders[48,49] through mechanisms such as non-allelic homologous recombination mediated through Alu transposition[46,47,50-54], palindromic AT-rich repeats[55-58] and other physiochemical properties of the DNA helix.[47] Non-homologous end-joining represents another mechanism by which these events can happen. Genomic copy number screens of various somatic and germline conditions, including cancers – prostate cancer,[59-67] breast cancer,[68-76] lung cancer,[77-80] melanoma,[81-84] liver cancer,[85-89] pancreatic cancer,[90-94] colon cancer,[88,95-104] brain cancers,[105-110] and blood cancers,[111-114] – and cognitive conditions – including mental retardation,[44,115-118] autism,[119-121] and schizophrenia[122,123] – have identified disrupted loci associated with the phenotype.

Tumor copy number analysis of prostate cancer with array CGH maps the duplicated or deleted chromosomal segments as registered through high-density arrays of oligonucleotide probes. This has been used to identify loss of heterozygosity (LOH) regions in prostate cancer.[34,64,124,125] Early application of array CGH to prostate cancer cell lines identified two novel regions of homozygous chromosomal loss at 17q21.31 and 10q23.1 in the PC3 cell line that may represent regions of tumor suppressor genes.[34] Another report using an early generation oligonucleotide probe platform with 100k SNPs differentiating probes compared 22 matched normals and tumors (Gleason 6-9).[59] A clear relationship between Gleason score and total number of copy number events

was demonstrated. For example, the median numbers of deletions were 17.5, 50, and 205 for the tumors of Gleason 6, 7, and 8, respectively. The median numbers of gains were 1.5, 35 and 94. Notably, they confirmed loci harboring the gene PTEN in 45% of samples and fusion products ERG and TMPRSS2 in 30% of the samples.[126] The 5' UTR of ERG and TMPRSS2 have also been shown to be fused and implicated in 49% of 118 prostate cancers and 49% of lymph node metastases,[127] causing the over expression of ERG transcript. These observations have been confirmed with novel deep sequencing technology, revealing the precise nature of the sequence abnormality.[20]

A study of 64 men at intermediate to high risk of recurrence following radical prostatectomy included 32 men who progressed biochemically, who were compared to 32 who did not.[64] Deletion of 8p23 was more common in progressors (50% vs. 31%) and gain of 11q13.1 was predictive of recurrence independent of stage and progression.

To examine the racial disparity in prostate cancer, somatic copy number alterations that contribute to the development and progression of prostate cancer were analyzed. As part of the efforts to identify somatic alterations that confer an ethnically-based enrichment for aggressive disease and to aid the identification of metastasis genes, we conducted an array CGH study using Affymetrix SNP Array 6.0 on paired normal and tumor tissue from 9 African American (AA) and 20 European American (EA) men. For added statistical power, 2 recently published datasets comprised of paired normal and tumor tissue from 20 AA[128] patients and multiple metastases (n=52) from 12 individuals[16] were included.

Whole-genome gene expression analysis has been given significant attention in the past decade.[129-134] Since its creation in 2004, Oncomine,[135] the leading academic repository for cancer gene expression, has accumulated data from 392 studies, including 18,000 genome profiles from 41 different cancer types. Twenty-two prostate cancer-related studies comprised of 747 individuals currently reside in their database. Because of the complexity of gene-expression data, very few meaningful discoveries have been made using this data-type alone.

Genome wide association studies (GWAS) use common variants to discover biomarkers that can be used to estimate the genetic component of disease risk and prognosis.[136] Ultimately, these biomarkers can be used to determine mechanistic and molecular details underlying a disease state and then design interventions.[137] The major limitation of one-marker-at-a-time analysis, when performed on typical GWAS datasets, is the limited sensitivity and specificity of low penetrance markers to account for genetic variance and guide clinical decision making for individual patients.[136] For example, three recent studies comprised of 63,000 cases and controls that attempted to associate height with single nucleotide polymorphisms (SNPs) resulted in a series of markers each explaining between 0.3% and 0.5% of the variance of the phenotypic data.[138-140] Detecting these very small effects at 80% power would require ~10,000 cases and controls.[141] For complex phenotypes involving cellular processes with many genes, an even larger number of cases and controls is needed, and population sizes are almost always much smaller than technically required. Sensitivity issues for detecting SNP associations are exacerbated in

traits such as prostate cancer that are genetically heterogeneous and have a relatively smaller heritable genetic variance (~57%) than that of height (~80%), as measured through studies of twin registries.[7,142] Under these estimates, GWAS of such complexity would require cohorts approaching 100,000 samples to achieve the power to identify significant associations of single markers with both strong and weak effects.[143]

Of the significant prostate cancer discoveries, SNPs on chromosome 8q24 and 17q12 have been replicated in several populations.[144-147] These SNPs are in proximity to the MYC regulatory sequence and are within the HNF1B gene, respectively, suggesting potential functional roles. High-density genotyping of 8q24 in five multi-ethnic populations revealed seven independent risk loci conferring a joint population attributable risk [PAR-multifactorial inherited component of a disease; R = K – Y/K, where K is the observed disease incidence and Y is the disease incidence in the absence of the genetic variant[136]] of 68% in African Americans and 32% in European Americans, respectively, figures consistent with their relative incidences of disease.[148] Targeting these SNPs along with 16 SNPs in the 17q24.3 region in a Swedish cohort of 2893 cases and 1781 controls, Zheng *et al* estimated a PAR of 46.34%. This was calculated from the cumulative PARs of the top 5 scoring SNPs (PAR = 40.45%) and family history (PAR = 9.89%).

These results, however, have poor clinical utility, because only 1.4% of the cases exhibited 5 or more SNPs conveying a risk that might influence clinical decision-making (OR = 9.46). Another 8.2% of cases had 4 SNPs, yielding an OR = 4.76, which is comparable to that of elevated prostate specific antigen

(PSA).[149] Other risk SNPs located around the KLK genes showed significant association,[150] but that may have been the result of bias in population sampling based on PSA levels.[151] These loci and their regional genes represent a major advancement to the understanding of prostate cancer genetic risk; however, a broader repertoire of genes and the understanding of their functional pathways will be necessary for making accurate predictions about the mechanistic relationship between genetic risk factors and cancer.

1.7 Thesis Statement

Prostate cancer has a genetic etiology of medical significance. Genomic regions involved in tumor genesis go through an accelerated evolution by recurrent rearrangements observed as copy number amplifications and deletions. These events reflect the genome's compensatory response to the stresses of a limiting local environment or exogenous insults incurred during treatment. African American and Caucasian American men exhibit a health disparity that leads African Americans to an increased incidence and mortality of disease. Utilization of the principles of evolutionary selection to build a model comparing the primary cancers of African American men and Caucasian American men with those of metastasis allows for the delineation of genes that select for metastatic potential versus those that oppose it. These inferences could be validated through data mining of putative networks and pathways of interaction and used to test complex hypotheses *in silico*, in the laboratory and ultimately in the clinic.

Chapter 2

2.1 Populations and samples studied

The rationale presented in this thesis for a study of somatic copy number was the clear observation that African American (AA) men have an earlier incidence and more aggressive (> 2-fold mortality) form of prostate cancer than do Caucasian American (CA) men.[1,9,10] Prostate cancer tumors typically exhibit genomic instability, characterized by increased rates of recurrent amplifications and deletions compared to normal genomes.[17] Therefore, it is reasonable to postulate that some of the events that are preferentially enriched in one of the two populations. Moreover, these events will harbor candidate genes that predispose that population to either an earlier onset and greater metastatic potential (in the case of the AA population) or to a less aggressive form of the disease and lower metastatic potential (in the case of the CA population). Ideally, this analysis would make use of data from primary prostate biopsies of patients with positive and negative metastatic outcomes who did not undergo radical prostatectomies; however, such samples are not readily available. Therefore, through the National Cancer Institute sponsored Cooperative Prostate Cancer Tissue Resource (CPCTR), 9 AA and 21 CA prostate tumors and matched normals dissected during radical prostatectomy from nearby prostate tissue were obtained for the analyses presented here. Additionally, 2 public datasets were used. The first was comprised of 20 AA normal-tumor pairs that also underwent radical prostatectomies.[128] The second contained 52 metastasis and paired normal tissue collected at autopsy from 12 CA individuals, each with 3 to 6

metastases.[16] In total, 61 individuals and 101 tumor copy number profiles were considered in this analysis (Table 1).

Table 1: Clinical parameters for prostate cancer individuals assayed by arrayCGH

Dataset	Number of Patients	Number of Tumors	Stage			Gleason score			Age at prostatectamy			Sample Type	Array Type
			T2	T3	*T4	<7	7-9	Met	<60	59><7	>6		
Primary African American/Caucasian	29	29	17	11	1	13	16	0	9	14	4	9 African American primary cancer, 20 Caucasian primary cancer	Affymetrix SNP 6.0
Primary African American (public data)	20	20	9	11	0	9	11	0	0*	5	4	20 African American primary cancer	Affymetrix 500k
Metastasis (public data)	12	52	0	0	55	0	0	55	na	na	na	52 metastasis samples**	Affymetrix SNP 6.0

*ages for only 9 of 20 individuals available for cohort **2 samples were from primary metastasis

Somatic prostate tumor copy number alterations have a functional influence over the cells by deregulating messenger RNA (mRNA) levels of genes through mutations and structural reorganization of the genome. These events may offer a selective advantage over the host environment or, conversely, prevent the cancer cell from forming metastasis.

Combining gene expression data with genomic copy number data helps to discern the genes that are altered as a function of the genomic instability in prostate cancer. Two gene expression prostate cancer datasets were collected for analysis. The first contained 50 primary tumors with matched normal prostate tissue profiles.[152] A second expression dataset (unpublished) with 19 AA and 14 CA age- and stage-matched primary somatic tumors was also analyzed (Table 2).

Table 2 Prostate cancer studies of gene expression				
Study	**tumors**	**norms**	**matched**	**chip type**
Sighn	50	50	50	U95A
Osman	19AA, 16CA	NA	NA	U133
AA: African American, CA: Caucasion American				

The third data type considered in this investigation was summarized single marker SNP association data from 1,172 non-Hispanic Caucasian prostate cancer cases and 1,157 ethnically matched controls. These were provided by the National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMS) genome wide SNP association studies (GWAS).[153] The data were integrated with the somatic copy number and gene expression data types to help delineate functional networks and pathways that predispose individuals to prostate cancer and increase their risk for aggressive disease.[146]

2.2 Prostate tissue sample processing for array CGH

Prostate cancer normal and tumor tissues were obtained from the CPCTR, and DNA was extracted using a Gentra DNA extraction kit. Purified genomic DNA (gDNA) was hydrated in reduced EDTA TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0). The DNA concentration was measured using the NanoDrop™ 2000 spectrophotometer at Optical Density (OD) wavelength of 260nm. Protein and organic contamination were measured at OD 280nm and 230nm, respectively. Samples that passed these quality control thresholds were then run on a 1% agarose gel to assess the intactness of the genomic DNA (Figure 4). 500ng of gDNA samples were run on the Affymetrix Human SNP Array 6.0 at the Rockefeller University Genomics Resource Center (New York, NY 10021) using standard operating procedures.

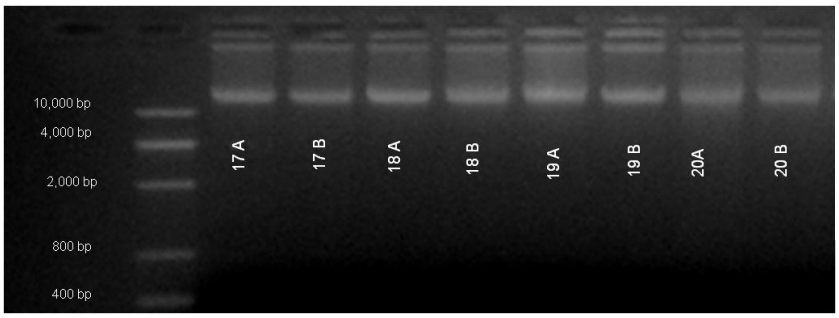

Figure 4: Purified prostate cancer gDNA run on 1% agarose gel. DNA ladder shown in first the first lane ranging from 400-10,000 basepairs (bp).

Table 3: Affymetrix Human SNP Array Quality Control

Dataset	Number of arrays	%Call rate			AA%Heterozygosity			CA%Heterozygosity			Genotyping algorithm	Array Type
		Norm	Primary	Met	Norm	Primary	Met	Norm	Primary	Met		
Primary African American/Caucasian	58	98.9	98.7	na	28.6	28.6	na	26.0	25.8	na	Birdseed v2	Affymetrix 6.0
Primary African American (public data)	40	97.7	97.2	na	29.1	28.6	na	na	na	na	BRLMM	Affymetrix 500k
Metastasis (public data)	68	98.0	97.5	96.7	27.6	na	20.9	25.2	22.3	22.0	Birdseed v2	Affymetrix 6.0

The average sample genotype call rate, as estimated by the birdseed algorithm[154] implemented through Affymetrix Power Tools-1.10.2, was 99%. The average heterozygosity rates for CAs and AAs were 26% and 29%, respectively. The public datasets were processed similarly and resulted in comparable call and heterozygosity rates (Table 3), with the exception of the dramatic reduction in heterozygosity rates for metastasis samples relative to their matched normals.

Population identification was then performed to confirm the racial identifications associated with either AA or CA samples. To accomplish this, a principal component analysis was run using the normal genotype profiles of each sample relative to a reference set of Nigerian (YRI), European (CEU) and Chinese (CHB) profiles (Figure 5) obtained from the International HapMap Project.[155] Additionally, except for one mis-annotated CHB individual, which was not included in further analysis, all normal and tumor profiles of patients were validated to be part of their self-identified racial assignment.

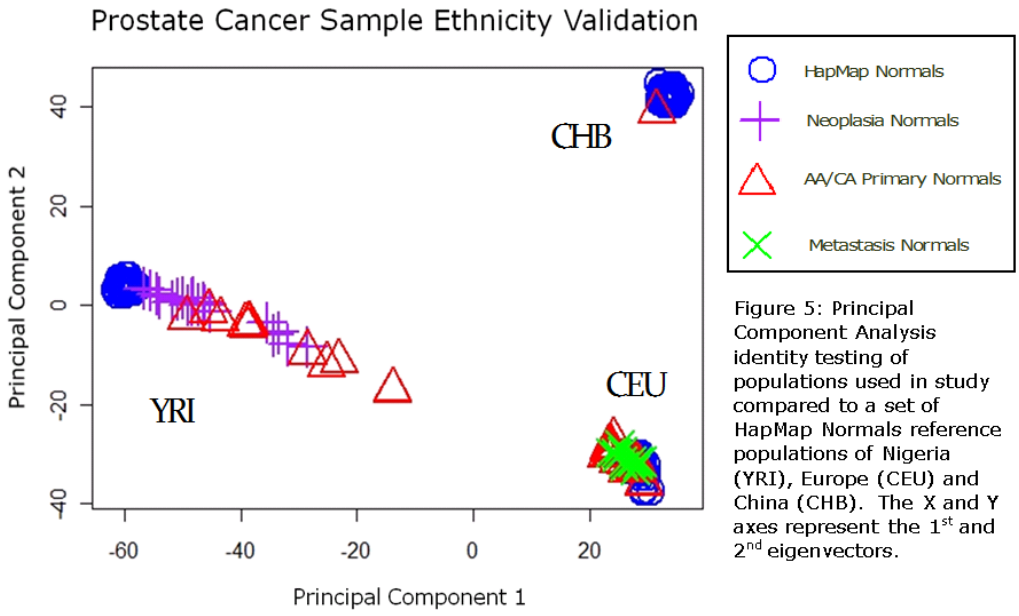

Figure 5: Principal Component Analysis identity testing of populations used in study compared to a set of HapMap Normals reference populations of Nigeria (YRI), Europe (CEU) and China (CHB). The X and Y axes represent the 1st and 2nd eigenvectors.

2.3 Significance testing procedures

Copy number, gene expression and GWAS data types provide a numerically optimal set of solutions for each respective dataset. Cancer biology and technical (assay-related) confounders, however, produce a large number of significant results that are numerically indistinguishable from the true positives. By combining the significant loci and associated genes from three orthogonal data types, the true positive signal may surface above the noise. To accomplish this, a Z-score-based procedure was used at multiple steps in the analysis – copy number, gene expression, GWAS – and at the integration of these steps to approximate the significance of genes in conforming to models about metastatic prostate cancer. The general formula for a Z-score is as follows:

$$Z = \frac{X - \mu}{\sigma}$$

where X is the observed measure and μ and σ are the mean and standard deviation of the population. In all of the following Z-score-based analyses, a

random sampling of the population was performed to estimate the parameters of the reference distribution supporting the model used to interpret the biological system under investigation.

At certain steps in the analysis, sets of ranks from the same data-type were combined. Once gene rankings were determined for a particular analysis, ranks *G* positions across *k* analyses were evaluated using a non-parametric ranking method:[156]

$$R(G) = \sum_{i=1}^{k} \log\left(\frac{1}{G_k}\right)$$

This method was selected as an improvement to a simple average of the ranks of each G across the *k* analyses because it gives more emphasis to having a high rank in any one of the analyses, regardless of rank in the others. This model of rank integration gives more weight, for example, to a gene ranked #1 and #1000 in two different analyses than to a gene ranked #500 in each.

To test for enrichment of a top-ranking group of genes in known pathways, gene-set enrichment was performed. A subset from the collection of 5451 genesets from the BROAD Institute's Molecular Signatures Database[157] (MSigDB) were used to test for enrichment against a group of prostate cancer related genes. This database consists of putative curated pathways (n=1892), Gene Ontology (GO) classified groups (n=1454), motif sets made up of transcription factor binding sites (n=500) and miRNA targets (n=222). Overlap of the observed set of genes with each of the MSigDB gene sets was evaluated using the hypergeometric equation:

$$p(x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

where n represents the total number of unique genes in the database, m represents the possible matches with all the genes in the database (n), x represents the number of genes in the observed set overlapping with a particular gene set, and k is the size of that particular gene set. P-values were obtained for each gene set representing the enrichment of the observed list of candidate genes; however, the distribution of these p-values showed a bias (Figure 6) in which the size of the gene set was inversely proportional to the p-value.

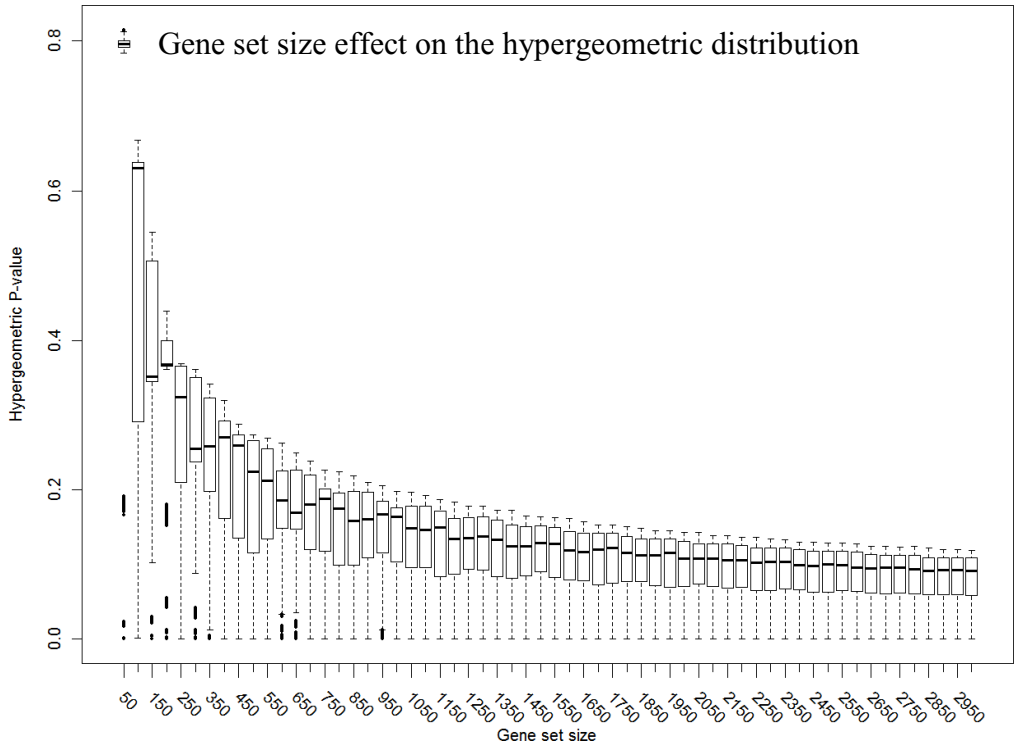

Figure 6: Simulation of hypergeometric p-value distributions by fixing the size of the input list of genes and windowing the size of the gene sets to reflect the gene set size distributions of the MSigDB database.

To correct for this, a Z-score was calculated for each geneses using a background of 5,000 randomly selected sets of genes, matching the sizes of the real gene sets. The gene sets were ranked by Z-score.

2.4 Copy number analysis

A copy number analysis pipeline was designed using the R-statistical software[158] (R) to process the data through a series of computational steps (Figure 7) resulting in ranked lists of genes and associated significance.

In **stage 1**, signal intensity files (.cel) for the Affymetrix SNP Array 6.0 or 500k mapping arrays were processed using the Affymetrix Power Tools, Birdseed V2[154] and BRLMM[159] algorithms, respectively, resulting in genotype allele calls and copy number signal intensity measures for each SNP and copy number probe. After the first stage, the genotype calls were prepared for downstream analyses such as PCA for identity and quality control testing.

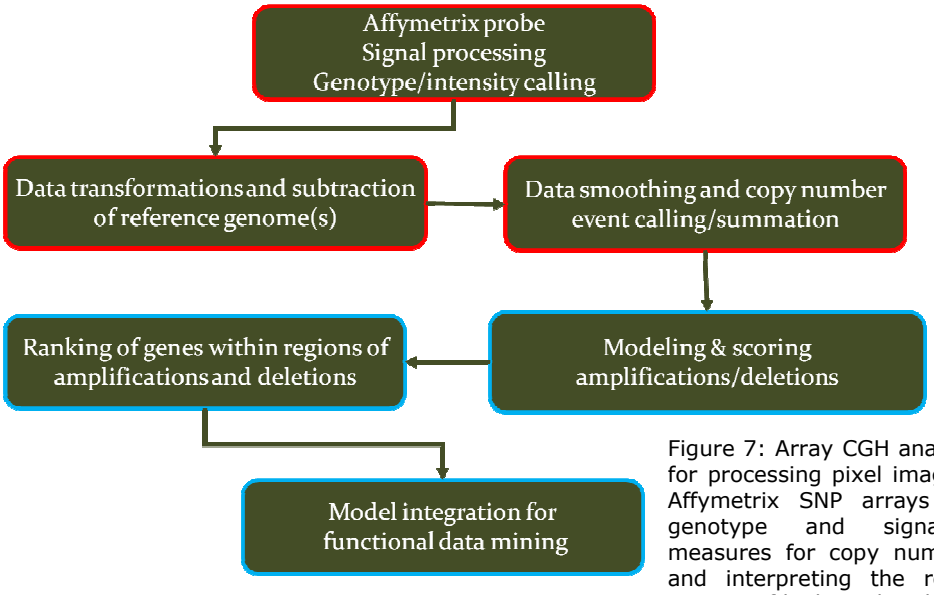

Figure 7: Array CGH analysis pipeline for processing pixel image data from Affymetrix SNP arrays to produce genotype and signal intensity measures for copy number analysis and interpreting the result in the context of biological pathways.

In **stage 2** of the pipeline, the probe-summarized intensity signals (I_k, where k represents the probe) were log-transformed and standardized (mean centered, standard deviation scaled) on an individual array basis. Then, 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.

At the beginning of **stage 3**, the probes were ordered as they appear in the genome. Then, 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 $S(CN)_k$ (where k represents a user-defined number of adjacent probes over which the running median was performed). The function $S(CN)_k$ 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 8.

The next part of this stage involved calling copy number events, either amplifications or deletions. A probe was called an event if its relative copy number after the application of the function *S(CN)* exceeded a fixed threshold of $\pm$1.7 standard deviation units (sdu). Based on a scheme of amplification, deletion or no event, a trinary event call was generated, with the value "1" or "-1" being assigned to any probe whose S(CN) value exceeded the amplification (amp) or deletion (del) threshold. A "0" was assigned to any probe with no event, or a "neutral" (neu) probe. Since an event call was applied to every

smoothing, there were k event calls per sample. These binned calls were summarized in a "ρ" profile, where *T()* represents the function of trinary binning:

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

Each individual's set of n ρ event call indices was then summarized on a probe basis by summation, resulting in a 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):

$$\rho' = \sum_{i=1}^{n} \rho_i$$

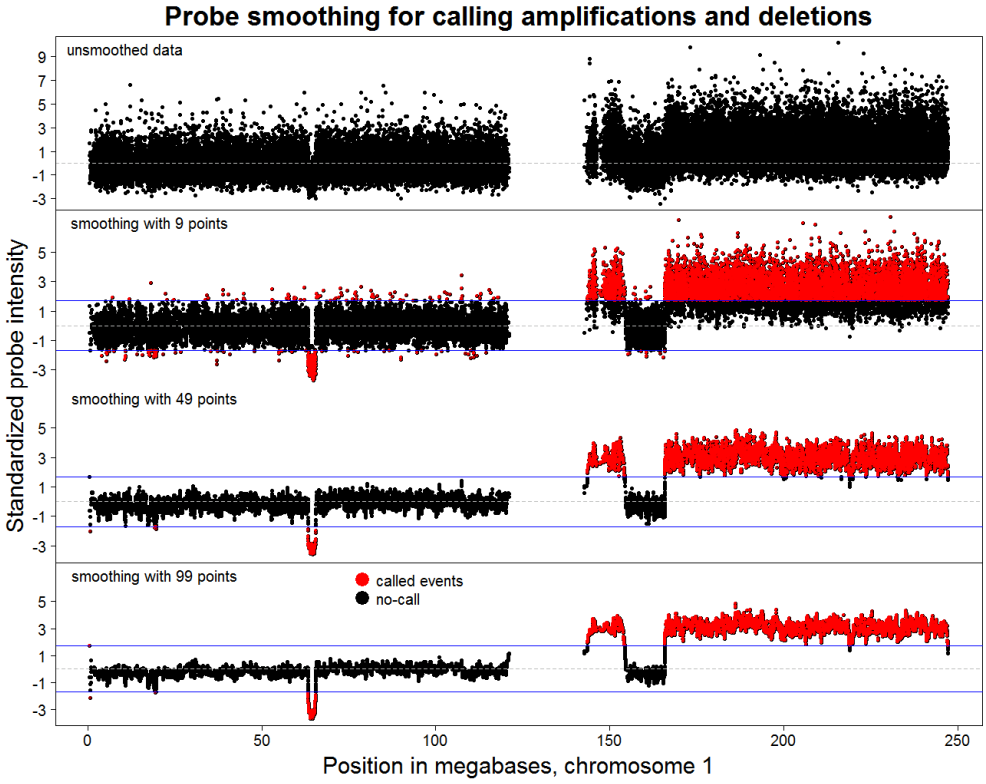

Figure 8: A representative primary tumor chromosome 1 copy number profile (top panel) and corresponding S(CN)n [n=9,49,99] in the bottom panels. Therefore, k=3 because three different smoothing lengths are used. Black probes represent events that are not called while red probes are the called events that exceed the predefined threshold of ±1.7 sdu.

One ρ' profile was thus generated per sample. Finally, summing across the ρ' profiles on a probe basis resulted in an event call index (ρ'') that reflected the frequency of amplification and deletions for a subgroup of samples or population. Two values of ρ'' were calculated for population or sub-population. 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}.

$$\rho''_{amp|del} = \sum \rho'_{i[amp|del]}$$

A summary of the different profiles used in this analysis is contained in table 4. An example of copy number ρ'' plot (figure 9), will be repeated in all results sections in a similar way. AA, CA or METS are displayed for a select region on chromosome 21 where the TMPRSS2-ERG deletion/fusion event is clearly observed. At each probe position, standardized values of both ρ''_{amp} and ρ''_{del} are shown. This region has been previously observed to be deleted, translocating the TMPRSS2 promoter, resulting in fusion with the ERG gene and amplification of ERG transcript in late stage primary tumors.[20,22] The chromosome 21 locus and corresponding genes confirmed those identified in previous copy number studies,[17] along with a series of novel candidates that have been associated with cancers and several with no known cancer function.

Table 4: Summary of Copy Number and Event Profiles	
Profile	**Description**
CN	Copy number profile derived by subtracting normal signal intensity from paired tumor signal intensity – 1 per sample
$S(CN)_n$	Smoothed copy number profiles derived by applying a running median to the copy number profile – *k* per sample, based on varying value of n, the number of probes used in the smoothing
ρ_n	Trinary bin of S(CN) profiles, with three values: amplification, deletion and no event – *k* per sample, as with S(CN)
ρ'	Sum of ρ values for each sample – ranges from $-\#n$ to #n, where n represents the number of different smoothings used – one per sample
ρ''_{amp}	Sum of all the ρ' values in a population which are amplified at a each probe – represents the extent to which the population has amplification events – one per population
ρ''_{del}	Equivalent of ρ''_{amp}, but represents the extent to which the population has deletion events – one per population

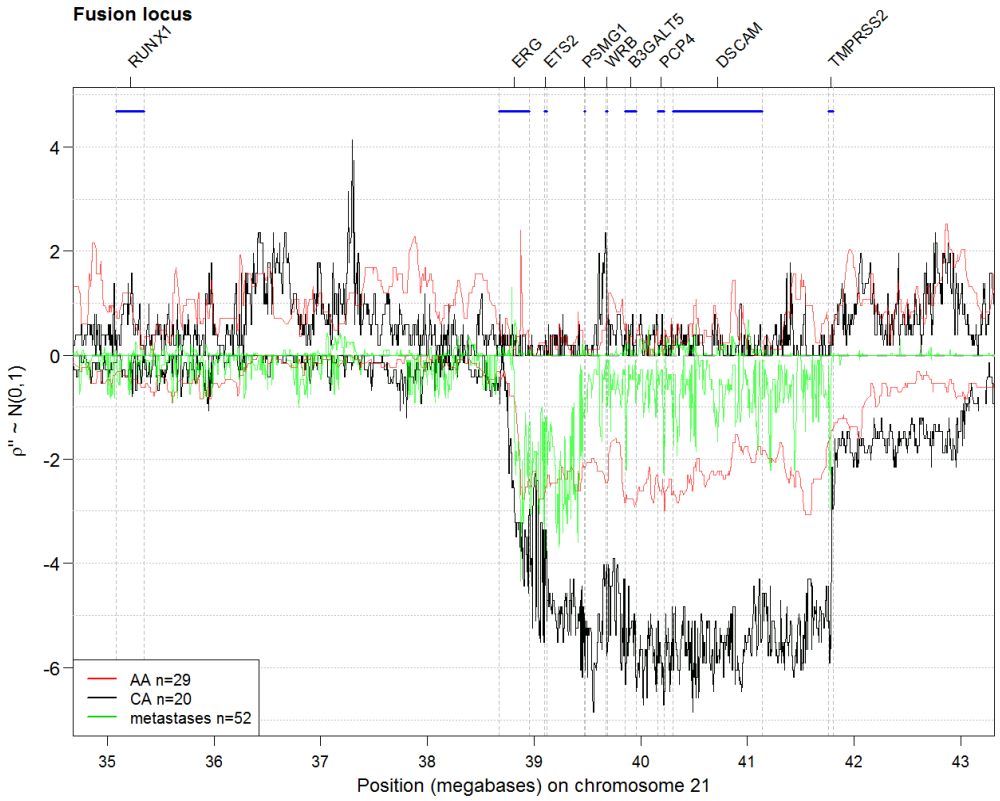

Figure 9: Copy number profile, ρ'' shows a deletion of a major chromosome 21 locus at the TMPRSS2-ERG boundary. The y-axis represents standardized population frequencies reflecting the number of samples exhibiting amplifications (above 0)/deletions (below 0). The populations are differentiated as red, black or green lines for AAs, CAs or METS respectively.

Stage 4 of the pipeline was designed to use the frequency profiles (ρ' and ρ'') to create evolutionary models of selection for and against prostate cancer metastasis. The profiles of AA, CA and METS subgroups were first processed through unsupervised hierarchical clustering. The rationale for unsupervised hierarchical clustering was the recurrently observed racial health disparity. Unsupervised clustering would be used to evaluate whether either of the two groups of primary tumors (AA or CA) preferentially segregated with the METS samples (12 individuals comprised of 52 METS). Knowledge of this racial disparity justified the expectation that AA individuals, having a greater propensity for aggressive disease, would preferentially cluster with the METS samples. CA individuals, meanwhile, would be expected to preferentially cluster away from the METS samples.

The clustering parameters (linkage and distance method) were optimized based on the model that the best parameters would result in the 52 METS ρ' profiles clustering by their respective individuals as nearly perfect as possible. Multiple metastases clustered [using the R function hclust() with the parameters of complete linkage and binary distance method] produced the best results, with only 2 METS clustering outside of their individuals cluster (Figure 10).

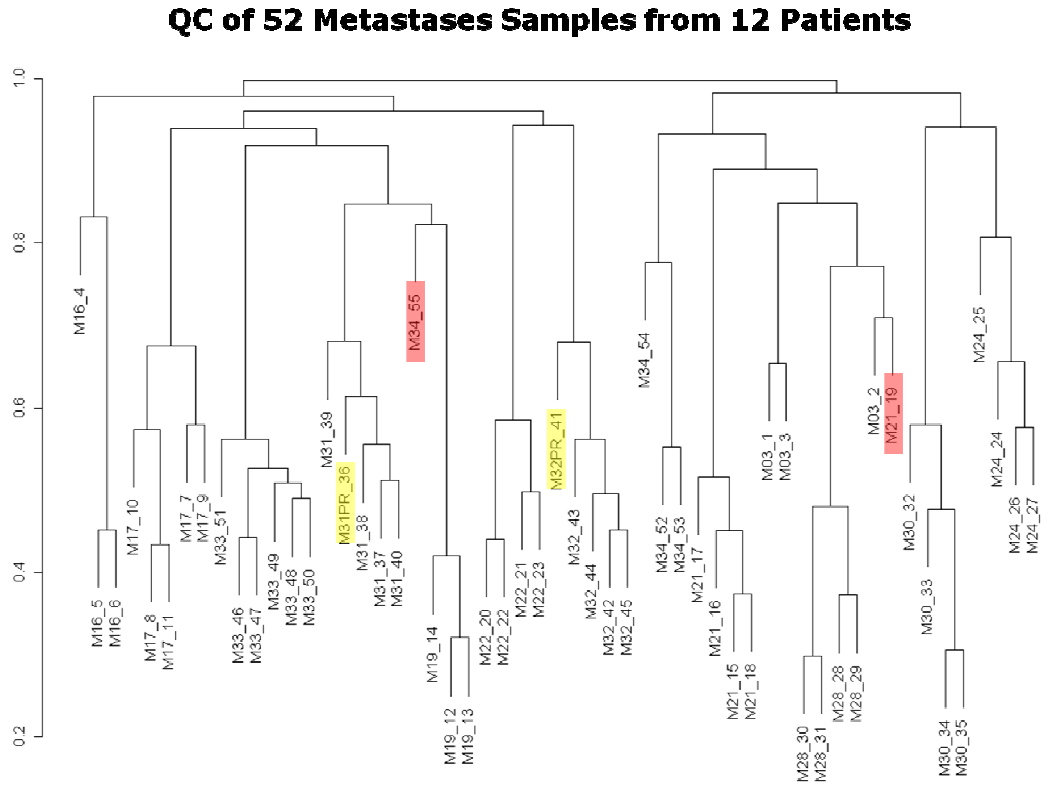

Figure 10: Quality control of hierarchical clustering using multiple metastases (n=52) from 12 individuals. Red highlights indicate the METS that segmented outside of their individual cluster node and the yellow highlights indicate the 2 primary tumors (PR_) available from 2 METS individuals. Dendrogram leaves are labeled by M_individual.index_sample.index

Next, to assess preferential segregation, ρ' profiles of AA and CA primary tumors were clustered with a version of a ρ'' profile for the METS subgroup. Unlike the standard ρ'', which was generated by a summation across all $\rho''_{amp|del}$ profiles from each sample in a population, this METS profile (ρ''_m) was created as a binary profile representing event or no event. For this analysis, an event was registered at those probe positions on the METS profile where at least 14 out of 52 METS ρ' profiles had an event register in either direction. The justification for the threshold of 14 was that it optimized the number of events in METS profile to be as close as possible to the average number of events in all the ρ' METS

profiles. The trinary bin was not used in this step of the investigation because the distance method used in the clustering algorithm was a binary method; therefore, the algorithm would convert a trinary bin to a binary one regardless of the user input.

After the primary profiles were clustered with the METS profile, an enrichment score E(AA) was calculated to assess whether the extent to which the ρ′ of AAs preferentially clustered in the dendrogram node that contained the METS profile as compared to the CAs:

$$E(AA) = \left(\frac{\#mPT_{AA}}{\#mPT_{CA}} \right) * \left(\frac{1}{1.45} \right)$$

AA/CA ρ′ that clustered within the node that contained the METS profile were classified as metastatic primary tumors (mPT), and the node was called the mPT node. The ρ′ that were outside of the MET cluster were classified as indolent primary tumors (iPT), and the node was called the iPT node. The score E(AA) was calculated for both the mPT node (as seen above) and the iPT node. A coefficient of $\left(\frac{1}{1.45} \right)$ was introduced into the E(AA) calculation in order to account for the difference in number of samples in the 2 subgroups (29 AA vs. 20 CA). A score of 1 would signify no AA or CA enrichment at a particular node. A significant score greater than 1 would indicate enrichment of AA over CA, and a significant score less than 1 would indicate enrichment of CA over AA.

To evaluate the robustness of this enrichment, a hierarchical clustering bootstrap methodology was employed by randomly sampling 50% of the

individual METS (n=6) 300 times to generate E(AA) scores for each mPT and iPT cluster node.

Genomic copy number alterations in advanced prostate tumors typically are numerous, systematic in their genomic placement and varied in size from a single nucleotide mutations to the amplification or deletion of an entire chromosome. Studying copy number alterations has clear informative value, reflecting the direction (i.e. the amount of protein made) in which the cell forces genes to reprogram their dose or structure for the purpose of maintenance and survival. When analyzing copy number data, however, a disproportionately large set of numerically significant passenger events representing weak-to-non disease associated copy number alterations mask the strong true-disease causing events. These passenger events may emerge by simply being in proximity to the genomic regions harboring the metastasis causing genes or by actually having weak causative members of pathways that follow the global trend of a compensatory response. Therefore, to demarcate the weak from strong events, models were developed based on the selective pressures influencing the progression toward metastatic disease.

Models were designed to score selection in two different contexts: unsupervised MET/iPT/mPT and supervised METS/CA/AA health disparity scenarios (Figure 11). The unsupervised (us) context was considered because the samples that clustered with the MET profile – the mPT samples – should be enriched for a poor outcome, whereas the iPT samples should be enriched for a better outcome. Therefore, it was worthwhile to analyze events that may distinguish mPT samples from the iPT samples. A racial disparity (rd) context was likewise considered in order to harness the model that certain events may

be overrepresented in primary prostate cancer samples across one race, affecting the odds of a favorable outcome and contributing to the disparity.

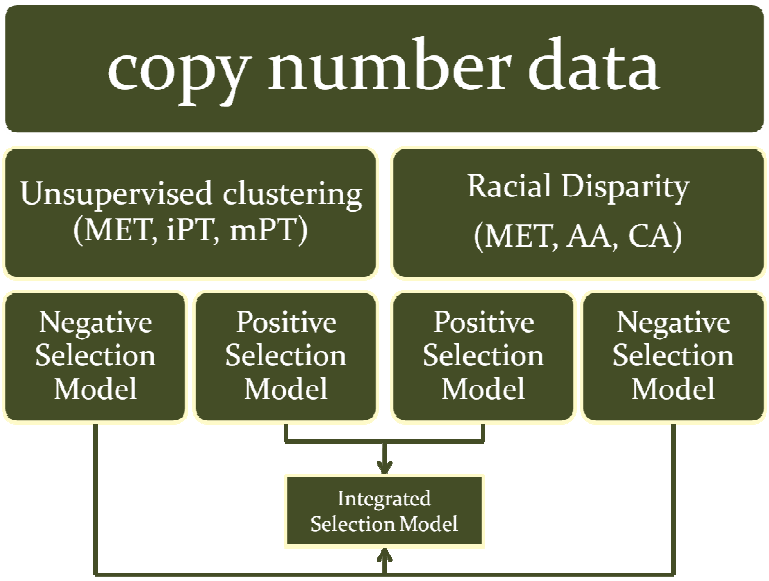

Figure 11: Scoring health disparity and metastatic potential through positive and negative selection modeling.

In this manuscript, the pipe "|" notation will be used to indicate that parallel selection analyses were performed in the us and rd contexts. Since the us and rd analyses were identical at many steps (eventually the two were integrated), the same formulae were applied to each one with the exception of that there were different input populations. A pipe in a formula indicates that the formula was applied to each context, each time using different inputs.

A negative selection model (NSM) based on rd|us was designed to detect enrichment for deletions or non-events (del or neu) in the METS and AA|mPT along with enrichment for amplifications (amp) in the CA|iPT. Probes exhibiting negative selection reflect a scenario in which there is a potential gene or functional locus that protects the individual from forming metastasis. This

represents selection *against* the primary carcinoma from metastasizing. Conversely, a positive selection model (PSM) based on rd|us detects probes that exhibit enrichment for amp in metastasis and AA|mPT but enrichment for del and neu in CA|iPT. In this scenario, there is a potential to detect genes that are deleterious for the individual because they positively select for metastatic cellular growth.

Because the SNP ranked most likely to undergo PSM should also be the SNP ranked least likely to undergo NSM, and vice versa, the NSM and PSM models can be reduced to a single integrated selection model (ISM). In order to create this model, an amplification enrichment score (*E(x)*) was first calculated to represent the relative amount of enrichment for amplifications versus deletions:

$$E(x) = \frac{(\#Amp - \#Del)}{\#Samples}$$

Using this amplification enrichment score, calculated for each SNP for each population, two selection models (SM) were developed. The first was to consider the unsupervised (us) hiearchical clustering scenario with subgroups mPT and iPT. The second was to consider the racial disparity (rd) subgroups, AA and CA. The coefficients (a,b,c) represent user-defined weights given to each subgroup that influence the outcomes based on nuances in the biological model (described in the results and discussion; the values used in this investigation were *a=3,b=1.5,c=3*):

$$SM(us) = a^{E(METS)} * b^{E(mPT)} * c^{-E(iPT)}$$

$$SM(rd) = a^{E(METS)} * b^{E(AA)} * c^{-E(CA)}$$

The first two terms being multiplied are designed to assign a higher score when the METS and AA|mPT samples have more amplifications than deletions. The greater the amplification enrichment, the higher the score; whereas the third term, due to the negative exponent, will be higher when the CA|iPT samples are enriched for deletions over amplifications. Therefore, a higher score will be an indication of positive selection, and a lower score will be an indication of negative selection.

At this point in stage 4, the us and rd selection models on the SNP level were ranked and combined to create an Integrated Selection Model (ISM):

$$ISM = \log\left(\sum_{i=1}^{2}\left(\log\left(\frac{1}{Rank(SM_i)}\right)\right) + k\right)$$

In this equation, SM_1 and SM_2 represent the us and rd selection models and *Rank(x)* represents a ranking of the values such that the highest rank represents most negative selection. *K* represents a scaling factor. This equation is designed so that a higher ISM score refers to greater probability of that SNP being involved in positive selection whereas a smaller ISM score indicates a greater probability of negative selection.

Figure 12 shows representations of copy number events in the SNPs located in genes that were later determined to be ranked very high for either

positive or negative selection. A sample with a value of zero represents a neutral event called at that SNP. A sample with a value of +1 or -1 represents either an amplified or deleted SNP, respectively. Figure 12a shows a SNP from PREX-2a, one of the top-ranked genes for positive selection. In this example, 27 METS samples exhibited amplifications with none exhibiting deletions, while all but one sample from the mPT node exhibited amplifications as well. The samples from the iPT no distribute equally between amplifications and deletions; nevertheless, this gene was ranked high due to the first two subgroups, which show strong evidence of positive selection.

Conversely, figure 12b shows a SNP located within the gene VASP. This marker was the #1 ranked SNP for negative selection according to the ISM. Of the 52 METS samples, 14 exhibit deletions with no amplifications. In the mPT samples, little is observed; however, the one sample that is deleted is AA and the one sample amplified is CA, thus fitting the rd context very well. Finally, there are more iPT samples (13) amplified than deleted (2), fitting the us context well and the rd context too, because the two deleted are AA. This clear example of negative selection fits well with a functional model of VASP.

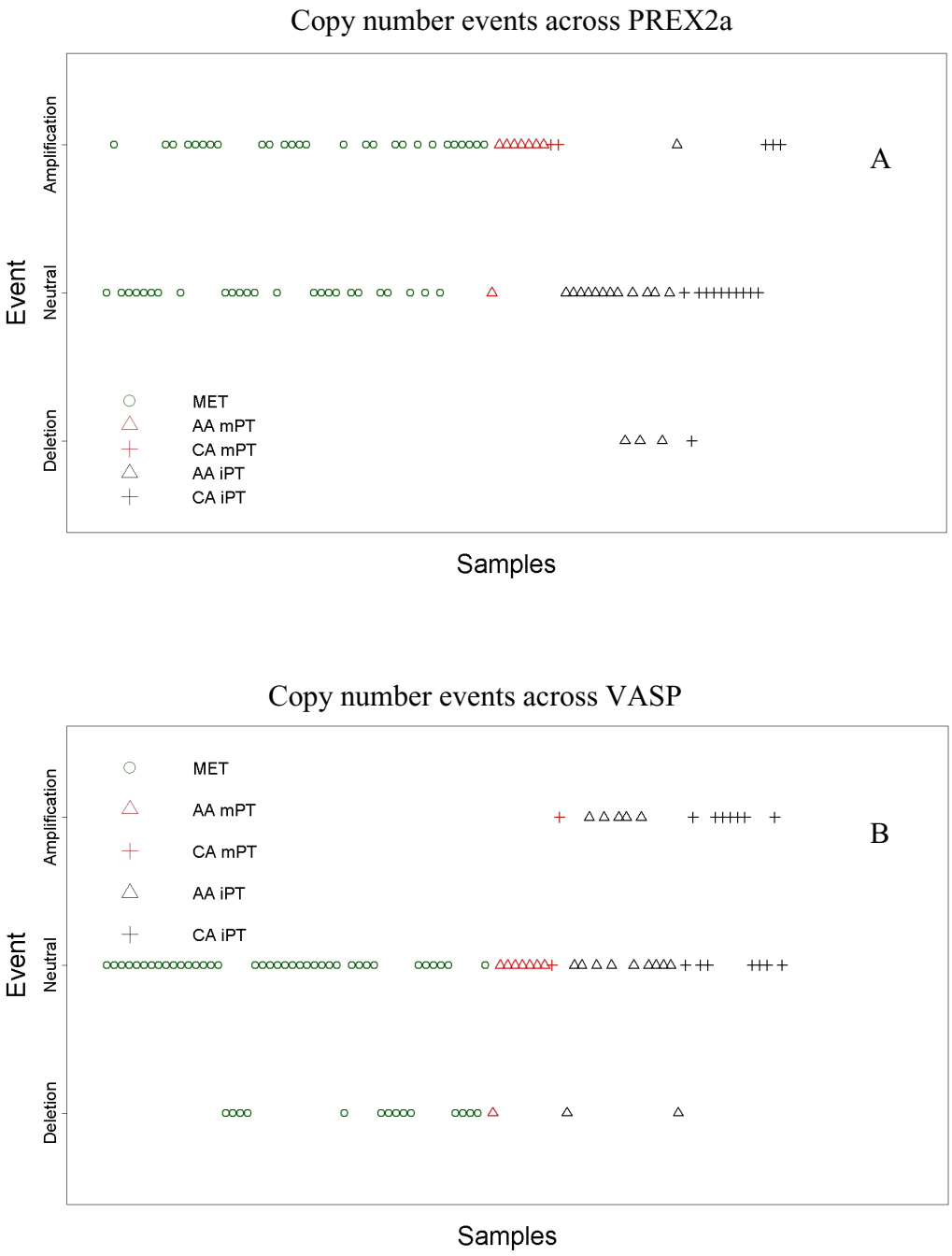

Figure 12: Top scoring SNP/gene for the health disparity and metastatic potential selection modeling, showing the top ranked candidate gene's amplification and deletion for (A) positive selection modeling (PREX2a) and (B) negative selection (VASP).

In the final part of stage 4, we decided that a model that simply ranks selection on a continuum from negative to positive could not entirely encompass all evolutionary scenarios. For example, in a positive selection (PS) scenario, there are two related but distinct possible manifestations of PS on the copy-number level. First, a gene that positively selects for metastatic disease could be one that is never or rarely deleted in METS samples. On the other hand, another scenario could be that CA samples in the area of this gene are never or rarely amplified. We thus created two different models reflecting the two "flavors" of positive selection mentioned:

$$PSM_{f1}(rd \mid us) = a^{E(METS)} * b^{E(AA|mPT)} * c^{-E(CA|iPT)} * d^{-\#met\_del}$$

$$PSM_{f2}(rd \mid us) = a^{E(METS)} * b^{E(AA|mPT)} * c^{-E(CA|iPT)} * d^{-(\#CA\_amp|\#iPT\_amp)}$$

In these flavors, an additional coefficient is multiplied each time to the original ISM. This coefficient serves as a "penalty." It reduces the measure of significance if there are deleted METS or amplified CAs, respectively. Like a- c, d is a user-defined parameter. Note that for each flavor, a us and rd model was created, which were integrated using the same integration formula applied to the plain ISM before being used in further analysis. According to this scheme, we also created two "flavors" of negative selection modeling:

$$NSM_{f1}(rd \mid us) = a^{-E(METS)} * b^{-E(AA|mPT)} * c^{E(CA|iPT)} * d^{-\#met\_amp}$$

$$NSM_{f2}(rd \mid us) = a^{-E(METS)} * b^{-E(AA|mPT)} * c^{E(CA|iPT)} * d^{-(\#CA\_del|\#iPT\_del)}$$

Here, a penalty is applied to decrease the calculated significance of a SNP if there are amplified METS or deleted CAs, respectively. As with the original ISM, each of these flavors attempt to tease out the loci exhibiting the most extreme form of positive and negative selection.

In **stage 5**, each SNP was designated as either a positively selected (PS) SNP or a negatively selected (NS) SNP, depending on whether it was ranked in the top or bottom 50% by the ISM. All the genes overlapping NS or PS SNPs generated were ranked by calculating a z-score for each, with X = ∑[ISM] 0.95 quantile of measures whose probes map to within a predefined distance away from the gene (we used 10kb in this analysis) although results were stable for distances of 0kb, 2kb, 10kb, 50kb, 100kb and 500kb. The background parameters (μ, σ) were estimated by random sampling an equal number from the 0.95 quantile of the corresponding measures of all mapped genes. The resulting ranked sets of genes and associated Z-scores were used for multi-data-type integration (sections 2.7 & 3.4). To measure the robustness of the resulting rankings of the candidate genes, we ran a bootstrap analysis on the model scores by randomly sampling 80% of the samples from each of the population groups (AA, CA and METS) and evaluating the reproducibility and range of rankings.

2.5 Expression Data Analysis

To complete an expression analysis, Affymetrix U95A or U133 human gene expression array (.cel) intensity files were processed using Affymetrix expression console v1.1. Probes were summarized using robust multi-array

average (RMA)[131] without normalization and exported as $\log_{10}$-signal intensities. Analyzing each data-set separately, the complete matrix of signal intensities was standardized N[0,1] on a sample basis. Significance testing was run using the Student T test followed by Bonferroni correction for multiple testing on a gene basis was calculated [using the R functions t.test() and p.adjust()]. Significant genes ($p < 0.05$) were further integrated with copy number and GWAS data and processed through network connectivity and gene-set enrichment analysis (described in chapter 2.7).

2.6 CGEMS GWAS data analysis

CGEMS GWAS data consists of a matrix of single marker p-values, representing the case/control association statistics run on the PLCO dataset evaluating 1,172 cases and 1,157 controls. The normal genomes of individuals diagnosed with prostate cancer (cases) were compared to genomes of individuals from an age and ethnically matched group of men without a prostate cancer diagnosis (controls). The cases and controls were interrogated by measuring the SNP frequency distributions and calculating a χ^2 statistic from data assayed on an Illumina 550K array. The p-values associated with each SNP were associated with genes that mapped to less than 500 kilobases (kb) away. Each corresponding p-value was converted to the –log(p). The –log(p) values were summed for each gene and a z-score was calculated with the parameters μ and σ estimated by random sampling of an equal number of –log p-values from the complete set of SNP probes that map to genes. The z-scores were integrated with copy number and gene expression data (described in Chapter

2.5) and further analyzed by network connectivity and gene set enrichment analysis.

2.7 Multi-data-type integration and functional data mining

Each of the data types: copy number, gene expression and GWAS data types yielded sets of numerically significant genes. As indicated earlier, in genomic analysis, the functionally meaningful loci may be peppered within a mound of numerically significant passengers or weakly associated loci. By combining these multiple orthogonal data types, we expected the most significant functional sets or pathways to emerge. A combined set E, for each locus E(L) was derived by taking the union of the significant (Z > 1.6sdu) CN, GE, and GWAS Z-scores (signified by S) as determined in their respective individual analyses:

$$E(L) = S_{CN} \text{ Y } S_{GE} \text{ Y } S_{GWAS}$$

The ranked list of genes as derived by E(L) were analyzed through networks of putative protein interactions and functional pathways to establish possible connectivity among these candidates. First, the top-ranking genes were analyzed for network connectivity using the Biogrid curated protein-protein interaction database[160]. A network connectivity score was calculated for the combined set of significant genes in the following way:

$$C(N) = \sum_{i=1}^{n} \left(l * \sum_{k=1}^{n-1} \frac{1}{D_{ik}} \right)$$

In this equation, *n* represents the total number of seed genes in the network. D_{ik} represents the network distance between seed genes *i and k,* and l represents the number of seed genes that are connected to seed gene *i* in network N. The implementation of the C(N) score required a network object, containing a symmetrical adjacency matrix of 0s and 1s to represent the interactions of the Biogrid database. In the adjacency matrix, the rows and columns represent genes found in the database and each cell x,y containing a 1 indicated an edge between genes x and y. A graph object for network N was created by isolating genes that share edges with the n seed genes in the adjacency matrix [using the R function graph.adjacency(); library(igraph)]. Once in a graph object, the shortest path between two vertices was calculated [using the R function get.shortest.paths();library(igraph)] to score the seed interactions in terms of network depth. Therefore, a seed gene connected indirectly to multiple other seed genes will achieve a higher C(N) than a seed gene with only one other primary connection.

Z-score parameters were estimated from a reference distribution of C(N) by generating 10,000 random networks, using the same number of seed genes as observed in E(L) set of genes.

CHAPTER 3

3.1 Copy number analysis of somatic tumors

Two parallel sets of copy number analyses were performed based on the type of Affymetrix SNP array on which the samples were run. The CA/AA primary and the public metastasis data-sets were both run on the version 6 array comprised of ~1.8 million SNP and copy number probes, all of which could be used for copy number analysis. The third public data-set of 20 AA primary tumors, however, was run on a older version Affymetrix array of ~500K SNP probes. Thus, in order to increase the statistical power, an integrated analysis of the ~472,000 probes common to all three datasets was considered. Copy number and Loss of Heterozygosity (LOH) frequency distributions were calculated for the AA/CA and MET data using the 6.0 probes (Table 5a). In addition, these were calculated for all three data-sets using probes common to both platforms (Table 5b).

Table 5a Copy number and LOH frequency distributions

Dataset	Number of tumors (AA/CA)	amplifications*		deletions*		LOH*		LOH/HET ratio	
		AA	CA	AA	CA	AA	CA	AA	CA
Primary African/Caucasian American**	9/20	114 (±37)	106 (±31)	101 (±32)	90 (±30)	5 (±6)	8 (±8)	0.009 (±0.010)	0.014 (±0.014)
Metastasis African/Caucasian American (public data)**	0/52		57 (±19)		40 (±34)		26 (±11)		0.113 (±0.047)

*Amplifications and deletions were called at any snp for which the rho score was a positive or negative number, respectively (Data in thousands of SNP probes)
**Distributions based on the 868157 SNPs common to all three datasets

Table 5b Copy number and LOH frequency distributions

Dataset	Number of tumors (AA/CA)	amplifications*		deletions*		LOH*		LOH/HET ratio	
		AA	CA	AA	CA	AA	CA	AA	CA
Primary African/Caucasian American**	9/20	63 (±20)	57 (±15)	54 (±17)	50 (±19)	2 (±3)	4 (±4)	0.008 (±0.010)	0.011 (±0.013)
Primary African American (public data)**	0/20	55 (±17)		51 (±11)		3 (±2)		0.023 (±0.013)	
Metastasis Caucasian American (public data)**	0/52		33 (±10)		20 (±16)		15 (±6)		0.115 (±0.051)

*Amplifications and deletions were called at any snp for which the rho score was a positive or negative number, respectively (Data in thousands of SNP probes)
**Distributions based on the 471012 SNPs common to all three datasets

The copy number event profiles (ρ'/ρ'') were generated for each sample (ρ') and for each population (ρ''): AA, CA and MET. The profiles were processed through

unsupervised hierarchical clustering and modeled on a probe basis to generate ten scores reflecting unsupervised metastatic potential and AA/CA racial disparity (usNSM, usPSM, rdNSM, rdPSM, iNSM, iPSM, $usNSM_{f1}$, $usPSM_{f1}$, $rdNSM_{f2}$, $rdPSM_{f2}$).

Observing the clustering indicated that in addition to the mPT and iPT nodes, it was necessary to define an intermediate "mid" node. Thus, the African American Enrichment score *E(AA)* was calculated for the mPT, iPT and mid nodes. The unsupervised hierarchical clustering of METS ρ''_M with CA/AA ρ' resulted in a compelling observation (Figure 13). The enrichment for AA was calculated to be 2.76 in the mPT node, consistent with the racial disparity of mortality reported by the American Cancer Society Cancer Facts and Figures for 2008[1]. The enrichment values for the mid and iPT nodes were, respectively, 0.86 and 0.79. To evaluate the stability of this result, a bootstrap of the

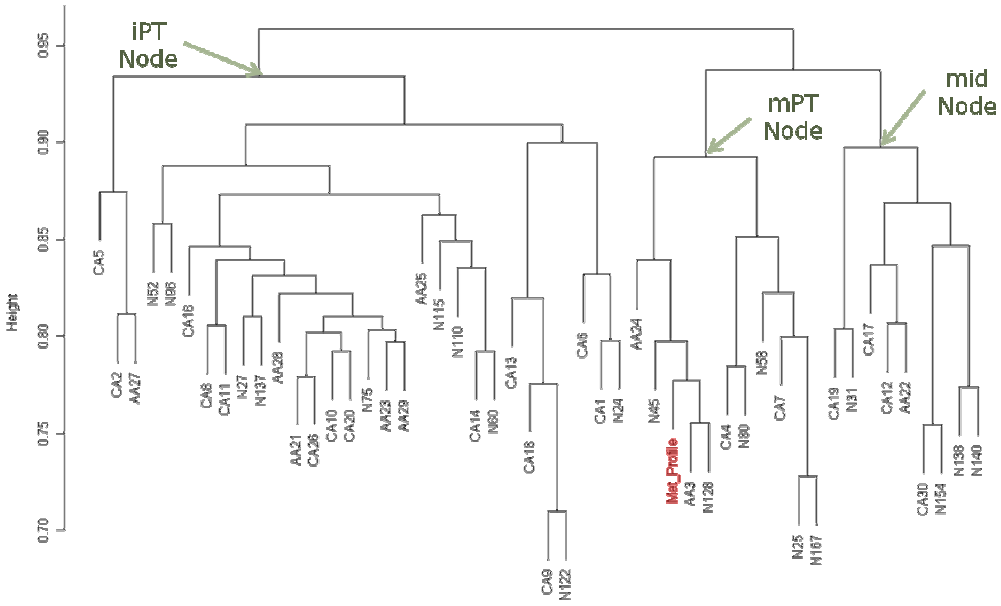

Figure 13: Unsupervised hierarchical clustering of METS, AAs and CAs

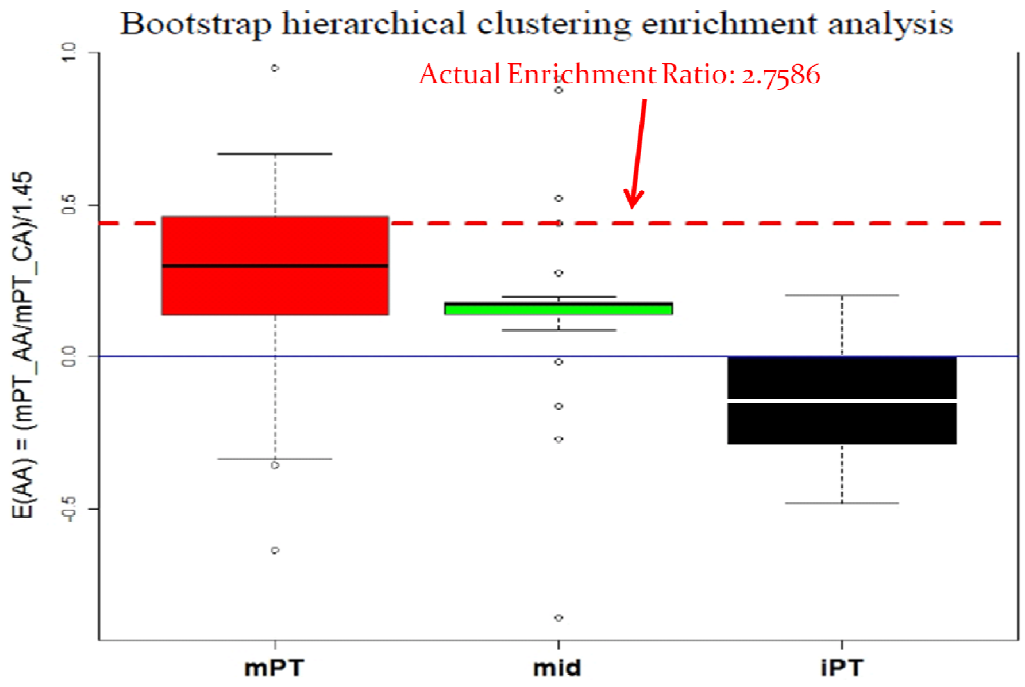

Figure 14: Bootstrap analysis of unsupervised hierarchical clustering of METS and primary tumors from AAs and CAs. Boxplots represent the AA enrichment scores (y-axis) for the mPT, mid and iPT cluster nodes.

clustering analysis using random samples from 50% of the METS was employed. The results in Figure 14 show the distribution of bootstrapped E(AA) estimates for the mPT, mid, and iPT cluster nodes. The boxplot distributions between the various cluster nodes reflect the number of bootstaps where the METS profile clustered within the respective cluster. Therefore the mPT node, with a 2.78 fold enrichment of AAs over CAs had a greater representation of the METS profile over the iPT node during the course of the 300 bootstrap iterations.

Next, a result of the combined unsupervised- and racial disparity-based negative selection score ($NSM_{rd|us}$) is shown in the bottom panel of Figure 15. The top ranking gene exhibiting negative selection model characteristics was

identified on chromosome 19q13.2. VASP, encoding the vasodilator-stimulated phosphoprotein was previously implicated along with a network of focal adhesion proteins to play a role in filamentous actin formation, functioning in the biological processes of adhesion and migration. VASP protein localizes to the mitochondrial membrane, negatively imparting control over metastatic potential in a variety functional cellular assays.[161-164]

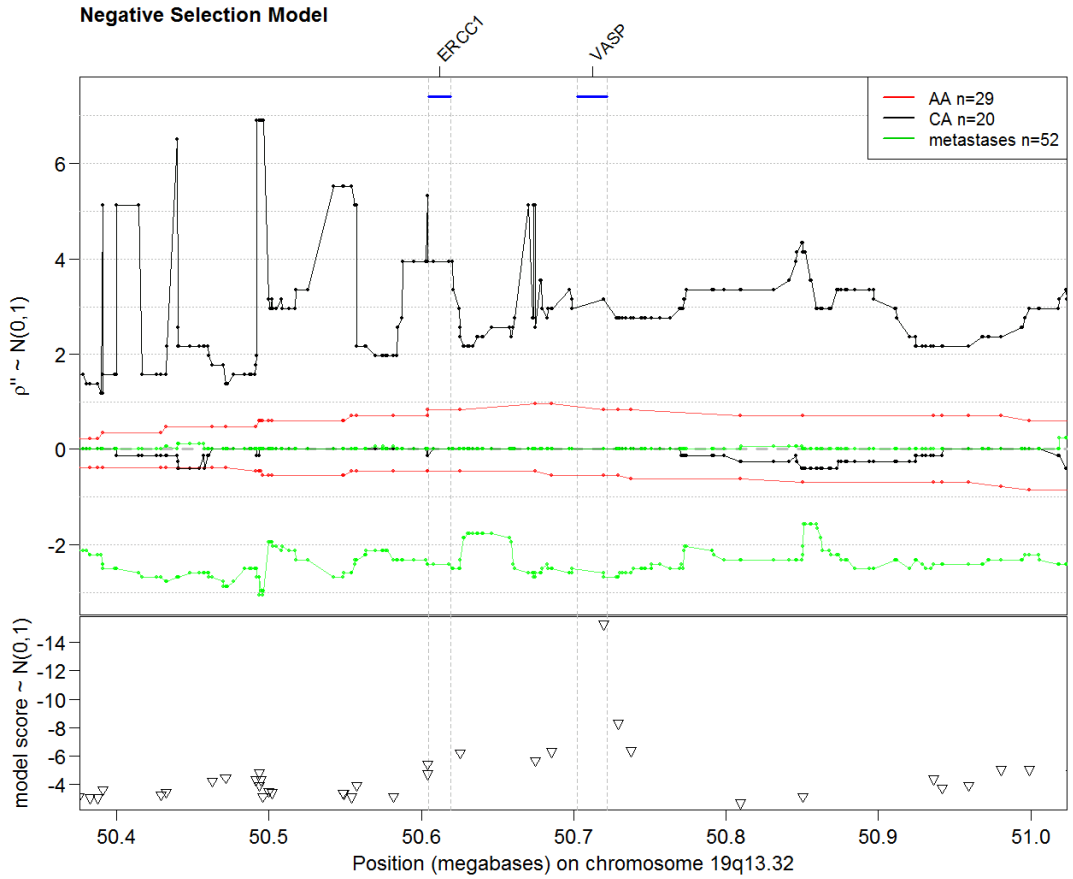

Figure 15: ρ'' profiles (top-panel) of the VASP and ERCC1 genomic region identified through the integrated selection model(bottom panel).

The negative selection forces characterizing VASP in our data model are consistent with reports claiming VASP inhibitory activity in a variety of cells such as fibroblasts, endothelial, epithelial and neuronal (for review[165]). Notably,

exogenous expression of VASP has been shown to inhibit cell migration in invasive breast cancer cells,[163] impeding their metastatic potential. VASP is located within a large stretch of the genome that exhibits the NSM scenario of CA_{amp} , $AA_{amp|neu}$ and $METS_{del|neu}$. Through bootstrap analysis of 80% of the population samples (at 80 iterations), we observed 47% of the bootstaps ranked VASP #1, 70% in the top 5 and 81% at least the top 10. Around 100kb upstream of VASP is the excision repair cross complementing 1 (ERCC1), an endonuclease subunit of the nucleotide excision repair complex, with a putative function in the repair of double strand breaks during homologous recombination. Interestingly, a study from lymphocyte samples taken from prostate cancer patients prior to radiotherapy attempted to correlate a SNP haplotype within the ERRC1 gene with levels of ERCC1 transcript showed a weak association.[166] A 12-fold variation in ERRC1 gene expression was observed among individuals; however, this variation in ERCC1 gene expression is most likely the result of genomic copy number variation,[166] as observed in our investigation. The close proximity of two potential candidates genes functioning in pathways of repair and motility can influence carcinogenesis and metastasis through mutation surveillance and control of mobility.

Conversely, a result of the combined unsupervised and racial disparity based on a positive selection score is shown in Figure 18 (plotted with AR in section of GWAS candidate genes). The top scoring $PSM_{rd|us}$ gene was oligophrenin-1 (OPHN1), located on chromosome Xq12, 318,597 bases downstream of the AR locus and part of the amplified region enriched only in METS. Boostrap analysis showed a consistent ranking for OPHN1 (40% of bootsraps ranked #1 and 100% ranked < 5). OPHN1 codes for a GTPase-

activating protein that functions in cell-matrix adhesion and membrane trafficking.[167] OPHN1 mRNA has been observed to be constitutively overexpressed in colon cancer[168], glial tumors[169], and overexpressed in gastic cancer exhibiting lymphovascular invasion.[170] Although these reports are consistent with our findings of OPHN1 ranking #1 in our positive selection model, the mechanism for this selection does not show a racial disparity, since the events around this region of chromosome Xq12 are specific for METS that underwent androgen ablation therapy. This is supported by the prevalence of amplifications in the METS and the lack or enrichment for AAs or CAs for either event. Interestingly another gene (RPNII), observed to be downregulated in lymphovascular invasion[170] was ranked as the #3137 NSM gene. Although seemingly uninformative in ranking, at this locus, only AAs showed deletions, whereas METS were all neutral and a equal number of AAs and CAs showed amplifications (similar to the NME4 distributions described below). Therefore, the complexity of the METS with the added component of androgen ablation requires careful interpretation of the ρ' distributions to qualify the selection model scores at any rank.

The #2 and #3 ranked PSM candidate genes are potassium channels KCNB2 and KCNQ3. KCNB2 potassium channels represent complex ion channels,[171] and have diverse functions ranging from regulation of insulin secretion to control of smooth muscles.[172] The KCNQ3 channel has been associated with neurological conditions such as epilepsy and neonatal convulsions.[173] Because of the high significance registered by these two genes in the positive selection model and the potential for pharmacologic utility

explained by the fact that they are protein channels, these two areas are worthy of further investigation.

More obvious is the fourth-ranked gene on the integrated PSM. Located roughly 4.5 megabases away from KCNB2 in the unstable and heavily amplified region of chromosome 8q13, PREX2a (phosphatidylinositol 3,4,5-trisphosphate RAC exchanger 2a, also known as DEPDC2) is an exchange factor that is believed to interact with the putative tumor suppressor PTEN.[174] Similarly to VASP and OPHN1, the bootstrap analysis of rankings showed 68% in top 5 and 93% in top 10 ranking genes. P-REX2a (Figure 16) is very consistent with a positive selection model favoring metastatic and tumor growth, as increased levels of the exchange factor have been shown to block PTEN in cancer cells (PTEN is discussed in section 3.2). P-REX2a protein decreased PTEN lipid phosphatase activity and affected other functions of PTEN, such as rescuing PTEN suppression of insulin signaling by inhibiting the P13K pathway of PTEN. P-REX2a thus may also influence the metastatic potential-affecting properties of PTEN by decreasing cell apoptosis and restoring cell growth. Mutations in *P-REX2a* have been observed in other tumors of the colon, pancreas and lung, consistent with our PSM findings based on ρ' copy data.

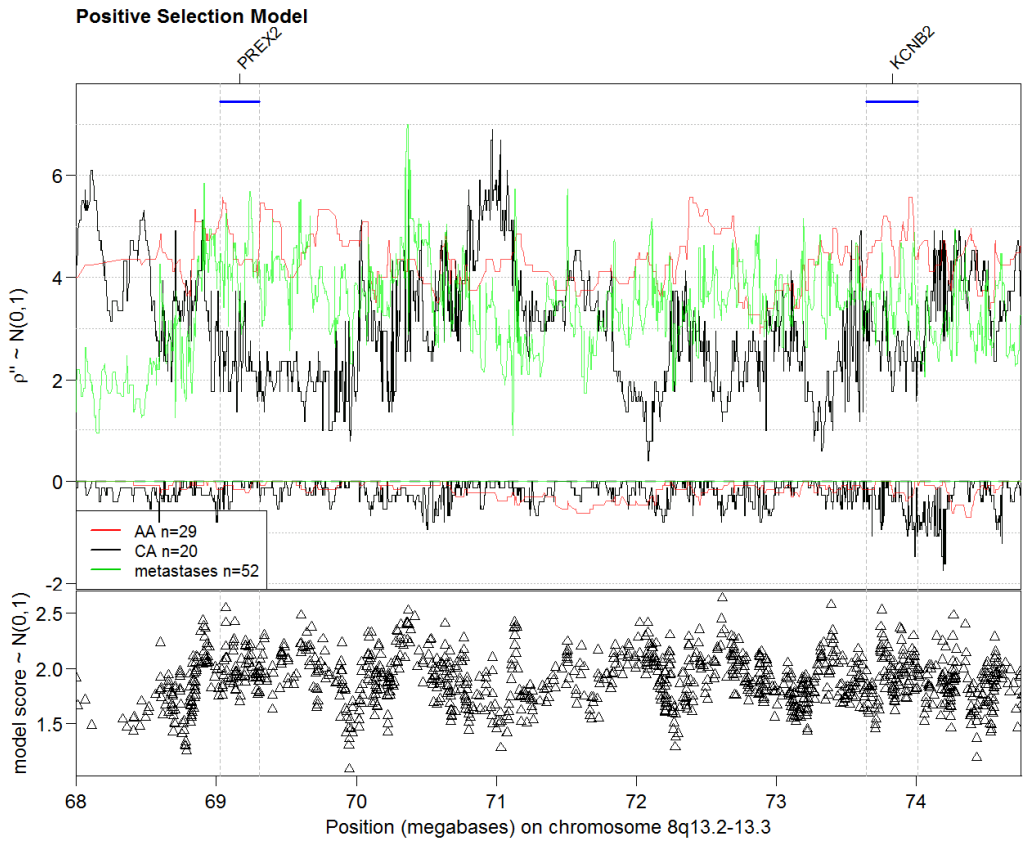

Figure 16: ρ'' profiles (top-panel) of the #2 and #3 ranked positive selection genomic region as calculalted by the integrated selection model (bottom panel) overlaps with the KCNB2 and PREX2 gene.

3.2 Gene expression of somatic tumors

Gene expression analysis indicated a number of potential genes with possible biological significance. Significant genes from a gene expression data-set comprised of AA/CA somatic tumors (n= 19 AAs, 16 CAs) and a data-set of CA normal and matched tumor tissue from the same patient (n=50) were evaluated for their correlation with genomic instability in the AA/CA primary tumors and METS. Figure 17 shows the relationship between the copy number and gene expression at the 10q24 loci harboring the putative tumor suppressor gene PTEN. PTEN regulates cell growth, apoptosis, cell adhesion and cell

migration. The most commonly proposed pathway by which PTEN operates is the phospholipid 3-phosphatase activity (PI3k/Akt pathway). PTEN, however, has been shown to influence Androgen Receptor (AR) through an AKt-independent pathway making it more vulnerable to degradation by enzymatic activity.[175] In the AA/CA gene expression data-set, PTEN was ranked 3rd ($p = 8.8E-6$, Bonferroni corrected) and shown to be increased in AA versus CA subgroups. PTEN was shown to be deleted in a large proportion of AAs, CAs and METS as indicated by the ρ''_{del} (figure 17: top panel). Most notably, PTEN copy number as indicated by ρ''_{amp} was observed to be elevated only in the CA subgroup.

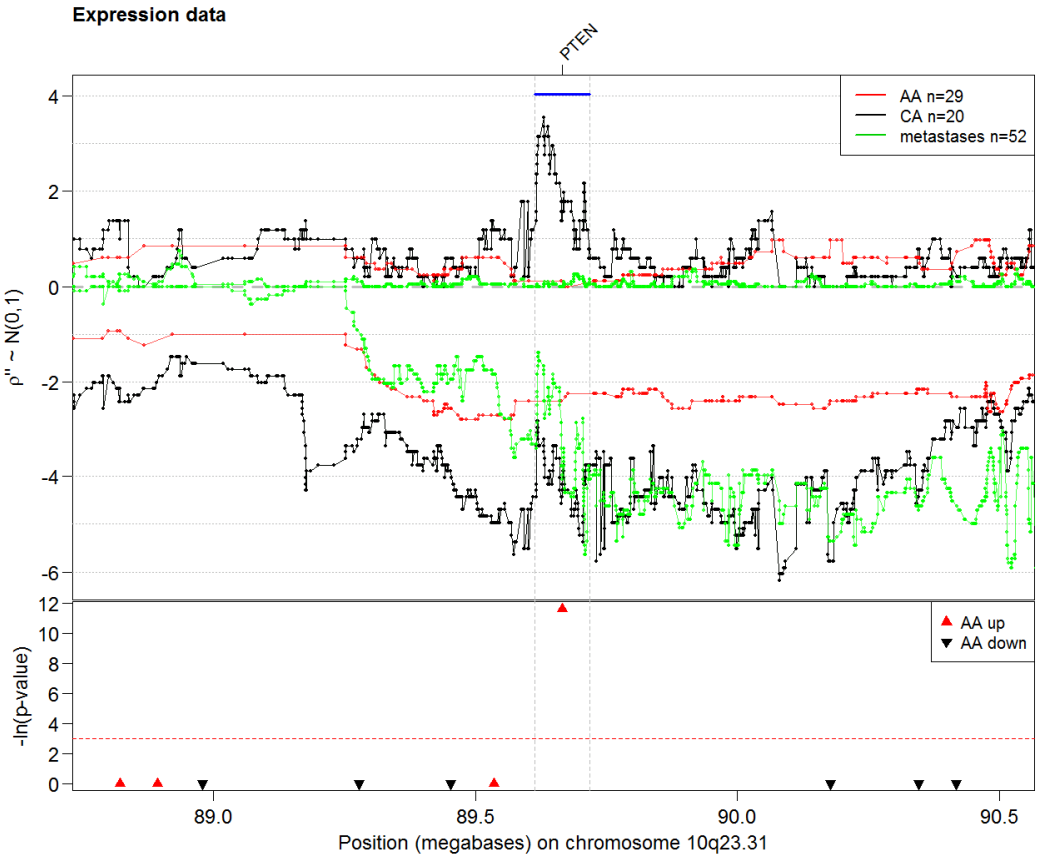

Figure 17: PTEN gene expression Student T derived -ln(p-value) (bottom panel). ρ'' profiles (top-panel) overlapping the PTEN gene.

Another interesting candidate gene called non-metastatic cells 4 (NME4), localized to chromosome 16q.13, is part of the nm23 nucleoside diphosphate kinase family and has an amino terminal domain that targets the protein to the mitochondria.[176] Members of the nm23 family catalyze the transfer of γ-phosphate from nucleoside triphosphates to nucleoside diphosphates and have a putative rolls in intracellular nucleotide homeostasis, differentiation, development, metastasis and cilia functions.[177] NME1 and NME2 have been observed to be upregulated in solid tissue tumors. Interestingly, in metastasis of melanoma, breast, liver, ovary and colon, a decreased expression of transcript was observed.[178] Our results show that NME4 gene expression was decreased in AA primary tumor subgroups as opposed to CA subgroups (rank = 17; p = 4.72E-04; Bonferroni corrected), whereas NME1 and NME2 were significantly upregulated in the primary tumor versus paired normal gene expression data-set[152] (NME1: rank = 16 ;p = 9.4E-07; NME2: rank = 70; p = 2.49E-04, Bonferroni corrected). Accordingly, the copy number profile of ($\rho''_{amp|del}$) at the NME4 locus showed preferential deletions in AA primary tumors and preferential amplifications in CA primary tumors, whereas METS from Caucasian individuals exhibited a ρ'' of neutral or no events (Figure 18).

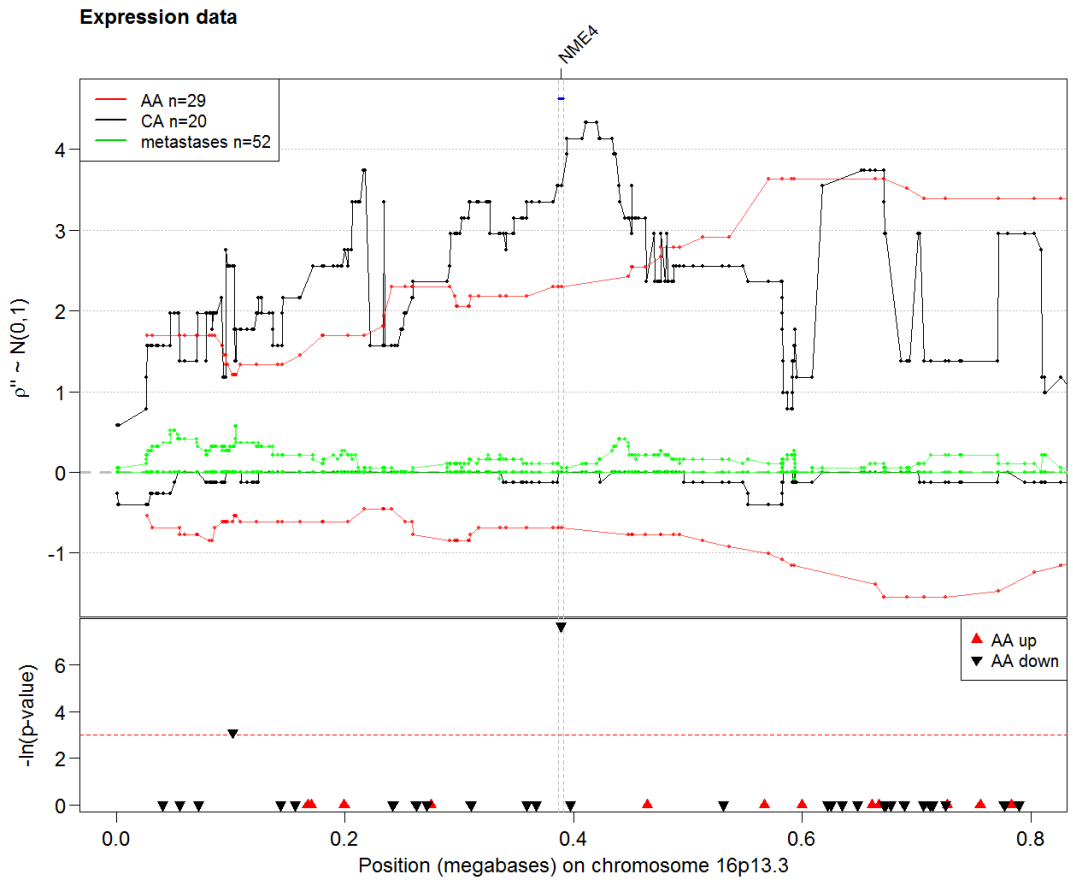

Figure 18: NME4 gene expression Student T derived -ln(p-value)—(bottom panel) and corresponding chromosome 16p13.3 region (top panel) displaying copy number ρ'' profiles.

3.3 Genome wide association analysis

Prostate cells require androgen stimulation in the form of testosterone and 5α-dihydrotestosterone (A) for normal growth and maintenance. The androgen receptor (AR) is the key receiver of this signal, the binding of which targets the complex AR-A to the nucleus where it acts in the transcription of androgen response genes.[179] These genes are involved a variety of functions including as metabolism, proliferation and stress response.[180] In the CGEMS-PLCO GWAS dataset,[153] the androgen receptor gene was ranked 7th based on its

proximity (<500kb) to a cluster of SNPs significantly associated with prostate cancer cases versus controls (Figure 19, bottom panel).

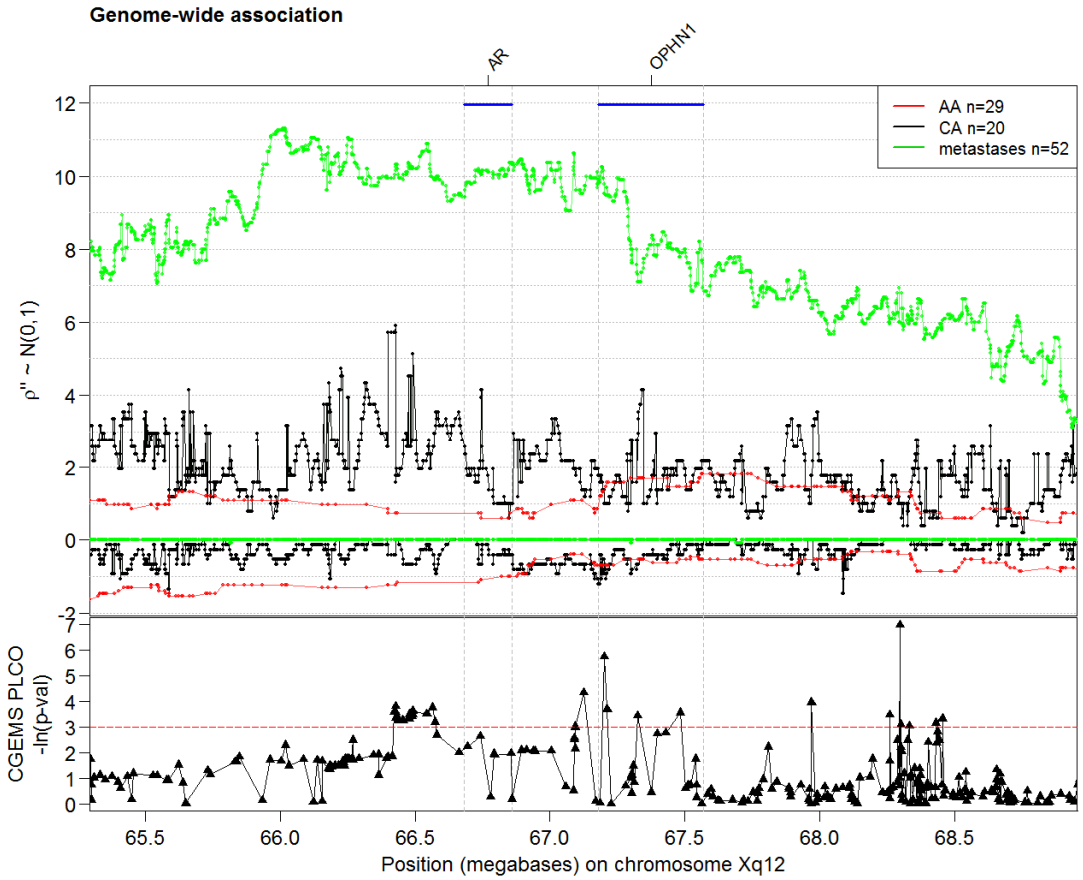

Figure 19: Androgen receptor and OPHN1 locus identified through GWAS (bottom panel), shows significant peaks as represented by the –ln(p-value) shown for SNPs on chromosome Xq12. Upper panels represents the copy number ρ'' profiles across the AR OPHN1 region.

This region of chromosome Xq12, shown to be highly amplified in METS (Figure 19, top panel), ranked #30- gene in the $PSM_{rd|us}$, driven by the fact that METS were highly amplified and majority of CA samples were neutral. Although this region was reported previously by the authors of the METS dataset,[16] there is an important feature that emerges when the METS $\rho''_{amp|del}$ are aligned with the $\rho''_{amp|del}$ of the primary CA and AA tumors. It appears that METS of individuals, who are typically treated with androgen ablation therapy, resulted in strong

amplifications around the AR gene. On the contrary, primary tumors from either AA or CA patients – who would not have been treated with androgen ablation – did not exhibit such amplification, indicating the evolutionary path of the METs had not been selected.

This discrepancy fits in well with an evolutionary model of metastatic response to androgen ablation therapy. Besides increasing the number of copies to promote a dosage effect, the amplification could be a form of accelerated evolution geared to reprogram the AR gene through semi-random recombination, rearrangement, and mutation. This may help the AR gene to relax its specificity for growth ligands or produce a variant resistant to negative intercellular forces leading to a androgen independent deregulated function. This theory is consistent with observations of rearrangements occurring within regions of amplifications made in MCF-7 breast cancer cell lines [181] and more recently in two lung cancer cell lines, NCI-H2171 and NCI-H1770 (small cell and neuroendocrine, respectively) using massively parallel sequencing.[182] In each study, a complex process of rearrangements was observed, including deletions, inversions and tandem duplications. It was suggested by the authors of the lung cancer study:

> "The complexity that emerges from the analysis of the NCI-H2171 amplicons implies that amplification involved an iterative process during which aberrant sister chromatid exchange to repair double-stranded DNA breaks led to progressive reorganization and expansion of the amplicons under selection pressure."[182]

This is further supported at the somatic gene expression level by the recent identification of 3 novel AR splice variants (AR3, AR4, AR5), each lacking their ligand binding domain. AR3 transcript, the major splice variant, was

observed to be upregulated during prostate cancer progression, correlated with poor outcome, and shown to be constitutively active in an androgen-independent manner.[183] Although AR does not reflect a copy number NSM/PSM score consistent with racial disparity during the primary tumor phase, the collection of genes presented above and others of yet unknown function may impart their selective forces on the way that AR ultimately manages to circumvent androgen ablation therapy.

3.4 Integrated analysis of copy number, gene expression and GWAS

Current complex genomics assays yield signals that are confounded by both technical and biological noise. The technical noise is handled by a simple scaling and standardization procedure that assumes a distribution close to normal. Biological noise, however, is more difficult to interpret. The distribution of numerically significant measures will vary on an individual and data-type basis. Since current genomics-based analyses range between thousands and millions of measures, procedures established to correct for multiple testing are generally encouraged. In the analysis of cancer genomes, however, the establishment of a significance threshold in the context of high-dimensionality and multiple testing is difficult to justify. Cancer genomes are highly variable and passenger signal dwarfs the driver signal in frequency, magnitude and recurrence. For example, megabase stretches of genome are amplified and deleted, along breakpoints of fragile site sequence (conserved regions of chromosome susceptible to recurrent DNA strand breaks) or origins of replication (starting position of transcription machinery). Only one gene, however, within a

region spanning 50kb – 5% of the numerically significant region – is a strong driver gene with utility in diagnosis or treatment of cancer.

In this study, the copy number data was integrated using a conditional heuristic based on evolutionary principles of selection and the health disparity observed between AAs and CAs, in the context of events observed in METS. The integrated model (ISM) score infers whether a gene has a propensity for positively or negatively driving the cancer (described in Chapter 2.6). In the case of GWAS, the sample genomes interrogated are derived from normal lymphocyte gDNA, and therefore reflect a more predictable and stable reference distribution, amenable to standard correction procedures such as Bonferroni. Gene expression distributions in somatic cancer genomes are similar to those observed in copy number in the sense that drivers and passengers of varying degrees of cancer causation result in up to 50% of the genes being significantly differentially expressed. To generate a manageable subset to integrate with copy number and GWAS, we apply a Bonferroni correction on the distribution of p-values as estimated by the Student T resulting in 145 significant genes in AA vs. CA data set (unpublished) and 260 genes from the tumor vs. matched normal data set ($p < 0.05$).

To further put into context the identified events, we combined all data types using the function E(L) described in chapter 2.6. The significant set of genes as derived from copy number, gene expression and GWAS data types (980 genes) were queried for primary protein-protein interactions in the Biogrid Database[160] which resulted in a highly significant network connectivity Z(C(N)) = 83sdu (as described in chapter 2.6). Following this, the 163 interconnected genes (those with interactions to at least one of the 980 seed genes) were

tested for enrichment against the hypergeometric distribution [using the R function dhyper()] of a MSig database made up of functional lists of genes and pathways. The genes overlapping the top ranking gene sets (both the Biogrid network interactions and the genes overlapping the top gene sets) are presented in Figures 20 and 21. The top ranking gene sets (Table 6) were clustered based on gene overlap, resulting in three significant ontology enrichment groups -- cell proliferation, focal adhesion, and cytoskeleton. Given that the candidate genes used to identify these functional interactions were selected through copy number, gene expression and GWAS analyses geared to uncover the aggressive nature of prostate cancer, the gene interactions and pathways involving cell proliferation, adhesion and cytoskeleton likely encompass the racial disparity observed between AAs and CAs, along with the selective path induced in the METS through androgen ablation therapy. Knowing the direction or selection propensity associated with each gene, as inferred by the ISM, will allow the next set of hypothesis to be tested through data mining of the published literature and encourage experimentation with these pathways in the context of epistatic or multi-genic responses. Robust experimental validation of metastatic potential will provide a basis for proposing multiple-synergistic interventions that could be readily to translated into clinical diagnostics and personalized therapy.

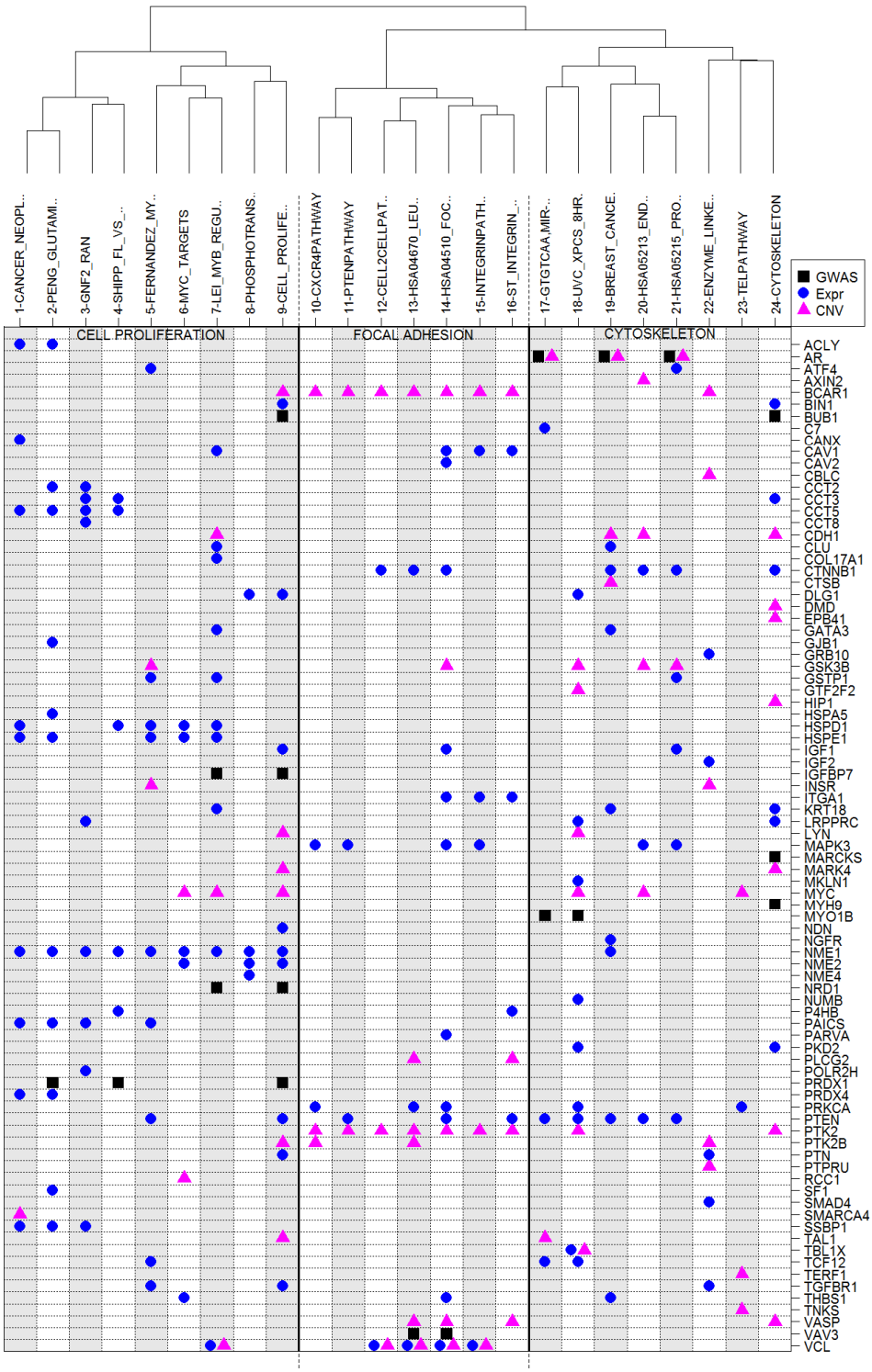

Figure 20 Gene set enrichment clusters identified from copy number, gene expression and GWAS-AA, CA and METS prostate cancer data sets.

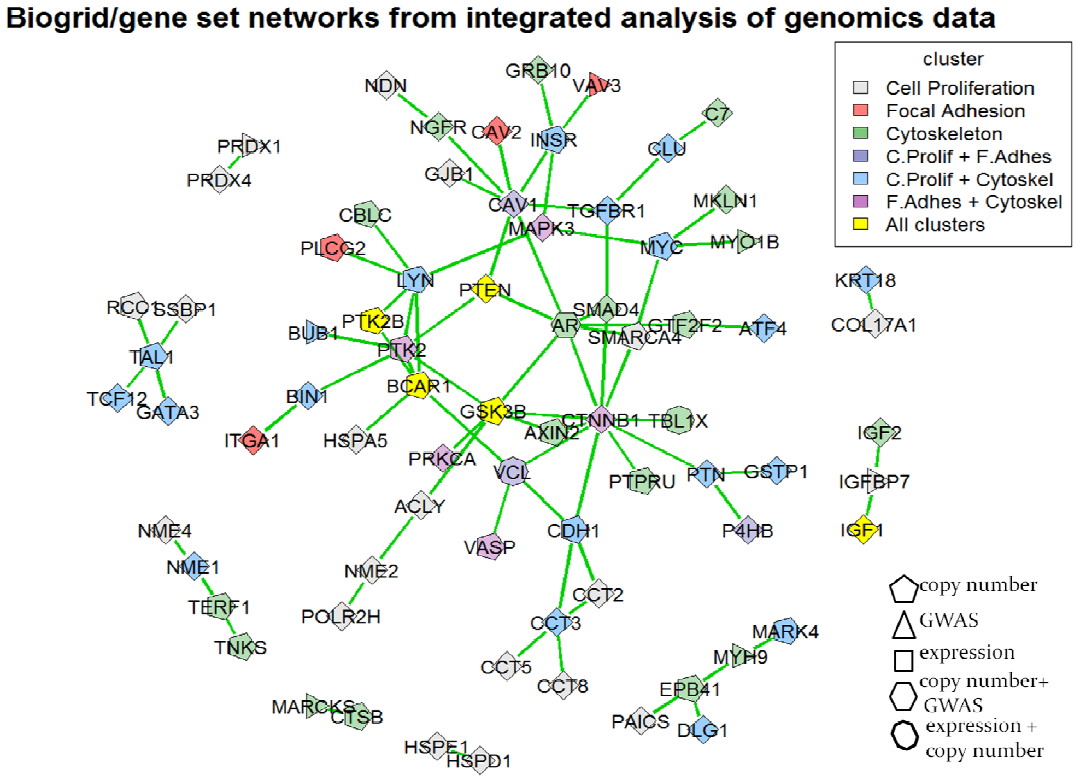

Figure 21: Biogrid protein-protein interaction network with gene set enrichment clusters. Shown in the network are 75 significant genes identified through analysis of data from prostate cancer copy number, gene expression and GWAS datasets.

Table 6: Gene sets identified using integrated genomics data

Gene set	rank	cluster	Overlap	Gene set size	Overlap	P value	Z score	Overlapping genes	PMID	Description
CANCER NEOPLASTIC META UP	2	Cell Proliferation	10	64	0.16	1.28E-13	22.76	SMARCA4; SSBP1; HSPD1; NME1; ACLY; CANX; PRDX4; CCT5; PAICS; HSPE1	15184677	Sixty-seven genes commonly upregulated in cancer relative to normal tissue, from a meta-analysis of the OncoMine gene expression database
GNF2 RAN	7	Cell Proliferation	9	86	0.10	9.12E-11	17.67	LRPPRC; SSBP1; NME1; POLR2H; CCT3; CCT5; PAICS; CCT2; CCT8		Neighborhood of RAN
MYC TARGETS	8	Cell Proliferation	7	42	0.17	4.09E-10	16.97	MYC; RCC1; THBS1; HSPD1; NME1; NME2; HSPE1	14519204	Myc-responsive genes reported in multiple systems
SHIPP FL VS DLBCL DN	10	Cell Proliferation	6	37	0.16	9.04E-09	15.40	HSPD1; NME1; CCT3; CCT5; P4HB; PRDX1	11786909	Genes upregulated in diffuse B-cell lymphomas (DLBCL) and downregulated in follicular lymphoma (FL) (fold change of at least 3)
FERNANDEZ MYC TARGETS	21	Cell Proliferation	11	180	0.06	2.47E-10	13.99	INSR; GSK3B; PTEN; TGFBR1; HSPD1; NME1; GSTP1; PAICS; TCF12; HSPE1; ATF4	12695333	MYC target genes by ChIP in U-937,HL60 (leukemia),P493 (B-cell),T98G (glioblastoma),WS1 (fibroblast)
PENG GLUTAMINE DN	31	Cell Proliferation	12	313	0.04	7.25E-09	13.04	SF1; SSBP1; NME1; ACLY; PRDX4; CCT5; PAICS; GJB1; CCT2; HSPE1; HSPA5; PRDX1	12101249	Genes downregulated in response to glutamine starvation
PHOSPHOTRANSFERASE ACTIVITY PHOSPHATE GROUP AS ACCEPTOR	32	Cell Proliferation	4	18	0.22	7.96E-07	12.83	NME4; DLG1; NME1; NME2		Genes annotated by the GO term GO:0016776. Catalysis of the transfer of a phosphorus-containing group from one compound (donor) to a phosphate group (acceptor).
CELL PROLIFERATION GO 0008283	33	Cell Proliferation	19	514	0.04	4.80E-13	12.82	MYC; BCAR1; PTK2B; MARK4; LYN; TAL1; PTEN; DLG1; TGFBR1; NME1; NME2; IGF1; BIN1; NDN; PTN; NRD1; PRDX1; IGFBP7; BUB1		Genes annotated by the GO term GO:0008283. The multiplication or reproduction of cells, resulting in the expansion of a cell population.
LEI MYB REGULATED GENES	35	Cell Proliferation	14	325	0.04	8.67E-11	12.70	MYC; CDH1; VCL; HSPD1; NME1; CAV1; KRT18; GSTP1; HSPE1; GATA3; COL17A1; CLU; NRD1; IGFBP7	15105423	Myb-regulated genes
BREAST CANCER ESTROGEN SIGNALING	3	Cytoskeleton	11	101	0.11	4.51E-13	20.92	AR; CDH1; CTSB; PTEN; CTNNB1; THBS1; NME1; KRT18; NGFR; GATA3; CLU		Genes preferentially expressed in breast cancers, especially those involved in estrogen-receptor-dependent signal transduction.
HSA05213 ENDOMETRIAL CANCER	14	Cytoskeleton	7	52	0.13	1.95E-09	14.80	MYC; CDH1; GSK3B; AXIN2; PTEN; CTNNB1; MAPK3		Genes involved in endometrial cancer
ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY	22	Cytoskeleton	10	140	0.07	3.61E-10	13.85	BCAR1; PTK2B; CBLC; INSR; PTPRU; GRB10; SMAD4; TGFBR1; IGF2; PTN		Genes annotated by the GO term GO:0007167. Any series of molecular signals initiated by the binding of an extracellular ligand to a receptor on the surface of the target cell, where the receptor possesses catalytic activity or is closely associated with a
TELPATHWAY	27	Cytoskeleton	4	18	0.22	7.96E-07	13.55	TERF1; MYC; TNKS; PRKCA		Telomerase is a ribonucleotide protein that adds telomeric repeats to the 3' ends of chromosomes.
CYTOSKELETON	28	Cytoskeleton	16	368	0.04	3.25E-12	13.49	DMD; PTK2; VASP; CDH1; MARK4; HIP1; EPB41; LRPPRC; CTNNB1; CCT3; BIN1; KRT18; PKD2; MARCKS; MYH9; BUB1		Genes annotated by the GO term GO:0005856. Any of the various filamentous elements that form the internal framework of cells, and typically remain after treatment of the cells with mild detergent to remove membrane constituents and soluble components of t
HSA05215 PROSTATE CANCER	29	Cytoskeleton	8	87	0.09	2.94E-09	13.42	AR; GSK3B; PTEN; CTNNB1; IGF1; MAPK3; GSTP1; ATF4		Genes involved in prostate cancer
GTGTCAA,MIR-514	37	Cytoskeleton	6	61	0.10	1.96E-07	12.24	AR; TAL1; PTEN; C7; TCF12; MYO1B		Targets of MicroRNA GTGTCAA,MIR-514
UVC XPCS 8HR DN	39	Cytoskeleton	15	408	0.04	1.62E-10	12.10	TBL1X; MYC; PTK2; GTF2F2; LYN; GSK3B; PTEN; LRPPRC; NUMB; DLG1; MKLN1; PRKCA; TCF12; PKD2; MYO1B	15608684	Down-regulated at 8 hours following treatment of XPB/CS fibroblasts with 3 J/m^2 UVC
HSA04510 FOCAL ADHESION	1	Focal Adhesion	16	200	0.08	2.65E-16	22.86	PTK2; VASP; BCAR1; VCL; GSK3B; PTEN; CTNNB1; THBS1; PARVA; PRKCA; CAV2; IGF1; CAV1; ITGA1; MAPK3; VAV3		Genes involved in focal adhesion
INTEGRINPATHWAY	12	Focal Adhesion	6	38	0.16	1.07E-08	14.95	PTK2; BCAR1; VCL; CAV1; ITGA1; MAPK3		Integrins are cell surface receptors commonly present at focal adhensions that interact with the extracellular matrix and transduce extracellular signaling.
CELL2CELLPATHWAY	13	Focal Adhesion	4	13	0.31	1.90E-07	14.90	PTK2; BCAR1; VCL; CTNNB1		Epithelial cell adhesion proteins such as cadherins transduce signals into the cell via catenins, which alter cell shape and motility.
CXCR4PATHWAY	16	Focal Adhesion	5	24	0.21	4.35E-08	14.30	PTK2; BCAR1; PTK2B; PRKCA; MAPK3		CXCR4 is a G-protein coupled receptor that responds to the ligand SDF-1 by activating Ras and PI3 kinase to promote lymphocyte chemotaxis.
ST INTEGRIN SIGNALING PATHWAY	17	Focal Adhesion	8	82	0.10	1.83E-09	14.28	PTK2; VASP; BCAR1; PLCG2; PTEN; CAV1; ITGA1; P4HB		Integrins are transmembrane receptors that mediate cell growth, survival, and migration by binding to ligands in the extracellular matrix.
HSA04670 LEUKOCYTE TRANSENDOTHELIAL MIGRATION	18	Focal Adhesion	9	115	0.08	1.25E-09	14.14	PTK2; VASP; BCAR1; PTK2B; VCL; PLCG2; CTNNB1; PRKCA; VAV3		Genes involved in Leukocyte transendothelial migration
PTENPATHWAY	36	Focal Adhesion	4	18	0.22	7.96E-07	12.69	PTK2; BCAR1; PTEN; MAPK3		PTEN suppresses AKT-induced cell proliferation and antagonizes the action of PI3K.

CHAPTER 4

Discussion

Cancer diagnosis is a life-altering experience. The instant association with death puts a severe emotional strain on the diagnosed individual and loved ones caring for him or her. This experience is compounded by the toxicity associated with most treatments and, of course, by the uncertainty of outcome. Current diagnostic methods for prostate cancer can predict whether carcinoma or localized growth is present and can place crude estimates on the natural history of disease. The CAPRA score is the most sophisticated series of clinical markers currently available in prostate cancer diagnostics.[32] Patients scored at the lowest level of the CAPRA scale (reflecting a diagnosis of early-stage cancer) ultimately had metastatic events 3% of the time. Patients with high CAPRA scores (reflecting late-stage carcinoma) ultimately contracted metastatic events 21% of the time. The nature of this range of certainty means that such a metric is of little utility to either physicians or patients when deciding upon a particular treatment. These outcome statistics therefore indicate that even late-stage carcinoma has a 79% chance of not metastasizing and the clues for metastatic potential are not apparent in the CAPRA series of markers. Thus clinical decisions will usually be driven instead by the age of the patient.

In the United Kingdom, there is a general consensus that the best response to diagnosis of primary carcinoma is watchful waiting – this allows for patients to go untreated until symptoms emerge and palliative care is

administered. Alternatively, the active surveillance approach calls for routine monitoring of PSA levels and biopsies for signals of progression.

This study's aims were to identify a series of nucleic acid based markers that could be used to 1) develop a diagnostic with strong predictive power, 2) reveal the mechanisms behind metastatic potential and 3) offer candidates for prophylactic and disease treatment. To accomplish this, a strategy based on integrating different populations of samples and orthologous data types was employed (Figure 22).

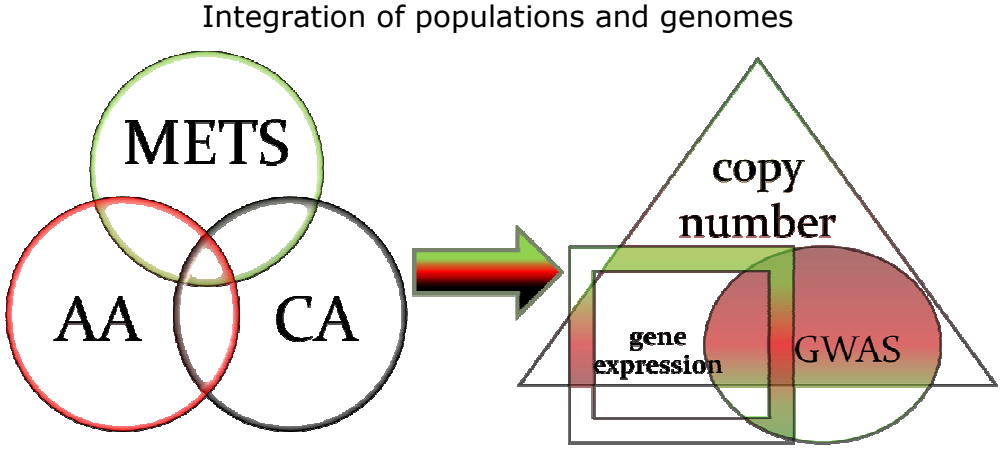

Figure 22 African American (AA) primary tumors, Caucasian American (CA) primary tumors and Caucasian American metasasis (METS) were used to study the genomics of prostate cancer. Three sample populations were applied to three different genomic profiling technologies analyzed as an integrated data set and evaluated for fuctional enrichment.

First, we generated a genomic copy number data set comprised of 9 AA and 20 CA pairs of matched primary cancer and normals. This data-set was created to assess whether the health-disparity between AAs and CAs can be captured through the amplification and deletion events typically observed in prostate carcinoma. Next, 2 published genomic copy number data sets were

included, adding an additional 20 AA pairs of primary cancer and matched control[128] along with 52 METS samples from 12 multiple-metastasis CA patients collected at autopsy following death from prostate cancer).[16] With these three data sets we harnessed the racial disparity feature of prostate cancer by comparing AAs to CAs in the context of the METS samples, because metastatic patients are a reference representing a *bona fide* poor outcome. One important factor to note for this three way comparison was that the METS come from twelve Caucasian American men.

Our initial analysis resulted in a fascinating observation, confirming that health disparity is inherent to the somatic copy number genomes of AAs and CAs. Through unsupervised clustering, we showed that AAs cluster preferentially with METS at an enrichment (EA) = 2.76, consistent with the greater than 2-fold increased incidence of death reported by the American Cancer Society's 2008, Cancer Facts and Figures.[1]

To delve further into the specific genomic regions influencing this enrichment and potentially harboring the genes responsible for the observed increase in death rate, we developed the ISM. This model utilized the principles of evolutionary selection to infer whether the distribution of signals from AAs, CAs and METS fit a scenario consistent with positive cellular growth and metastasis, or one that imposed negative forces on the cell. In this model, a score was calculated on a probe basis. The ISM score combined both directions of selection force into a single function, resulting in high positive values when METS and AAs were preferentially amplified whereas CAs are deleted and in high negative values when there were amplifications in CAs and preferential deletions or in METS and AAs. The ISM(rd) score in the context of positive selection is

made up of three terms that have 3 user defined weight constants (a,b,c) respectively (a weighs enrichment for METS amplifications; b weighs enrichment for AA amplifications; c weighs enrichment for CA deletions) placing emphasis on different aspect(s) of the selective criteria that drive the analysis. For the ISM scores used in this investigation for both the rd and us components, we set *a=3, b=1.5, and c=3* resulting in more emphasis placed on METS amplification enrichment and CA|iPT deletion enrichment over AA|mPT amplification enrichment. This scenario rewards for greater separation between METS and CA|iPT events while it is more lenient on the direction of AA|mPT copy number. An optional fourth penalty term was added to the ISM four different times in order to model different "flavors" of evolutionary selection.

The probes representing significant model scores were associated with genes and integrated with gene expression and GWAS significant genes. At the level of individual candidate genes, our models proved reliable in predicting whether genes would be found to be over or under expressed and whether they would have a positive or negative influence on promoting metastatic disease. Both our top ranked PSM (OPHN1) and NSM (VASP) scoring genes were shown be associated with metastatic cancer of the $breast_{VASP}$, $colon_{OPHN1}$, $gastric_{OPHN1}$ and $glial_{OPHN1}$ and reported to be involved in the processes of cell-matrix adhesion and membrane trafficking.[165,167] Consistent with their SM classifications, OPHN1 and VASP are over expressed and under expressed respectively in the context of metastatic potential. While VASP has clear racial disparity characteristics, as mentioned earlier, metastatic potential in the case of OPHN1 is most likely a function of androgen ablation therapy in METS patients and not due to racial disparity. Within 100kb away from VASP, the ERCC1 locus

exhibits a similar pattern of copy number events, therefore, also reflecting NSM characteristics. This is consistent for what is know about this endonuclease component of the nucleotide excision repair complex, functioning in the repair of double strand breaks. Presumably, deletions in this locus would result in increase mutation rates, whereas amplifications may have be protective against mutation. For both VASP and ERCC1, METS are preferentially deleted with no amplifications, AAs are neutral and CAs have several amplifications with no deletions.

The 4th ranked PSM gene, PREX2a, has been shown to antagonize the putative tumor suppressor gene PTEN[174] through its actions on PI3K, resulting in the accumulation of the downstream messenger PIP3 and activation of the AKT pathway. Uninhibited AKT leads to uncontrolled cellular growth and proliferation. The distribution of copy number signal among the subgroups was consistent with PSM revealing greater enrichment for amplifications in METS and AAs over CAs.

Next, through differential gene expression analysis of AA vs. CA primary tumors, we identified the putative prostate cancer gene PTEN as being ranked 3rd and significantly elevated in AAs over CAs (rank = 3; p = 8.8 x 10^-6, Bonferroni corrected). Considering that PTEN is a tumor suppressor gene, looking at only the gene expression, this result went against the expectation that AAs would exhibit a greater propensity for deletion. However, when we explored the copy number ρ'' distributions (Figure 17), we observed that CAs were indeed preferentially deleted (in a different set of samples from those used for gene expression) over AAs, conforming with the gene expression result. The best PTEN ranking #6728 for PSM, which proved unreliable for determining whether

the locus is NS or PS leaning. However, a racial disparity does appear to be acting in the NS direction, exhibited by only CAs having amplifications, a potential protective effect for the CA subgroup. With this candidate both the gene expression results and PS score were misdirected and uninformative, respectively, but, a careful inspection of the ρ'' provided justification for a NS direction consistent with PTEN putative tumor suppressive activity.

Another compelling candidate gene, NME4, identified through gene expression analysis in the CA vs. AA data set was decreased in AA primary tumor subgroups as opposed to CA subgroups (rank = 17; p = 4.72E-04, Bonferroni corrected). NME4, is part of the nm23 nucleoside diphosphate kinase family and has an amino terminal domain that after degradation, activates the protein and targets it to mitochondria where it forms a hexameric structure spanning the inner-to-outer membrane junction.[177] NME4 was reported to be over expressed in primary colon and renal tumors,[184] whereas NME1 and NME2 were over expressed in solid tumors and decreased in metastasis of melanoma, breast, liver, ovary and colon.[178] In our analysis of the tumor versus paired normal gene expression data-set[152], NME1 and NME2 were significantly over expressed (NME1, rank, p = 9.4E-07; NME2, rank = , p = 2.49E-04; Bonferroni corrected). Accordingly, the copy number profile of ($\rho''_{amp|del}$) at the NME4 locus showed preferential deletions in AA primary tumors and preferential amplifications in CA primary tumors, whereas METS from Caucasian individuals exhibited a ρ'' of neutral or no events. As with PTEN, the selection score for NME4 yielded an uninformative NSM rank = 5653; however, the distribution of signals matched expectations for negative selection. The composite of

information from primary somatic tumor gene expression, copy number and data mining of functional information, all in the context of racial disparity and metastatic potential, clearly shows that PTEN and NME4 have negative selection properties, while having dual effects in carcinoma. PTEN is preferentially deleted in carcinoma and METS, where its inhibitory control over the PI3K-AKT pathway is relinquished. Alternatively, it may be amplified in carcinoma, offering a protective effect, reducing metastatic potential by inhibiting the actions of AR and other metastasis-promoting genes, such as cell division cycle 6 (CDC6) and cyclin E2 (CCNE2).[185] The mechanism of protection could be a dosage effect or a structural reorganization of the genome, yielding a novel isoform with superior activity against metastasis. Clinically, PTEN deletions along with TMPRSS2-ERG fusion products have been associated with a poor outcome.[186] NME4 has a different presentation of NS, where over expression is observed in the primary carcinoma, most likely as a negative regulatory response to carcinogenesis. This response is seen in both AA and CA primaries versus normal tissue, and CAs as a group show over expression in mRNA and copy number amplification relative to AAs. Copy number events were not observed in METS, whereas 3 out of 29 AAs showed deletions. Since active NME4 has been observed to be localized to the mitochondrial membrane, it would be of interest to see if its NME4 activity has a direct effect on VASP which also co-localizes to the mitochondrial membrane.

Next, from the PLCO-CGEMS GWAS[153] data set an interesting candidate, AR, was associated with prostate cancer and ranked as the 7th most significant gene. This study reveals genetic associations from prostate cancer cases of both aggressive (n=688) and non-aggressive (n=484) outcomes. AR has been heavily studied in the context of prostate carcinoma progression through

androgen independent metastatic disease.[179] However, a strong association may reflect predisposing ancestral structures that may influence either or both stages of disease. At the copy number it is not clear whether AR contributes to the racial disparity between AAs and CAs. Although it ranked #34 for positive selection, the score was driven by the extreme amplification in METS, presumably as a result of androgen ablation therapy. AA and CA primary tumors have a low frequency of events, however, it is uncertain whether a more high-resolution view would change the landscape. It is appears that androgen independent prostate tumor growth is a function of the androgen ablation therapy that induces an evolutionary state of selection and survival not present in primary tumors of either race.

Understanding the functions of genes in the context of pathways is a critical step in developing concerted schemes to influence the vitality of a cell. In the case of cancer progression and metastasis, prior information about the direction in which pathway components regulate the cell to proliferate and mobilize can aid in the experimental design of synergistic and epistatic assays. For example, VASP and NME4 exhibit NSM characteristics that are clearly based on the racial disparity, both proteins localize to the mitochondrial membrane to perform their functional duties and both have been shown to be under expressed in functional assays studying metastatic potential. This series of observations, based on copy number, gene expression and data mining could be used to experimentally test new hypotheses for prostate cancer metastatic progression.

To get a more global perspective of how a subset of the most significantly ranked genes segregate in terms of function, we analyzed a set of 980 genes through network connectivity of putative protein-protein interactions followed by

gene set enrichment analysis. Out of a total of 980 genes that were applied to the network analysis, 163 emerged with at least 1 primary interaction with another seed gene. The 163 seed genes were analyzed for enrichment against the MSigDB gene sets and shown to be enriched for 3 major pathways, cell proliferation, focal adhesion and cytoskeleton. Interestingly, a report comparing androgen dependent vs. androgen independent primary tumors identified cell adhesion as a significantly enriched ontology.[187]

Future work will involve experimental validation of the candidate genes identified in this study through functional assays and deep sequencing through of amplified and deleted regions to get a more resolved understanding of the structure of the altered genomes. Automated data mining procedures of literature and public data repositories would allow for rapid validation of hypotheses generated through novel primary data sets. Most importantly, the use of the selection models to predict the risk associated with having a particular distribution of positive versus negative events, allowing clinicians to make more informed decisions for treating patients and drug companies to add candidates to their pipeline.

Bibliography

1. Society, A.C. Cancer facts & figures. (ed. 2008CAFFfinalsecured.pdf) (Atlanta, 2008).
2. Page, W.F., Braun, M.M., Partin, A.W., Caporaso, N. & Walsh, P. Heredity and prostate cancer: a study of World War II veteran twins. *Prostate* **33**, 240-5. (1997).
3. Lichtenstein, P. et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85. (2000).
4. Carter, B.S., Beaty, T.H., Steinberg, G.D., Childs, B. & Walsh, P.C. Mendelian inheritance of familial prostate cancer. *Proc Natl Acad Sci U S A* **89**, 3367-71 (1992).
5. Keetch, D.W., Rice, J.P., Suarez, B.K. & Catalona, W.J. Familial aspects of prostate cancer: a case control study. *J Urol* **154**, 2100-2 (1995).
6. McLellan, D.L. & Norman, R.W. Hereditary aspects of prostate cancer. *Cmaj* **153**, 895-900 (1995).
7. Zeegers, M.P., Jellema, A. & Ostrer, H. Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma: a meta-analysis. *Cancer* **97**, 1894-903 (2003).
8. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-3 (1971).
9. Cheng, I. et al. Socioeconomic status and prostate cancer incidence and mortality rates among the diverse population of California. *Cancer Causes Control* (2009).
10. Freedland, S.J. & Isaacs, W.B. Explaining racial differences in prostate cancer in the United States: sociology or biology? *Prostate* **62**, 243-52 (2005).
11. Roddam, A.W., Allen, N.E., Appleby, P. & Key, T.J. Endogenous sex hormones and prostate cancer: a collaborative analysis of 18 prospective studies. *J Natl Cancer Inst* **100**, 170-83 (2008).
12. Tsai, C.J. et al. Sex steroid hormones in young manhood and the risk of subsequent prostate cancer: a longitudinal study in African-Americans and Caucasians (United States). *Cancer Causes Control* **17**, 1237-44 (2006).
13. Weiss, J.M. et al. Endogenous sex hormones and the risk of prostate cancer: a prospective study. *Int J Cancer* **122**, 2345-50 (2008).
14. Sakr, W.A., Haas, G.P., Cassin, B.F., Pontes, J.E. & Crissman, J.D. The frequency of carcinoma and intraepithelial neoplasia of the prostate in young male patients. *J Urol* **150**, 379-85 (1993).
15. Abate-Shen, C. & Shen, M.M. Molecular genetics of prostate cancer. *Genes Dev* **14**, 2410-34 (2000).
16. Liu, W. et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med* **15**, 559-65 (2009).
17. Sun, J. et al. DNA copy number alterations in prostate cancers: a combined analysis of published CGH studies. *Prostate* **67**, 692-700 (2007).

18. Knudson, A.G., Jr. Hereditary cancer, oncogenes, and antioncogenes. *Cancer Res* **45**, 1437-43 (1985).
19. Collins, N. et al. Consistent loss of the wild type allele in breast cancers from a family linked to the BRCA2 gene on chromosome 13q12-13. *Oncogene* **10**, 1673-5 (1995).
20. Maher, C.A. et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* (2009).
21. Mosquera, J.M. et al. Prevalence of TMPRSS2-ERG fusion prostate cancer among men undergoing prostate biopsy in the United States. *Clin Cancer Res* **15**, 4706-11 (2009).
22. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).
23. Wang, M.C., Valenzuela, L.A., Murphy, G.P. & Chu, T.M. Purification of a human prostate specific antigen. *Invest Urol* **17**, 159-63 (1979).
24. Lilja, H. A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. *J Clin Invest* **76**, 1899-903 (1985).
25. Cathelineau, X. P.W. Kantoff, P.R. Carroll, A.V. D'Amico (eds). Prostate Cancer: Principles and Practice 10.1093/annonc/mdg276. *Ann Oncol* **14**, 1157- (2003).
26. Smith, D.S., Bullock, A.D., Catalona, W.J. & Herschman, J.D. Racial differences in a prostate cancer screening study. *J Urol* **156**, 1366-9 (1996).
27. Andriole, G.L. et al. Mortality results from a randomized prostate-cancer screening trial. *N Engl J Med* **360**, 1310-9 (2009).
28. Fang, J., Metter, E.J., Landis, P. & Carter, H.B. PSA velocity for assessing prostate cancer risk in men with PSA levels between 2.0 and 4.0 ng/ml. *Urology* **59**, 889-93; discussion 893-4 (2002).
29. Vickers, A.J., Savage, C., O'Brien, M.F. & Lilja, H. Systematic review of pretreatment prostate-specific antigen velocity and doubling time as predictors for prostate cancer. *J Clin Oncol* **27**, 398-403 (2009).
30. Humphrey, P.A. Gleason grading and prognostic factors in carcinoma of the prostate. *Mod Pathol* **17**, 292-306 (2004).
31. Iczkowski, K.A. Current prostate biopsy interpretation: criteria for cancer, atypical small acinar proliferation, high-grade prostatic intraepithelial neoplasia, and use of immunostains. *Arch Pathol Lab Med* **130**, 835-43 (2006).
32. Cooperberg, M.R. et al. The University of California, San Francisco Cancer of the Prostate Risk Assessment score: a straightforward and reliable preoperative predictor of disease recurrence after radical prostatectomy. *J Urol* **173**, 1938-42 (2005).
33. Cooperberg, M.R., Broering, J.M. & Carroll, P.R. Risk assessment for prostate cancer metastasis and mortality at the time of diagnosis. *J Natl Cancer Inst* **101**, 878-87 (2009).
34. Clark, J. et al. Genome-wide screening for complete genetic loss in prostate cancer by comparative hybridization onto cDNA microarrays. *Oncogene* **22**, 1247-52 (2003).
35. Bussemakers, M.J. et al. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res* **59**, 5975-9 (1999).

36. Mearini, E. et al. The combination of urine DD3PCA3 mRNA and PSA mRNA as molecular markers of prostate cancer. *Biomarkers* **14**, 235-43 (2009).
37. Berthold, D.R., Sternberg, C.N. & Tannock, I.F. Management of advanced prostate cancer after first-line chemotherapy. *J Clin Oncol* **23**, 8247-52 (2005).
38. Shamash, J. et al. Chlorambucil and lomustine (CL56) in absolute hormone refractory prostate cancer: re-induction of endocrine sensitivity an unexpected finding. *Br J Cancer* **92**, 36-40 (2005).
39. Shaw, G. & Prowse, D.M. Inhibition of androgen-independent prostate cancer cell growth is enhanced by combination therapy targeting Hedgehog and ErbB signalling. *Cancer Cell Int* **8**, 3 (2008).
40. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).
41. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).
42. Lucito, R. et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* **13**, 2291-305 (2003).
43. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).
44. Sharp, A.J. et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* (2006).
45. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).
46. Bailey, J.A., Liu, G. & Eichler, E.E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73**, 823-34 (2003).
47. Zhou, Y. & Mishra, B. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* **102**, 4051-6 (2005).
48. Lupski, J.R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**, e49 (2005).
49. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* **3**, 199-242 (2002).
50. Antonell, A., de Luis, O., Domingo-Roura, X. & Perez-Jurado, L.A. Evolutionary mechanisms shaping the genomic structure of the Williams-Beuren syndrome chromosomal region at human 7q11.23. *Genome Res* **15**, 1179-88 (2005).
51. Ciccarelli, F.D. et al. Complex genomic rearrangements lead to novel primate gene function. *Genome Res* **15**, 343-51 (2005).
52. Jin, H. et al. Structural evolution of the BRCA1 genomic region in primates. *Genomics* **84**, 1071-82 (2004).
53. Eichler, E.E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res* **9**, 1048-58 (1999).

54. Babcock, M. et al. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res* **13**, 2519-32 (2003).
55. Kato, T. et al. Genetic variation affects de novo translocation frequency. *Science* **311**, 971 (2006).
56. Gotter, A.L., Shaikh, T.H., Budarf, M.L., Rhodes, C.H. & Emanuel, B.S. A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2. *Hum Mol Genet* **13**, 103-15 (2004).
57. Nimmakayalu, M.A., Gotter, A.L., Shaikh, T.H. & Emanuel, B.S. A novel sequence-based approach to localize translocation breakpoints identifies the molecular basis of a t(4;22). *Hum Mol Genet* **12**, 2817-25 (2003).
58. Kurahashi, H., Shaikh, T., Takata, M., Toda, T. & Emanuel, B.S. The constitutional t(17;22): another translocation mediated by palindromic AT-rich repeats. *Am J Hum Genet* **72**, 733-8 (2003).
59. Liu, W. et al. Comprehensive assessment of DNA copy number alterations in human prostate cancers using Affymetrix 100K SNP mapping array. *Genes Chromosomes Cancer* (2006).
60. Beroukhim, R. et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput Biol* **2**, e41 (2006).
61. Brookman-Amissah, N. et al. Genome-wide screening for genetic changes in a matched pair of benign and prostate cancer cell lines using array CGH. *Prostate Cancer Prostatic Dis* **8**, 335-43 (2005).
62. Chaudhary, J. & Schmidt, M. The impact of genomic alterations on the transcriptome: a prostate cancer cell line case study. *Chromosome Res* **14**, 567-86 (2006).
63. Paris, P.L. et al. High-resolution analysis of paraffin-embedded and formalin-fixed prostate tumors using comparative genomic hybridization to genomic microarrays. *Am J Pathol* **162**, 763-70 (2003).
64. Paris, P.L. et al. Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Hum Mol Genet* **13**, 1303-13 (2004).
65. Saramaki, O.R., Porkka, K.P., Vessella, R.L. & Visakorpi, T. Genetic aberrations in prostate cancer by microarray analysis. *Int J Cancer* (2006).
66. Verhagen, P.C. et al. The PTEN gene in locally progressive prostate cancer is preferentially inactivated by bi-allelic gene deletion. *J Pathol* **208**, 699-707 (2006).
67. Watson, J.E. et al. Integration of high-resolution array comparative genomic hybridization analysis of chromosome 16q with expression array data refines common regions of loss at 16q23-qter and identifies underlying candidate tumor suppressor genes in prostate cancer. *Oncogene* **23**, 3487-94 (2004).
68. Yao, J. et al. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res* **66**, 4065-78 (2006).
69. Bergamaschi, A. et al. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-

expression subtypes of breast cancer. *Genes Chromosomes Cancer* (2006).
70. Fridlyand, J. et al. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* **6**, 96 (2006).
71. Han, W. et al. Genomic alterations identified by array comparative genomic hybridization as prognostic markers in tamoxifen-treated estrogen receptor-positive breast cancer. *BMC Cancer* **6**, 92 (2006).
72. Mastracci, T.L. et al. Genomic alterations in lobular neoplasia: A microarray comparative genomic hybridization signature for early neoplastic proliferationin the breast. *Genes Chromosomes Cancer* (2006).
73. Naylor, T.L. et al. High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res* **7**, R1186-98 (2005).
74. Rha, S.Y. et al. Alteration of hTERT full-length variant expression level showed different gene expression profiles and genomic copy number changes in breast cancer. *Oncol Rep* **15**, 749-55 (2006).
75. Shadeo, A. & Lam, W.L. Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Res* **8**, R9 (2006).
76. van Beers, E.H. & Nederlof, P.M. Array-CGH and breast cancer. *Breast Cancer Res* **8**, 210 (2006).
77. Zhao, X. et al. Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res* **65**, 5561-70 (2005).
78. Nymark, P. et al. Identification of specific gene copy number changes in asbestos-related lung cancer. *Cancer Res* **66**, 5737-43 (2006).
79. Tonon, G. et al. Common and contrasting genomic profiles among the major human lung cancer subtypes. *Cold Spring Harb Symp Quant Biol* **70**, 11-24 (2005).
80. Tonon, G. et al. High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A* **102**, 9625-30 (2005).
81. Garraway, L.A. & Sellers, W.R. Lineage dependency and lineage-survival oncogenes in human cancer. *Nat Rev Cancer* **6**, 593-602 (2006).
82. Hughes, S. et al. Microarray comparative genomic hybridisation analysis of intraocular uveal melanomas identifies distinctive imbalances associated with loss of chromosome 3. *Br J Cancer* **93**, 1191-6 (2005).
83. Kim, M. et al. Comparative oncogenomics identifies NEDD9 as a melanoma metastasis gene. *Cell* **125**, 1269-81 (2006).
84. Bauer, J. & Bastian, B.C. Distinguishing melanocytic nevi from melanoma by DNA copy number changes: comparative genomic hybridization as a research and diagnostic tool. *Dermatol Ther* **19**, 40-9 (2006).
85. Hashimoto, K. et al. Analysis of DNA copy number aberrations in hepatitis C virus-associated hepatocellular carcinomas by conventional CGH and array CGH. *Mod Pathol* **17**, 617-22 (2004).
86. Kawaguchi, K., Honda, M., Yamashita, T., Shirota, Y. & Kaneko, S. Differential gene alteration among hepatoma cell lines demonstrated by cDNA microarray-based comparative genomic hybridization. *Biochem Biophys Res Commun* **329**, 370-80 (2005).

87. Patil, M.A. et al. Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma. *Carcinogenesis* **26**, 2050-7 (2005).
88. Tanami, H. et al. Involvement of cyclin D3 in liver metastasis of colorectal cancer, revealed by genome-wide copy-number analysis. *Lab Invest* **85**, 1118-29 (2005).
89. Zender, L. et al. Identification and validation of oncogenes in liver cancer using an integrative oncogenomic approach. *Cell* **125**, 1253-67 (2006).
90. Holzmann, K. et al. Genomic DNA-chip hybridization reveals a higher incidence of genomic amplifications in pancreatic cancer than conventional comparative genomic hybridization and leads to the identification of novel candidate genes. *Cancer Res* **64**, 4428-33 (2004).
91. Bashyam, M.D. et al. Array-based comparative genomic hybridization identifies localized DNA amplifications and homozygous deletions in pancreatic cancer. *Neoplasia* **7**, 556-62 (2005).
92. Gysin, S., Rickert, P., Kastury, K. & McMahon, M. Analysis of genomic DNA alterations and mRNA expression patterns in a panel of human pancreatic cancer cell lines. *Genes Chromosomes Cancer* **44**, 37-51 (2005).
93. Mahlamaki, E.H. et al. High-resolution genomic and expression profiling reveals 105 putative amplification target genes in pancreatic cancer. *Neoplasia* **6**, 432-9 (2004).
94. Aguirre, A.J. et al. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A* **101**, 9067-72 (2004).
95. Alrawi, S.J. et al. Genomic instability of human aberrant crypt foci measured by inter-(simple sequence repeat) PCR and array-CGH. *Mutat Res* (2006).
96. Camps, J. et al. Genome-wide differences between microsatellite stable and unstable colorectal tumors. *Carcinogenesis* **27**, 419-28 (2006).
97. Cardoso, J. et al. Genomic profiling by DNA amplification of laser capture microdissected tissues and array CGH. *Nucleic Acids Res* **32**, e146 (2004).
98. Swede, H. et al. Genomic profiles of colorectal cancers differ based on patient smoking status. *Cancer Genet Cytogenet* **168**, 98-104 (2006).
99. Gaasenbeek, M. et al. Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex changes and multiple forms of chromosomal instability in colorectal cancers. *Cancer Res* **66**, 3471-9 (2006).
100. Ying, J. et al. Functional epigenetics identifies a protocadherin PCDH10 as a candidate tumor suppressor for nasopharyngeal, esophageal and multiple other carcinomas with frequent methylation. *Oncogene* **25**, 1070-80 (2006).
101. Davison, E.J., Tarpey, P.S., Fiegler, H., Tomlinson, I.P. & Carter, N.P. Deletion at chromosome band 20p12.1 in colorectal cancer revealed by high resolution array comparative genomic hybridization. *Genes Chromosomes Cancer* **44**, 384-91 (2005).
102. Westbrook, T.F. et al. A genetic screen for candidate tumor suppressors identifies REST. *Cell* **121**, 837-48 (2005).

103. Jones, A.M. et al. Array-CGH analysis of microsatellite-stable, near-diploid bowel cancers and comparison with other types of colorectal carcinoma. *Oncogene* **24**, 118-29 (2005).
104. Tsafrir, D. et al. Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res* **66**, 2129-37 (2006).
105. de Stahl, T.D. et al. Chromosome 22 tiling-path array-CGH analysis identifies germ-line- and tumor-specific aberrations in patients with glioblastoma multiforme. *Genes Chromosomes Cancer* **44**, 161-9 (2005).
106. McCabe, M.G. et al. High-resolution array-based comparative genomic hybridization of medulloblastomas and supratentorial primitive neuroectodermal tumors. *J Neuropathol Exp Neurol* **65**, 549-61 (2006).
107. Rossi, M.R. et al. Array CGH analysis of pediatric medulloblastomas. *Genes Chromosomes Cancer* **45**, 290-303 (2006).
108. Mendrzyk, F. et al. Genomic and protein expression profiling identifies CDK6 as novel independent prognostic marker in medulloblastoma. *J Clin Oncol* **23**, 8853-62 (2005).
109. Ichimura, K. et al. Small regions of overlapping deletions on 6q26 in human astrocytic tumours identified using chromosome 6 tile path array-CGH. *Oncogene* **25**, 1261-71 (2006).
110. Magnani, I. et al. Identification of oligodendroglioma specific chromosomal copy number changes in the glioblastoma MI-4 cell line by array-CGH and FISH analyses. *Cancer Genet Cytogenet* **161**, 140-5 (2005).
111. Carrasco, D.R. et al. High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell* **9**, 313-25 (2006).
112. Chen, W. et al. Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas. *Blood* **107**, 2477-85 (2006).
113. Nakashima, Y. et al. Genome-wide array-based comparative genomic hybridization of natural killer cell lymphoma/leukemia: different genomic alteration patterns of aggressive NK-cell leukemia and extranodal Nk/T-cell lymphoma, nasal type. *Genes Chromosomes Cancer* **44**, 247-55 (2005).
114. Tagawa, H. et al. Comparison of genome profiles for identification of distinct subgroups of diffuse large B-cell lymphoma. *Blood* **106**, 1770-7 (2005).
115. Friedman, J.M. et al. Oligonucleotide Microarray Analysis of Genomic Imbalance in Children with Mental Retardation. *Am J Hum Genet* **79**, 500-513 (2006).
116. Wang, N.J., Liu, D., Parokonny, A.S. & Schanen, N.C. High-resolution molecular characterization of 15q11-q13 rearrangements by array comparative genomic hybridization (array CGH) with detection of gene dosage. *Am J Hum Genet* **75**, 267-81 (2004).
117. Harvard, C. et al. A variant Cri du Chat phenotype and autism spectrum disorder in a subject with de novo cryptic microdeletions involving 5p15.2 and 3p24.3-25 detected using whole genomic array CGH. *Clin Genet* **67**, 341-51 (2005).

118. de Vries, B.B. et al. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* **77**, 606-16 (2005).
119. Sahoo, T. et al. Microarray based comparative genomic hybridization testing in deletion bearing patients with Angelman syndrome: genotype-phenotype correlations. *J Med Genet* **43**, 512-6 (2006).
120. Koochek, M. et al. 15q duplication associated with autism in a multiplex family with a familial cryptic translocation t(14;15)(q11.2;q13.3) detected using array-CGH. *Clin Genet* **69**, 124-34 (2006).
121. Jacquemont, M.L. et al. Array- based comparative genomic hybridization identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J Med Genet* (2006).
122. Moon, H.J. et al. Identification of DNA copy-number aberrations by array-comparative genomic hybridization in patients with schizophrenia. *Biochem Biophys Res Commun* **344**, 531-9 (2006).
123. Wilson, G.M. et al. DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. *Hum Mol Genet* **15**, 743-9 (2006).
124. Kraus, J., Pantel, K., Pinkel, D., Albertson, D.G. & Speicher, M.R. High-resolution genomic profiling of occult micrometastatic tumor cells. *Genes Chromosomes Cancer* **36**, 159-66 (2003).
125. van Dekken, H. et al. Evaluation of genetic patterns in different tumor areas of intermediate-grade prostatic adenocarcinomas by high-resolution genomic array analysis. *Genes Chromosomes Cancer* **39**, 249-56 (2004).
126. Tomlins, S.A. et al. Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer 10.1126/science.1117679. *Science* **310**, 644-648 (2005).
127. Perner, S. et al. TMPRSS2:ERG Fusion-Associated Deletions Provide Insight into the Heterogeneity of Prostate Cancer 10.1158/0008-5472.CAN-06-1482. *Cancer Res* **66**, 8337-8341 (2006).
128. Castro, P. et al. Genomic profiling of prostate cancers from African American men. *Neoplasia* **11**, 305-12 (2009).
129. Schadt, E.E., Li, C., Ellis, B. & Wong, W.H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl* **Suppl 37**, 120-5 (2001).
130. Seo, J. et al. Optimizing signal/noise ratios in expression profiling: project-specific algorithm selection and detection p value weighting in affymetrix microarrays. *Bioinformatics* (2004).
131. Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-64 (2003).
132. Li, C. & Hung Wong, W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* **2**, RESEARCH0032 (2001).
133. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**, 31-6 (2001).
134. Irizarry, R.A. et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).

135. Rhodes, D.R. et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6 (2004).
136. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**, 695-701 (2008).
137. Wang, K., Li, M. & Bucan, M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet* **81**(2007).
138. Gudbjartsson, D.F. et al. Many sequence variants affecting diversity of adult human height. *Nat Genet* **40**, 609-15 (2008).
139. Weedon, M.N. et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**, 575-83 (2008).
140. Lettre, G. et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* **40**, 584-91 (2008).
141. Visscher, P.M. Sizing up human height variation. *Nat Genet* **40**, 489-90 (2008).
142. Perola, M. et al. Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet* **3**, e97 (2007).
143. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
144. Witte, J.S. Multiple prostate cancer risk variants on 8q24. *Nat Genet* **39**, 579-80 (2007).
145. Eeles, R.A. et al. Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* **40**, 316-21 (2008).
146. Thomas, G. et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nat Genet* **40**, 310-5 (2008).
147. Zheng, S.L. et al. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* **358**, 910-9 (2008).
148. Haiman, C.A. et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* **39**, 638-44 (2007).
149. Klotz, L. Re: cumulative association of five genetic variants with prostate cancer. *Eur Urol* **53**, 1298-9 (2008).
150. Eeles, R., Giles, G., Neal, D., Muir, K. & Easton, D.F. Reply to "Variation in KLK genes, prostate-specific antigen and risk of prostate cancer". *Nat Genet* **40**, 1035-1036 (2008).
151. Ahn, J. et al. Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. *Nat Genet* **40**, 1032-4 (2008).
152. Singh, D. et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-9 (2002).
153. Yeager, M. et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).
154. Korn, J.M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-60 (2008).
155. The International HapMap Project. *Nature* **426**, 789-96 (2003).
156. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**, 83-92 (2004).

157. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
158. Team, R.D.C. R: A language and environment for statistical computing. 2.9 edn (R Foundation for Statistical Computing, Vienna, Austria, 2009).
159. Rabbee, N. & Speed, T.P. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7-12 (2006).
160. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).
161. Philippar, U. et al. A Mena invasion isoform potentiates EGF-induced carcinoma cell invasion and metastasis. *Dev Cell* **15**, 813-28 (2008).
162. Hasegawa, Y., Murph, M., Yu, S., Tigyi, G. & Mills, G.B. Lysophosphatidic acid (LPA)-induced vasodilator-stimulated phosphoprotein mediates lamellipodia formation to initiate motility in PC-3 prostate cancer cells. *Mol Oncol* **2**, 54-69 (2008).
163. Turner, D.P., Findlay, V.J., Kirven, A.D., Moussa, O. & Watson, D.K. Global gene expression analysis identifies PDEF transcriptional networks regulating cell migration during cancer progression. *Mol Biol Cell* **19**, 3745-57 (2008).
164. Bear, J.E. et al. Negative regulation of fibroblast motility by Ena/VASP proteins. *Cell* **101**, 717-28 (2000).
165. Bear, J.E. & Gertler, F.B. Ena/VASP: towards resolving a pointed controversy at the barbed end. *J Cell Sci* **122**, 1947-53 (2009).
166. Woelfelschneider, A. et al. A distinct ERCC1 haplotype is associated with mRNA expression levels in prostate cancer patients. *Carcinogenesis* **29**, 1758-64 (2008).
167. Fauchereau, F. et al. The RhoGAP activity of OPHN1, a new F-actin-binding protein, is negatively controlled by its amino-terminal domain. *Mol Cell Neurosci* **23**, 574-86 (2003).
168. Pinheiro, N.A., Caballero, O.L., Soares, F., Reis, L.F. & Simpson, A.J. Significant overexpression of oligophrenin-1 in colorectal tumors detected by cDNA microarray analysis. *Cancer Lett* **172**, 67-73 (2001).
169. Ljubimova, J.Y. et al. Gene expression abnormalities in human glial tumors identified by gene array. *Int J Oncol* **18**, 287-95 (2001).
170. Dicken, B.J. et al. Lymphovascular invasion is associated with poor survival in gastric cancer: an application of gene-expression and tissue array techniques. *Ann Surg* **243**, 64-73 (2006).
171. Wei, A. et al. K+ current diversity is produced by an extended gene family conserved in Drosophila and mouse. *Science* **248**, 599-603 (1990).
172. Xu, C., Lu, Y., Tang, G. & Wang, R. Expression of voltage-dependent K(+) channel genes in mesenteric artery smooth muscle cells. *Am J Physiol* **277**, G1055-63 (1999).
173. Soldovieri, M.V. et al. Decreased subunit stability as a novel mechanism for potassium current impairment by a KCNQ2 C terminus mutation causing benign familial neonatal convulsions. *J Biol Chem* **281**, 418-28 (2006).
174. Fine, B. et al. Activation of the PI3K pathway in cancer through inhibition of PTEN by exchange factor P-REX2a. *Science* **325**, 1261-5 (2009).

175. Lin, H.K., Hu, Y.C., Lee, D.K. & Chang, C. Regulation of androgen receptor signaling by PTEN (phosphatase and tensin homolog deleted on chromosome 10) tumor suppressor through distinct mechanisms in prostate cancer cells. *Mol Endocrinol* **18**, 2409-23 (2004).
176. Boissan, M. et al. The mammalian Nm23/NDPK family: from metastasis control to cilia movement. *Mol Cell Biochem* **329**, 51-62 (2009).
177. Milon, L. et al. The human nm23-H4 gene product is a mitochondrial nucleoside diphosphate kinase. *J Biol Chem* **275**, 14264-72 (2000).
178. Milon, L. et al. nm23-H4, a new member of the family of human nm23/nucleoside diphosphate kinase genes localised on chromosome 16p13. *Hum Genet* **99**, 550-7 (1997).
179. Heinlein, C.A. & Chang, C. Androgen receptor in prostate cancer. *Endocr Rev* **25**, 276-308 (2004).
180. Nelson, P.S. et al. The program of androgen-responsive genes in neoplastic prostate epithelium. *Proc Natl Acad Sci U S A* **99**, 11890-5 (2002).
181. Volik, S. et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* **16**, 394-404 (2006).
182. Campbell, P.J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-9 (2008).
183. Guo, Z. et al. A novel androgen receptor splice variant is up-regulated during prostate cancer progression and promotes androgen depletion-resistant growth. *Cancer Res* **69**, 2305-13 (2009).
184. Hayer, J., Engel, M., Seifert, M., Seitz, G. & Welter, C. Overexpression of nm23-H4 RNA in colorectal and renal tumours. *Anticancer Res* **21**, 2821-5 (2001).
185. Abidin, M.R. & Eisele, D.W. Angioedema after long-term use of angiotensin-converting enzyme inhibitor. *Arch Otolaryngol Head Neck Surg* **117**, 1059 (1991).
186. Yoshimoto, M. et al. Absence of TMPRSS2:ERG fusions and PTEN losses in prostate cancer is associated with a favorable outcome. *Mod Pathol* **21**, 1451-60 (2008).
187. Best, C.J. et al. Molecular alterations in primary prostate cancer after androgen ablation therapy. *Clin Cancer Res* **11**, 6823-34 (2005).