CHARACTERIZATION OF COPY NUMBER ABERRATIONS AND EPIGENETIC MODIFICATIONS IN PROSTATE CANCER

by

Jung H. Kim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2010

Doctoral Committee:

Professor Arul M. Chinnaiyan, Co-Chair
Professor Kerby Shedden, Co-Chair
Professor Thomas W. Glover
Associate Professor Jill A. Macoska
Assistant Professor Sami N. Malek

ACKNOWLEDGEMENTS

I would first like to thank my mentor, Dr. Arul M. Chinnaiyan, for his expert guidance, support, encouragement, and inspiration. I would also like to thank Dr. Kerby Shedden, Dr. Thomas W. Glover, Dr. Jill A. Macoska, and Dr. Sami N. Malek for serving on my Doctoral committee and giving me valuable suggestions on this thesis work.

I would like to thank Saravana Dhanasekaran for his great advice, and Jindan Yu, Jianjun Yu, Xuhong Cao, Saravana Dhanasekaran, Daniel Rhodes, Scott A. Tomlins, and Xiaojun Jing, who have contributed to most of the work described here. I would like to express my gratitude to all the members in the Chinnaiyan lab for their support. Without them, none of the work described here could have been completed.

I would like to express my deepest gratitude to my husband and son. Thanks for putting up with my late nights and giving me unconditional love and encouragement through my Doctoral study and the writing of this work. I would also like to thank my mom, parents-in-laws, and my grandmother for giving constant support and love. And last but not least, I would like to thank my friends and all whose support helped me completing this thesis in time.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

CHARACTERIZATION OF COPY NUMBER ABERRATIONS AND EPIGENETIC MODIFICATIONS IN PROSTATE CANCER

by

Jung H. Kim

Co-Chairs: Arul M. Chinnaiyan and Kerby Shedden

Prostate cancer (PCa) is the most common epithelial cancer and second leading cause of cancer death among men in the US. Prostate cancer tumorigenesis is associated with numerous molecular events including transcriptomic, epigenetic, and copy number alterations. In this thesis, we characterized two major types of genome-wide events to understand their global implications and contributions in PCa.

First, we assessed genome-wide copy number variations (CNVs) using array comparative genomic hybridization of laser-capture microdissected prostate cancer samples representing multiple stages of PCa progression. Minimal common regions (MCR) of CNVs, including novel regions, were defined for each sample type. Integrative analysis of MCRs with matched gene expression profiles revealed genes with coordinated CNV and altered transcript expression during PCa progression. We also identified MCRs that distinguished PCa samples harboring or lacking an ETS gene fusion.

Secondly, we characterized genome-wide DNA methylation patterns, an epigenetic mark known to repress gene transcription. We employed a novel technology termed Methylplex-Next Generation Sequencing (M-NGS) which uses methylation sensitive restriction enzymes to enrich methylated genomic regions. The performance of M-NGS to characterize global methylation was tested in LNCaP prostate cancer and PrEC benign prostate epithelial cells. Multiple techniques, including bisulfite sequencing, validated the results. Detailed analyses revealed diverse promoter methylation patterns which correlated with transcriptional repression. Interestingly, integration of DNA methylation and H3K4me3 ChIP-Seq data in LNCaP identified differential epigenetic regulation of specific transcript isoforms.

We next employed M-NGS to characterize PCa tissue samples. We identified 2,481 cancer-specific methylated regions, including WFDC2 promoter methylation, which served as a PCa biomarker and was validated on an independent tissue cohort. Finally we used a 3-way integrative analysis of these genome-wide events and identified regions with co-occurrence of copy number gain/loss and DNA methylation, associated with aberrant gene expression in PCa .

In summary, this thesis work presents a comprehensive analyses of genome-wide copy number and methylation changes and its global implications on transcriptional regulation for the first time in prostate cancer. The datasets generated here will be a valuable public resource for genome-wide analysis in the future.

CHAPTER 1

INTRODUCTION

The Prostate

Prostate (Greek: prohistani, meaning to stand in front of) was described by Herophilus in 355 B.C.E. as a small organ located in front of the bladder. The prostate, a male accessory sex gland located below the bladder is found exclusively in mammals and produces many components such as fructose, zinc ions and various proteins that maintain the fluid nature of semen. The prostate gland has been extensively studied as a model for androgen action in regulating epithelial cell growth. Androgens are important regulators of male sexual differentiation including their essential role in directing the development of prostate (1-2). The effects of androgens have been extensively studied in prostate models due to the relationship between androgens and the development of widely prevalent neoplastic diseases of the prostate such as benign prostatic hyperplasia (BPH) and prostate adenocarcinoma (PCa)(3).

In light of this, prostatic diseases remain widely prevalent in the Western world. An estimated 75% of men show histologic evidence of BPH by age eighty, and PCa remains the second most common cancer-related death in the US (4). Several factors such as age, lifestyle and diet as well as inheritable factors are known to play a role in the development of prostate cancer (4). The mortality rate of prostate cancer varies among races, and African American men have 5 fold higher mortality frequency compared to Asian Americans and 2 fold greater compared to Caucasians. The incidence of prostate cancer also correlates with age, where around 63% of new cases occur among men over 65 years of age. According to 2009 report from American Cancer Society, the number of estimated new cases of prostate cancer is 192,280 and estimated death is 27,360 in US males. The improvement in early prostate cancer diagnosis and treatment made during last two decades raised the 5-year survival rate from 69% to 99%.

Stages of Prostate Cancer

The cellular architecture of a human prostate duct primarily consists of three cell types organized in two layers 1) secretory luminal cells (cytokeratin, 8, 18 and CD57 positive) that form the inner layer and 2) cytokeratin5, 14 and CD44 positive basal cells found between the luminal layer and the underlying basement membrane **(Fig. 1.1)**. A minor population of neuroendocrine cells that make the third cell type are found dispersed in the basal layer and specifically expresses a maker protein called chromagraninA.

The commonly used Gleason grading system of prostate cancer was developed by Dr. Donald Gleason in 1974 based on the breach that occurs in the tissue glandular architecture (shape, size and differentiation of glands) during tumorigenesis (5). By histologic evaluation pathologists determine a primary (representing the majority of tumor) and secondary Gleason grades (assigned to the minority of tumor), such as Gleason grades 3, 4 or 5. The sum of the primary and secondary Gleason grades are then considered as Gleason Score (GS) which is a number that ranges from 2 to 10 (primary grade + secondary grade = GS). For example a primary grade of 3 and a secondary grade of 4 will give a Gleason score of 7.

A high Gleason score (≥8) represents severe disease with a poorly differentiated glandular pattern. Currently, the histologic grade defined by Gleason score, along with serum levels of Prostate Specifc Antigen (PSA) and clinical stage are the most commonly used parameters for treatment decisions, prediction of organ confined disease and disease progression after treatment. However patients with similar clinicopathologic findings frequently display widely heterogeneous disease courses and clinical responses to treatment. Compounding these observations is the conundrum that patients with prostate cancer often have highly heterogeneous disease, with several distinct foci of tumor (which may have differing Gleason grades) and preneoplastic lesions (such as Prostatic Intraepithelial Neoplasia (PIN) and Proliferative Inflammatory Atrophy (PIA) are found in a single patient. Hence an effort to fully understand the genetics of disease initiation and progression is a top priority in prostate cancer research. Advent of high throughput technologies have provided a wealth of information in transcriptome changes and genomic abberations that occur in cancer, hence a review of the available literature under the headings Pre and post microarray era in these areas are presented below.

Molecular Characterization of Prostate Cancer

Pre-microarray Era

Earlier detection of prostate cancer is important. While the tumor from earlier stage is small and non-invasive and easily curable, the advanced form of this disease is often associated with cancer-related mortality and morbidity. Prostate-specific antigen (PSA) is a tumor marker with high sensitivity used for the earlier detection of prostate cancer. Yet it has low specificity as there is significant overlap between PSA levels found in cancer and benign prostatic hyperplasia, and it often fails to detect cancer in significant number of patients undergoing prostate biopsy (6).

Several studies have characterized genetic events that underlie prostate cancer which include PTEN genomic deletion. PTEN (Phosphatase and tensin homolog) is a tumor suppressor biomarker that negatively regulates cell migration and cell survival (7-9). PTEN induces G1 cell-cycle arrest in prostate cancer development (9) acting as an “off” switch for PI3K/AKT signaling pathway, which plays a role in proliferation, apoptosis, nutrient response, DNA damage, and cell size (10). PTEN is located within 10q23 and is one of the most frequently deleted and mutated genes in prostate cancer. As a result of PTEN loss, PIP3 is accumulated and activates PI3K/AKT pathway (8-11). Loss of PTEN activity in prostate cancer is implicated in several studies. The paraffin tissues of 109 cases were immunostained with PTEN antiserum, and the absence of PTEN expression was correlated with Gleason score, especially those with gleason score of 7 or higher (9). PTEN has a haploinsufficient property for prostate cancer progression. While the transgenic adenocarcinoma of mouse prostate model containing PTEN (+/-) progressed significantly faster than its wildtype control, there was no statistical significance between the progression rate of PTEN (+/-) and PTEN(-/-) mouse prostate model, although the author has mentioned that it was somewhat slower in PTEN (+/-) than PTEN (-/-) model (12). The researchers have suggested the dosage effect of PTEN in prostate cancer progression and proposed the role of homozygous PTEN loss in invasive cancer development. In a murine prostate cancer model, the mice with heterogeneous PTEN loss developed PIN within 12-16 months time frame, while the mice with homozygous loss took only 9 weeks to develop invasive form of carcinoma (11). Comparing the mouse model with decreasing PTEN activity (*PTEN* +/+, *PTEN* hypomorphic/+, *PTEN* +/-, and *PTEN* hypomorphic/-), the PTEN downstream genes such as Akt, p27, mTOR, and FOXO3 are affected in dosage-dependent manner (13).

Similar to loss of PTEN expression observed in earlier stage of prostate cancer, the NKX3.1 loss in pre-cancerous PIN is widely reported (14-17),and is also a useful biomarker for earlier detection of prostate cancer. NKX3.1 is a prostate-specific homeobox gene regulated by Androgen (18). This gene is only expressed in androgen-dependent cell line such as LNCaP and plays an opposing role to androgen-driven differentiation (18). The loss of heterozygosity (LOH) in 8p21, where NKX3.1 is located, is a well reported event during prostate cancer progression and speculated to play an important role during earlier stage of cancer development (19-20). It is reported that LOH in 8p21 was present in 34 out of 54 PIN samples, and consequently more than half of PIN samples showed either reduced or absent NKX3.1 expression (20). NKX3.1 is a tumor suppressor gene, and the mouse model with NKX3.1 deletion developed PIN (15, 17). Importantly the loss of NKX3.1 expression correlated with tumor progression, and loss of NKX3.1 expression is more notable in advanced tumors and hormone refractory prostate cancers (14). The combined study of FISH and methylation-specific PCR (MS-PCR) on 48 primary prostate cancer samples summarized the silencing mechanism of NKX3.1 in dosage dependent manner (16). NKX3.1 is not only subjected to a deletion, but prone for methylation changes as well. The LOH was reported in 27 out of 43 samples, and the increased level of methylation is also reported in 33 out of 40 malignant compared to adjacent normal samples. The samples with combined LOH and increased level of methylation had lowest expression (16). Interestingly, NKX3.1 and PTEN together has a synergistic effect on AKT pathway activation (21) playing a role in cell growth and survival, and this activation is only restricted to prostate, such that AKT pathway activation is not observed in non-NKX3.1 expressing tissues (21).

Another biomarker that is used for earlier detection of prostate cancer is GSTP1 (22). The decrease in GSTP1 expression is studied using immunohistochemical staining on 91 prostate cancer samples with GSTP1 antibodies, where 88 of them failed to show detectable amount of enzymes (23). The repression via hypermethylation of GSTP1 is addressed from both functional analysis and sequencing-based validation (24-25). The 5-aza-deoxycytidine (5Aza) is an inhibitor to DNA methylation, which acts as a demethylating agent. The reactivation of GSTP1 transcription was demonstrated using 5Aza treatment on prostate cancer cell lines, implicating the role of CpG methylation in GSTP1 gene repression (25). GSTP1 is especially useful for earlier detection of prostate cancer, because more than 70% of HGPIN samples already exhibits hypermethylation in its promoter, and in localized prostate cancer, the percentage harboring GSTP1 promoter hypermethylation increases to over 90% (22, 24). GSTP1 serves as non-invasive biomarker, since its hypermethylation status can be detected in urine, ejaculate, and plasma from men with prostate cancer (22).

Post-microarray Era

With the advent of the microarray technology, the gene expression profiling on large number of cancer samples provided new insights to the prostate cancer disease progression and development, as well as a new understanding of the molecular events that occur. The microarray analysis had been successfully used to characterize various types of cancer since 1995. To date, Oncomine, an online gene expression database consortium (26) with data-mining and visualization tools, features 39,000+ samples from more than 500 profiling studies on various types of cancer (2,244 samples from 43 different studies on prostate cancer). This reveals the extent to which microarray technology has been used in cancer biology. Recently, the copy number profiling of 3,490 samples from 27 studies has also been added to this application, and this data can be integrated with existing gene expression microarray data.

Using this genome-wide approach, several genes have been identified based on their differential expression in prostate cancer and evaluated as biomakers as shown in the case of Hepsin, AMACR, and EZH2 (27-31). From the gene expression analysis on more than 50 tissues samples (normal adjacent, BPH, localized and metastatic prostate cancer), Hepsin is found to be correlated with the clinical outcome (27), and is differentially expressed between localized and metastatic samples, and while it is over-expressed in primary tumor, a decreased expression is observed in metastatic cancer samples (28). The role of Hepsin in prostate cancer progression is shown in a mouse model study, where the over-expression of Hepsin has caused the disorganization of basement membrane, promoting tumor progression and metastasis (32). It is also reported by quantitative-PCR (qPCR) analysis on 90 patient samples, that there is a change in its expression with statistical significance between samples that are with low risk compared to those with high risk for relapse (33).

AMACR is a proven biomarker originally nominated from multiple microarray datasets including Dhanasekaran, Luo, Welsh studies, (27, 34-35), where it was up-regulated in prostate cancer. Moreover, the elevated level of protein is also detected from tissue microarray study (29). AMACR performed better as the serum biomarker than PSA with higher sensitivity and specificity (29, 36). As a non-invasive approach, the urinalysis on AMACR protein showed 100% sensitivity and 58% specificity in urine samples collect from 26 individuals (37).

EZH2 is a transcriptional repressor and a marker of aggressive cancer as shown in tissue microarray on over 900 patient samples (31). Functionally, a polycomb-group (PcG) protein, EZH2, is a gene silencer and methylates histone repression marker H3K27me3, which leads to chromatin condensation (38). It is shown from microarray data analysis, that EZH2 is over-expressed in metastatic prostate cancer, especially in hormone-refractory metastatic samples (30-31), and the prostate cancer samples with higher level of EZH2 were noted to have a poorer prognosis (30). Later, the mechanism of EZH2 over-expression is explained by miRNA (hsa-mir-101) deletion. The expression of EZH2 is inhibited by hsa-mir-101, and the genomic loss of hsa-mir-101 in prostate tumors lead to over-expression of EZH2 (39). Yu et al. have identified the 14 direct targets of PcG called “polycomb repression signature” using chromatin immunoprecipitation (ChIP) with EZH2, SUZ12, H3K27me3 antibodies and established the relationship between the PcG repression signature with poor clinical outcome in multiple microarray data sets of breast and prostate cancer (38).

In addition to the biomarker discovery, the microarray analysis can be used to classify and predict the patient outcome in various cancers including prostate using “underlying gene expression differences” (40). Using the microarray profiling data, Singh et al. have accurately distinguished prostate tumor samples from normal with up to 92% accuracy (40). Similarly, using the gene expression of 12,625 transcripts on microarray, Glinsky et al. have identified the clusters of genes that can differentiate recurrent from non-recurrent prostate cancer samples, where 88% of recurred prostate samples were correctly classified into poor-prognostic group based on Kaplan-Meier analysis (41).

The pathways that are often dysregulated in prostate cancer can be also revealed through microarray analysis (42-43). One of the examples is shown from the meta-analysis performed across multiple independent microarray dataset of prostate cancer. In this study, the cohort of genes involved in polyamines and purine biosynthesis pathway from KEGG database are found to be consistently and significantly dysregulated across multiple prostate cancer microarray studies (42). The gene expression analysis on oligo-array with 63,175 probes identified the genes involved in cell cycle and DNA replication and repair, and the samples with the alteration in these genes indeed had high proliferation rate (43).

Gene Fusion

Recently paradigms of prostate cancer have been radically altered by the discovery of recurrent gene fusion events involving ETS family of transcriptional factors. Through fusion event, a dormant gene can be activated, when repositioned to tissue-specific or ubiquitously active genome loci (44). The path that leads to the gene fusion discovery in prostate cancer is described below. Microarray analysis of prostate cancer tissues identified transcriptional dysregulation that specifically occurred in the cancer epithelial compartment (27, 45). The Cancer Outlier Profile Analysis (COPA) of multiple PCa microarray dataset revealed the presence of distinct molecular subtypes among the cancer samples. Unlike a standard statistical test, COPA analysis was specifically designed to identify highly expressed genes in a subset of samples (46). It revealed overexpression of the transcription factor ERG in 40-50% of PCa cases and ETV1 in 10% of the specimens, and their expression was found to be mutually exclusive. A subsequent measurement by qPCR revealed disparity in exon level expression of these ETS genes in corresponding ETS over-expressing samples, where the 5' exons had very low/undetectable expression, while the 3' exons, recorded exceedingly high expression values which ranked them as number one outlier in the cancer samples. This intriguing finding necessitated Rapid Amplification of CDNA Ends (RACE) analysis to examine the 5' region of these outlier ETS family genes to explain the disparate exon level expression in cancer. Results from the RACE analysis provided the first confirmation of a gene fusion event in prostate cancer between ETS family genes and TMPRSS2, a prostate specific androgen regulated gene. The gene fusion event explained the outlier expression pattern of these genes and copy number analysis in prostate cancer by aCGH (discussed in detail under Copy Number and Microarray) explained the low expression of 5' exons. There are around 30 genes harbored in the genomic region between ERG and TMPRSS2 genes on human chromosome 21. Array CGH data revealed genomic deletion of this region and deletion boundaries fell within TMPRSS2 and ERG genes exactly accounting for the disparate exon expression levels. Several studies have now looked at the prevalence of these events and their value as prognostic/ diagnostic makers in prostate cancer (47-48). The role of these gene fusions in prostate cancer initiation and progression has also been studied (44, 49). A complete list of various types of gene fusion identified in prostate cancer is schematically represented in Figure 1.2. The above described is just one example to demonstrate the important contribution of high throughput technologies and its integrative analysis in understanding disease biology and much more.

Genomic Aberrations in Prostate Cancer Progression

Background in Copy Number Aberration in Prostate Cancer

In cancer cells, chromosomal aberrations may result in activation of oncogenes and inactivation of tumor suppressor genes (50). Amplification of oncogenes such as AKT2 in ovarian cancer, REL in Hodgkin lymphoma, ERBB2 in breast and ovarian tumors, MYCL1 in small cell lung cancer, MYCN in neuroblastoma, EGFR in glioma and non-small cell lung cancer and MYC in various cancers have been studied in detail (51).Understanding the contours of a cancer genome, may provide valuable mechanistic insights into the disease as well as in identifying therapeutic gene targets. This fact is highlighted by the development of Trastuzumab (Herceptin) a targeted therapy for metastatic breast cancer patients with *ERBB2* gene amplification (52).

Traditionally, to study the copy number variations among given samples, the techniques such as comparative genomic hybridization (CGH), fluorescence in situ (FISH) hybridization, and PCR-based LOH analysis were employed (19-20, 53-54). Several important findings were made using these techniques including the copy number aberration such as deletion in 8p12-21 (19-20, 53-54). PTEN and NKX3.1 are located within this frequently deleted 8p region (16, 55). Besides recurrent regions of alterations such as gains in 7q, 8q, 18q, and Xq and losses in 1p, 8p, 10q, 13q, 16q, 19, and 22 , the altered genes such as AR, MYC, and Cyclin-D1 in prostate cancer are also reported (53-54, 56).

Androgen receptor (AR), a transcription factor of major importance in normal prostate function, is reported to be frequently amplified and mutated in prostate cancer (57). The AR gene amplification status is detected using FISH in hormone-refractory samples, and cancer samples with AR amplification had 2-fold increase in AR level compared with the ones without AR amplification (58). The increased level of AR mRNA is shown to be both necessary and sufficient for the transformation into hormone-refractory cancer, also known as androgen-independent prostate cancer. While the hormone therapy with AR antagonists is initially effective at first in tumor growth, eventually the prostate cancer becomes androgen-independent, thus fails to respond. The hormone-sensitive and hormone-refractory prostate cancer xenograft pairs were profiled on the microarray to uncover the mechanism behind the resistance to hormone therapy, and AR cDNA was the most differentially expressed probe. When mice were implanted with LAPC4 cells infected with androgen knockdown, the growth of tumor is stopped (59). Recently, the compounds named RD162 and MDV3100 is developed as a second-generation anti-androgen treatment (60). Unlike previously developed anti-androgen drug for hormone therapy, they work under the increased level of AR, and these compounds are currently being clinically tested. In addition to the increased AR, the mutation in AR also plays a role in androgen independency in prostate cancer (61-62).

In prostate cancer, a dramatic increase in number of aberrations by more than 50% in recurrent tumor compared to corresponding primary tumor has been previously reported (54). In general, the metastatic samples exhibit wide-ranges of amplified and deleted regions from various sites in entire sets of chromosomes. This genome-wide events support the hypothesis that “the accumulation of multiple genetic changes, perhaps as a result of genomic instability, is associated with prostate cancer progression” (63). The new findings and understanding of this event not only will help to uncover the pathogenesis of prostate cancer, but to design new therapeutic targets for clinical use as well (64).

Copy Number and Microarray

Identifying the regions of genomic aberration such as amplification and deletion in genome-wide manner is also benefited from the microarray technology. Unlike the traditional approaches limited by low resolution of several megabases or single-gene approach, array CGH (aCGH) based on microarray technology, was developed to overcome this limitation and began new era of genome-wide copy number analysis. In brief, more than decade ago, Pinkel and Albertson successfully showed the copy number variation on aCGH platform using chromosome 20 and four copies of X chromosome from breast cancer (65). Since then, aCGH is widely performed on BAC, PAC, cosmid, cDNA and SNP arrays using samples from various sources including tissues, cell lines, and xenografts (66-71). The tumor suppressor genes located in 16q23-q terminus is identified from 16 prostate cancer tissues using aCGH technique on the microarray platform made from BAC, PAC, cosmid clone contigs covering 78% of entire q arm of chromosome16 (66). Using similar approach, the 8q amplification containing MYC and TPD52 oncogenes with elevated expression in tumor is identified using BAC and SNP arrays (67-69).

In addition to the identification of aberration-prone genomic regions in cancer, the result from aCGH could be integrated with the gene expression array analysis from matched samples to give the insights of underlying biology and in identifying the genes of interest and studying the effect of copy number changes on gene expression in a genome-wide scale (70-71). Genome-wide integrative analysis between aCGH and gene expression on matched samples in breast and prostate cancer show 40-60% correlation rate between genes that are highly amplified and over-expressed (72-73). More recently, the recurrent gene fusion events involving ETS transcriptional family in prostate cancer is reported. According to initial report, TMPRSS2 is fused to either ETV1 or ERG in mutually exclusive manner in 23 out of 29 prostate cancer samples (74). Using high-density BAC aCGH, the locus of TMPRSS2 from 21q22.3 is shown to be a hotspot for rearrangement with 75% genomic alteration rate (75). Furthermore, the use of high-resolution tiling array allowed to pin-point the breakpoints along the chromosomes (76).

Diagnostic and Prognostic Use of Copy Number Aberrations

Several groups have attempted to determine the types of cancer based on copy number alterations observed in each sample group or classify samples with bad prognosis (77-79). Using aCGH technology, the recurrent copy number alterations in samples that are linked to good prognosis vs. bad prognosis are addressed (77). Among genomic alterations that are linked with the samples with bad prognosis include gain in 8q24 containing MYC and loss at 10q23 containing PTEN (77). In a similar study, the copy number analysis from the cohort of 64 patients on aCGH platform proposed the gain at 11q13.1 to be a predictor of recurrence independent of tumor stage and grade and the loss at 8p23.2 to be the marker for advanced form of the disease (79). The DNA copy number analysis on androgen-sensitive (AS) and androgen-insensitive (AI) cell lines revealed vast differences between these two sample types (78). AI samples exhibited more extensive degree of amplification and higher number of genomic aberrations compared to AS samples. Transcriptional regulation differences between these two were previously known, and some of these transcriptomic differences can now be explained by underlying genomic copy number variations (78).

Epigenetic Modification in Prostate Cancer Progression.

Background on DNA Methylation – Age, Race, Cancer, and Field Effects

In addition to the genomic aberrations, the epigenetic modifications such as DNA methylation and histone modifications also occur simultaneously during prostate cancer progression. The epigenetic changes tend to occur earlier in cancer development, believed to be therapeutically reversible (80) and plays an essential role in transcript regulation. Unfortunately, the mechanism of epigenetic changes in prostate cancer development still remains to be unveiled.

The hyper and hypomethylation events during cancer progression is widely known, however, the variation in methylation between populations also depends on age and ethnicity (81-82). The age-related methylation is observed in normal tissues from older population. Although, the matched tumor tissues exhibited significantly higher level of methylation than the matched normal pairs, there was a strong correlation between the level of methylation and the age among the genes (such as GSTP1, RASSF1, and RARbeta2) monitored by pyrosequencing technique (82). The occurrence of prostate cancer differs among African-American, Caucasian, and Asians. While the African-Americans have the highest susceptibility of the disease, the Asians have the lowest incidents of prostate cancer. The GSTP1 promoter methylation is a diagnostic marker in prostate cancer and is hypothesized to play a role in cancer progression via gene inactivation resulting from promoter methylation, and its methylation status is monitored in each ethnic group. The highest difference between the rate of GSTP1 promoter methylation in cancer and benign samples is observed among African Americans who have the highest occurrence rate of the disease (81).

In addition to the age- and ethnicity-driven methylation differences among general public, the field effect of cancer also should be taken in concern. The field effects have been reported in gene expression, protein expression, and gene promoter methylation in tumor-surrounding tissues (82-83). For example, the methylation status of 5 known prostate cancer methylation-target genes in surrounding tissues (up to 3mm from the malignancy) is monitored. A total of 4 genes out of 5 harbored the promoter methylation as a result of field effect. The use of distanced benign tissues or tissues from cancer-free organs in paired sample analysis should be useful for field effect minimization.

Hypermethylated Genes as Diagnosis and Prognostic Biomarker

Promoter methylation resulting in loss of gene expression especially in tumor suppressor genes are widely reported in prostate cancer. The frequently hypermehylated genes can serve as diagnostic and prognostic markers for prostate cancer as shown in well-characterized GSTP1. The epigenetic modification such as DNA methylation occurs earlier in prostate cancer progression (84-85), such that the majority of HGPIN samples already retained GSTP1 and Cyclin D2 promoter hypermethylation (85-87). Because of its possible advantage as an early detection marker for prostate cancer and ability to be detected from urine samples as a non-invasive method, the methylation marker is gaining in strength (88-89).

Instead of using methylation marker from a single gene, the use of hypermethylation status in multiple loci such as TIG1, APC, PTGS2, and GSTP1 not only improved the specificity and sensitivity in disease diagnosis, it also showed stronger correlation with the disease progression and gleason score as well (90). The increased risk for prostate-specific mortality in the patients with RUNX3 and APC promoter hypermethylation is reported (91). The hypermethylated genes such as ASC and CDH13 showed increased risk for recurrence after radical prostatectomy (92). The target genes of DNA methylation can be used as a prognostic tool to determine the clinical outcome of the disease as shown above (90-92).

Therapeutic Relevance in Methylation Study

Identifying methylation target genes from earlier stage of prostate cancer also has a therapeutic value for the treatment of the disease. The DNA sequence is not altered during epigenetic modification, thus it is believed to be reversible (80, 93). Promoter methylation alters the gene expression especially if it resides on gene promoter regions. In vitro, the use of demethylating agent 5-Aza demonstrated its ability to re-activate silenced genes (94-95). Although the use of such approach in clinical setting is still under trial and at developmental stage, several drugs such as 5-azacytidine (Vidaza®) and 5-aza-2'-deoxycytidine (decitabine) have been clinically tested and FDA-approved to be used in terminal blood cancer patients.

Hypomethylation during Cancer Progression

Unlike promoter hypermethylation starting to accumulate during earlier stage of prostate cancer, DNA hypomethylation arises at much later stage (96). Using the antibody for 5-methylcytosine (5MeC), the immunocytochemistry levels of stained tumor sections from different stages of prostate cancer were quantified, and revealed that the metastatic samples showed significant reduction of 5 MeC compared to normal and localized prostate cancer samples (96). While the hypermethylation event is more targeted for gene promoter, the hypomethylation occurs in genome-wide manner, especially within repeat elements such as LINE1 and Alu (96-97). Also, with the integrative analysis with gene expression microarray, it was revealed that the genes that are over-expressed and hypomethylated include subset of testis antigen genes, whose change was present among localized samples, but more pronounced among metastatic samples (96). Lastly, in contrast to the hypermethylation events especially targeting CpG islands, the hypomethylation is more heterogeneous, such that the samples from same patient evidently showed wide variability in hypomethylation sites (96-97).

DNA Methylation and Histone Modification

The histone modifications along with DNA methylation may play an important role in tumorigenesis. While the histone acetylation opens up the chromatin for transcriptional activation, the histone methylation is generally associated with transcriptional repression as seen in H3K9 and H3K27 trimethylation. H3K4me3, however, is associated with active chromatin, thus gene activation. The previous studies indicate the rare presence of H3K4me3 in the regions of DNA methylation (98). In rice genome, while strong association between DNA methylation and transcriptional repression is established, in presence of both H3K4me3 and DNA methylation, the transcriptional activity was reduced by lesser degree (99). The tri-methylated histone H3 methylated at Lys27 (H3K27me3) is a key histone marker for epigenetic repression. As seen in H3K4me3 histone marks, DNA methylation and H3K27me3 is also occupying mutually exclusive regions on the genome (100). The independence between H3K27me3 histone repressive mark and DNA methylation is demonstrated using chromatin immunoprecipitation microarrays (ChIP-chip). Among the genes that contain H3K27me3 and silenced in prostate cancer, only small portion of these gene contained promoter DNA methylation. The downregulation of EZH2 also had no effect on the genes containing promoter DNA methylation, while it successfully restored the expression of H3K27me3 target genes (101). DNA methylation and histone modification is believed to be two independent mechanisms for gene silencing, and the association between these two mechanisms are not yet fully understood (101).

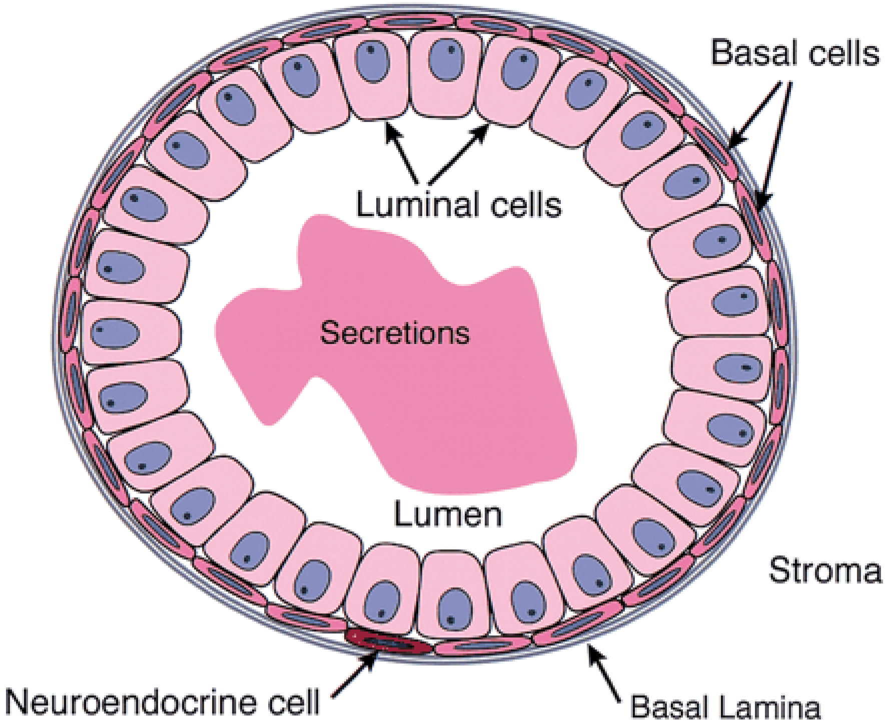
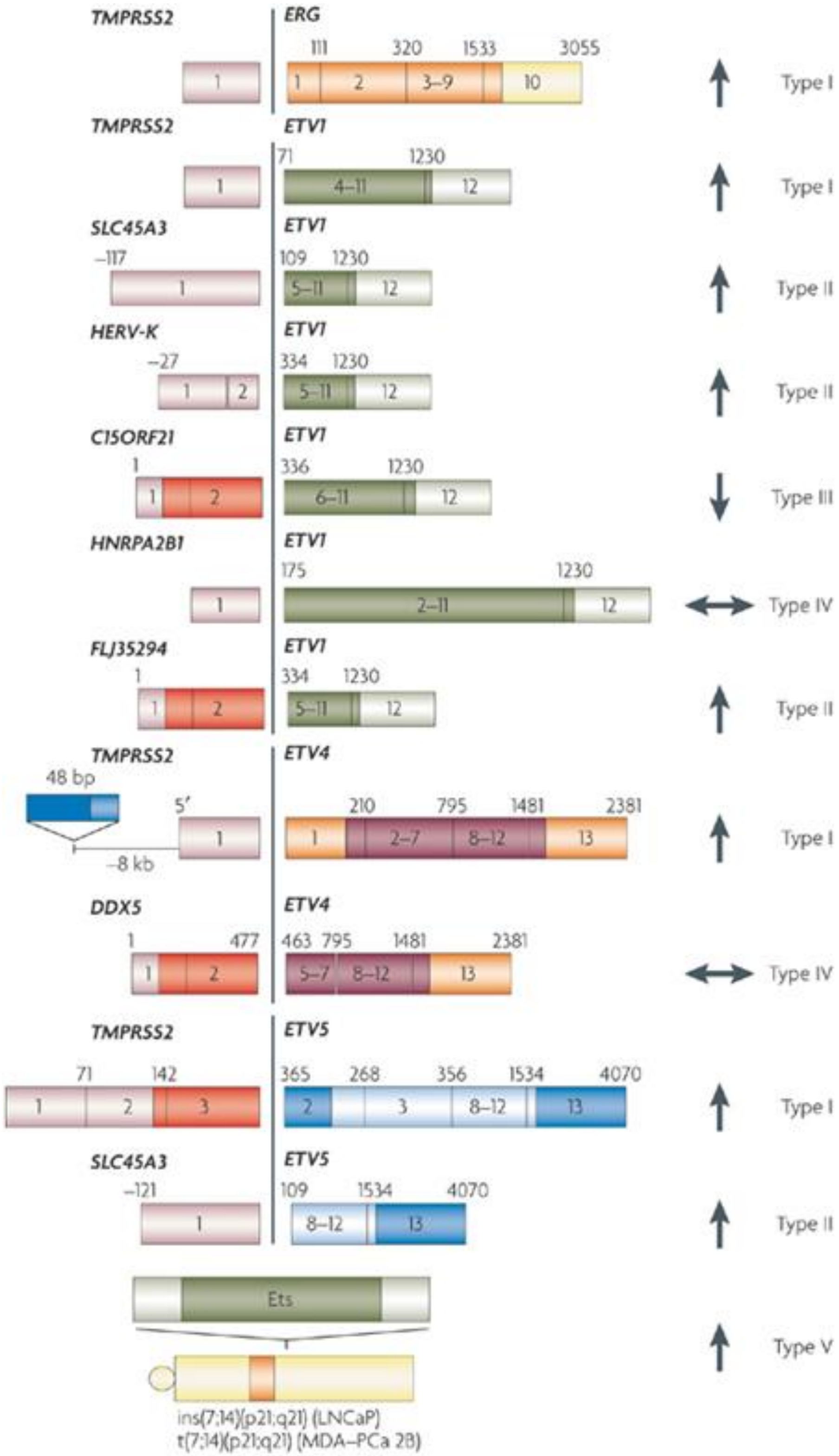


Figure 1.1 Schematic depiction of the cell types found in a human prostatic duct. Secretory luminal cells that form the inner layer and basal layer found encased by the basement membrane are indicated. The figure has been reproduced from the review by Shen and Shen et al (102).



Nature Reviews | Cancer

Figure 1.2 List of gene fusions in prostate cancer. The schematic represents the gene fusions characterized in prostate cancers. The 5' fusion partners and their corresponding

3’ partners are depicted on the right and left respectively. . Coding exons are depicted in darkly shaded boxes and non-coding regions in lighter shade. Arrows indicate androgen regulation ; Upregulation- upward pointing arrows , downregulation- downward pointing arrow, no influence- horizontal arrows. This figure has been reproduced from review published by Kumar et.al., on gene fusions in prostate cancer (47).

References

1. Marker PC, Donjacour AA, Dahiya R, & Cunha GR (2003) Hormonal, cellular, and molecular control of prostatic development. *Dev Biol* 253(2):165-174 .
2. Yeh S*, et al.* (2002) Generation and characterization of androgen receptor knockout (ARKO) mice: an in vivo model for the study of androgen functions in selective tissues. *Proc Natl Acad Sci U S A* 99(21):13498-13503 .
3. Cunha GR*, et al.* (1987) The endocrinology and developmental biology of the prostate. *Endocr Rev* 8(3):338-362 .
4. Nelson WG, De Marzo AM, & Isaacs WB (2003) Prostate cancer. *The New England journal of medicine* 349(4):366-381 .
5. Gleason DF & Mellinger GT (1974) Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J Urol* 111(1):58-64 .
6. Stamey TA, Johnstone IM, McNeal JE, Lu AY, & Yemoto CM (2002) Preoperative serum prostate specific antigen levels between 2 and 22 ng./ml. correlate poorly with post-radical prostatectomy cancer morphology: prostate specific antigen cure rates appear constant between 2 and 9 ng./ml. *J Urol* 167(1):103-111 .
7. Li J*, et al.* (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* 275(5308):1943-1947 .
8. Backman SA*, et al.* (2004) Early onset of neoplasia in the prostate and skin of mice with tissue-specific deletion of Pten. *Proc Natl Acad Sci U S A* 101(6):1725-1730 .
9. McMenamin ME*, et al.* (1999) Loss of PTEN expression in paraffin-embedded primary prostate cancer correlates with high Gleason score and advanced stage. *Cancer Res* 59(17):4291-4296 .
10. Sansal I & Sellers WR (2004) The biology and clinical relevance of the PTEN tumor suppressor pathway. *J Clin Oncol* 22(14):2954-2963 .
11. Wang S*, et al.* (2003) Prostate-specific deletion of the murine Pten tumor suppressor gene leads to metastatic prostate cancer. *Cancer Cell* 4(3):209-221 .
12. Kwabi-Addo B*, et al.* (2001) Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. *Proc Natl Acad Sci U S A* 98(20):11563-11568 .
13. Trotman LC*, et al.* (2003) Pten dose dictates cancer progression in the prostate. *PLoS biology* 1(3):E59 .
14. Bowen C*, et al.* (2000) Loss of NKX3.1 expression in human prostate cancers correlates with tumor progression. *Cancer Res* 60(21):6111-6115 .
15. Kim MJ*, et al.* (2002) Nkx3.1 mutant mice recapitulate early stages of prostate carcinogenesis. *Cancer Res* 62(11):2999-3004 .
16. Asatiani E*, et al.* (2005) Deletion, methylation, and expression of the NKX3.1 suppressor gene in primary human prostate cancer. *Cancer Res* 65(4):1164-1173 .
17. Abdulkadir SA*, et al.* (2002) Conditional loss of Nkx3.1 in adult mice induces prostatic intraepithelial neoplasia. *Molecular and cellular biology* 22(5):1495-1503 .

18. He WW*, et al.* (1997) A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics* 43(1):69-77 .

19. Vocke CD*, et al.* (1996) Analysis of 99 microdissected prostate carcinomas reveals a high frequency of allelic loss on chromosome 8p12-21. *Cancer Res* 56(10):2411-2416 .

20. Emmert-Buck MR*, et al.* (1995) Allelic loss on chromosome 8p12-21 in microdissected prostatic intraepithelial neoplasia. *Cancer Res* 55(14):2959-2962 .

21. Kim MJ*, et al.* (2002) Cooperativity of Nkx3.1 and Pten loss of function in a mouse model of prostate carcinogenesis. *Proc Natl Acad Sci U S A* 99(5):2884-2889 .

22. Nakayama M*, et al.* (2004) GSTP1 CpG island hypermethylation as a molecular biomarker for prostate cancer. *Journal of cellular biochemistry* 91(3):540-552 .

23. Lee WH*, et al.* (1994) Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc Natl Acad Sci U S A* 91(24):11733-11737 .

24. Nakayama M*, et al.* (2003) Hypermethylation of the human glutathione S-transferase-pi gene (GSTP1) CpG island is present in a subset of proliferative inflammatory atrophy lesions but not in normal or hyperplastic epithelium of the prostate: a detailed study using laser-capture microdissection. *Am J Pathol* 163(3):923-933 .

25. Lin X*, et al.* (2001) GSTP1 CpG island hypermethylation is responsible for the absence of GSTP1 expression in human prostate cancer cells. *Am J Pathol* 159(5):1815-1826 .

26. Rhodes DR*, et al.* (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6(1):1-6 .

27. Dhanasekaran SM*, et al.* (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature* 412(6849):822-826.

28. Srikantan V, Valladares M, Rhim JS, Moul JW, & Srivastava S (2002) HEPSIN inhibits cell growth/invasion in prostate cancer cells. *Cancer Res* 62(23):6812-6816 .

29. Rubin MA*, et al.* (2002) alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA* 287(13):1662-1670 .

30. Varambally S*, et al.* (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 419(6907):624-629 .

31. Kleer CG*, et al.* (2003) EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A* 100(20):11606-11611 .

32. Klezovitch O*, et al.* (2004) Hepsin promotes prostate cancer progression and metastasis. *Cancer Cell* 6(2):185-195 .

33. Stephan C*, et al.* (2004) Hepsin is highly over expressed in and a new candidate for a prognostic indicator in prostate cancer. *J Urol* 171(1):187-191 .

34. Luo J*, et al.* (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* 61(12):4683-4688 .

35. Welsh JB*, et al.* (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* 61(16):5974-5978 .
36. Sreekumar A*, et al.* (2004) Humoral immune response to alpha-methylacyl-CoA racemase and prostate cancer. *Journal of the National Cancer Institute* 96(11):834-843 .
37. Rogers CG*, et al.* (2004) Prostate cancer detection on urinalysis for alpha methylacyl coenzyme a racemase protein. *J Urol* 172(4 Pt 1):1501-1503 .
38. Yu J*, et al.* (2007) A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Res* 67(22):10657-10663 .
39. Varambally S*, et al.* (2008) Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science* 322(5908):1695-1699 .
40. Singh D*, et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1(2):203-209 .
41. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, & Gerald WL (2004) Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* 113(6):913-923.
42. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, & Chinnaiyan AM (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 62(15):4427-4433.
43. LaTulippe E*, et al.* (2002) Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* 62(15):4499-4506 .
44. Tomlins SA*, et al.* (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* 448(7153):595-599 .
45. Tomlins SA*, et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39(1):41-51.
46. MacDonald JW & Ghosh D (2006) COPA--cancer outlier profile analysis. *Bioinformatics* 22(23):2950-2951 .
47. Kumar-Sinha C, Tomlins SA, & Chinnaiyan AM (2008) Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* 8(7):497-511 .
48. Tomlins SA*, et al.* (2009) ETS gene fusions in prostate cancer: from discovery to daily clinical practice. *European urology* 56(2):275-286 .
49. Tomlins SA*, et al.* (2008) Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia* 10(2):177-188 .
50. Myllykangas S & Knuutila S (2006) Manifestation, mechanisms and mysteries of gene amplifications. *Cancer Lett* 232(1):79-89 .
51. Futreal PA*, et al.* (2004) A census of human cancer genes. *Nat Rev Cancer* 4(3):177-183 .
52. Vogel CL & Franco SX (2003) Clinical experience with trastuzumab (herceptin). *Breast J* 9(6):452-462 .
53. Bubendorf L*, et al.* (1999) Survey of gene amplifications during prostate cancer progression by high-throughout fluorescence in situ hybridization on tissue microarrays. *Cancer Res* 59(4):803-806 .

54. Nupponen NN, Kakkola L, Koivisto P, & Visakorpi T (1998) Genetic alterations in hormone-refractory recurrent prostate carcinomas. *Am J Pathol* 153(1):141-148 .

55. Sircar K, *et al.* (2009) PTEN genomic deletion is associated with p-Akt and AR signalling in poorer outcome, hormone refractory prostate cancer. *J Pathol* 218(4):505-513 .

56. Visakorpi T, *et al.* (1995) Genetic changes in primary and recurrent prostate cancer by comparative genomic hybridization. *Cancer Res* 55(2):342-347 .

57. Han G, *et al.* (2005) Mutation of the androgen receptor causes oncogenic transformation of the prostate. *Proc Natl Acad Sci U S A* 102(4):1151-1156 .

58. Linja MJ, *et al.* (2001) Amplification and overexpression of androgen receptor gene in hormone-refractory prostate cancer. *Cancer Res* 61(9):3550-3555 .

59. Chen CD, *et al.* (2004) Molecular determinants of resistance to antiandrogen therapy. *Nat Med* 10(1):33-39 .

60. Tran C, *et al.* (2009) Development of a second-generation antiandrogen for treatment of advanced prostate cancer. *Science* 324(5928):787-790 .

61. O'Mahony OA, *et al.* (2008) Profiling human androgen receptor mutations reveals treatment effects in a mouse model of prostate cancer. *Mol Cancer Res* 6(11):1691-1701 .

62. Steinkamp MP, *et al.* (2009) Treatment-dependent androgen receptor mutations in prostate cancer exploit multiple mechanisms to evade therapy. *Cancer Res* 69(10):4434-4442 .

63. Rubin MA & De Marzo AM (2004) Molecular genetics of human prostate cancer. *Mod Pathol* 17(3):380-388 .

64. Gurel B, *et al.* (2008) Molecular alterations in prostate cancer as diagnostic, prognostic, and therapeutic targets. *Advances in anatomic pathology* 15(6):319-331 .

65. Pinkel D, *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 20(2):207-211 .

66. Watson JE, *et al.* (2004) 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(19):3487-3494 .

67. Rubin MA, *et al.* (2004) Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res* 64(11):3814-3822 .

68. van Duin M, *et al.* (2005) Construction and application of a full-coverage, high-resolution, human chromosome 8q genomic microarray for comparative genomic hybridization. *Cytometry A* 63(1):10-19 .

69. van Duin M, *et al.* (2005) High-resolution array comparative genomic hybridization of chromosome arm 8q: evaluation of genetic progression markers for prostate cancer. *Genes Chromosomes Cancer* 44(4):438-449 .

70. Brookman-Amissah N, *et al.* (2005) 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(4):335-343 .

71. Saramaki OR, Porkka KP, Vessella RL, & Visakorpi T (2006) Genetic aberrations in prostate cancer by microarray analysis. *Int J Cancer* 119(6):1322-1329 .
72. Kim JH*, et al.* (2007) Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res* 67(17):8229-8239 .
73. Pollack JR*, et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 99(20):12963-12968.
74. Tomlins SA*, et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310(5748):644-648 .
75. Watson SK*, et al.* (2009) Minimum altered regions in early prostate cancer progression identified by high resolution whole genome tiling path BAC array comparative hybridization. *Prostate* 69(9):961-975 .
76. Demichelis F*, et al.* (2009) Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosomes Cancer* 48(4):366-380 .
77. Lapointe J*, et al.* (2007) Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. *Cancer Res* 67(18):8504-8510 .
78. Zhao H*, et al.* (2005) Genome-wide characterization of gene expression variations and DNA copy number changes in prostate cancer cell lines. *Prostate* 63(2):187-197 .
79. Paris PL*, et al.* (2004) Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Hum Mol Genet* 13(13):1303-1313 .
80. Nelson WG*, et al.* (2007) Abnormal DNA methylation, epigenetics, and prostate cancer. *Front Biosci* 12:4254-4266 .
81. Enokida H*, et al.* (2005) Ethnic group-related differences in CpG hypermethylation of the GSTP1 gene promoter among African-American, Caucasian and Asian patients with prostate cancer. *Int J Cancer* 116(2):174-181 .
82. Mehrotra J*, et al.* (2008) Quantitative, spatial resolution of the epigenetic field effect in prostate cancer. *Prostate* 68(2):152-160 .
83. Nonn L, Ananthanarayanan V, & Gann PH (2009) Evidence for field cancerization of the prostate. *Prostate* 69(13):1470-1479 .
84. Cho NY, Kim JH, Moon KC, & Kang GH (2009) Genomic hypomethylation and CpG island hypermethylation in prostatic intraepithelial neoplasm. *Virchows Arch* 454(1):17-23 .
85. Meiers I, Shanks JH, & Bostwick DG (2007) Glutathione S-transferase pi (GSTP1) hypermethylation in prostate cancer: review 2007. *Pathology* 39(3):299-304 .
86. Henrique R*, et al.* (2006) Hypermethylation of Cyclin D2 is associated with loss of mRNA expression and tumor development in prostate cancer. *J Mol Med* 84(11):911-918 .
87. Jeronimo C*, et al.* (2004) A quantitative promoter methylation profile of prostate cancer. *Clin Cancer Res* 10(24):8472-8478 .
88. Hoque MO (2009) DNA methylation changes in prostate cancer: current developments and future clinical implementation. *Expert Rev Mol Diagn* 9(3):243-257 .

89. Pfeifer GP & Dammann R (2005) Methylation of the tumor suppressor gene RASSF1A in human tumors. *Biochemistry (Mosc)* 70(5):576-583 .

90. Ellinger J*, et al.* (2008) CpG island hypermethylation at multiple gene sites in diagnosis and prognosis of prostate cancer. *Urology* 71(1):161-167 .

91. Richiardi L*, et al.* (2009) Promoter methylation in APC, RUNX3, and GSTP1 and mortality in prostate cancer patients. *J Clin Oncol* 27(19):3161-3168 .

92. Alumkal JJ*, et al.* (2008) Effect of DNA methylation on identification of aggressive prostate cancer. *Urology* 72(6):1234-1239 .

93. Manoharan M, Ramachandran K, Soloway MS, & Singal R (2007) Epigenetic targets in the diagnosis and treatment of prostate cancer. *Int Braz J Urol* 33(1):11-18 .

94. Chuang JC*, et al.* (2005) Comparison of biological effects of non-nucleoside DNA methylation inhibitors versus 5-aza-2'-deoxycytidine. *Mol Cancer Ther* 4(10):1515-1520 .

95. Lodygin D, Epanchintsev A, Menssen A, Diebold J, & Hermeking H (2005) Functional epigenomics identifies genes frequently silenced in prostate cancer. *Cancer Res* 65(10):4218-4227 .

96. Yegnasubramanian S*, et al.* (2008) DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. *Cancer Res* 68(21):8954-8967 .

97. Cho NY*, et al.* (2007) Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. *J Pathol* 211(3):269-277 .

98. Okitsu CY, Hsieh JC, & Hsieh CL (Transcriptional Activity Affects the H3K4me3 Level and Distribution in the Coding Region. *Molecular and cellular biology* .

99. Li X*, et al.* (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell* 20(2):259-276 .

100. Lindroth AM*, et al.* (2008) Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS Genet* 4(8):e1000145 .

101. Kondo Y*, et al.* (2008) Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet* 40(6):741-750 .

102. Abate-Shen C & Shen MM (2000) Molecular genetics of prostate cancer. *Genes Dev* 14(19):2410-2434 .

CHAPTER 2

GENOME-WIDE COPY NUMBER CHANGES IN PROSTATE CANCER UNRAVELED USING ARRAY COMPARATIVE GENOMIC HYBRIDIZATION (aCGH)

Chromosomal aberrations due to genome instability are a characteristic of human solid tumors (1) and are considered the primary drivers in the development and progression of cancer (2). Precise measurements of gene copy number alterations with high resolution are now possible with array CGH (aCGH) performed on BAC arrays, cDNA microarrays, or oligoCGH arrays (3). Several tumors including breast, prostate, and lung cancers among others have been analyzed using aCGH technology (4-5). More recently, studies have been documenting genome-wide copy number changes with parallel mRNA expression profiling for various cancers using microarray platforms (6-10).

aCGH analyses of human prostate cancer cell lines (11-13), xenografts(14-15) and prostate cancer tissues (16-17) have been reported. While all of the above studies have used grossly dissected tissues, profiling of laser captured, microdissected prostate cancer specimens has been shown to resolve cancer specific genomic aberrations with higher sensitivity (17), Recently Hughes *et al*. reported aCGH profiling of a small set of laser captured prostate cancer specimens (18). In this chapter, we carried out a comprehensive characterization of cytogenetic profiles of 62 prostate cell populations using aCGH on a cDNA microarray platform as described by Pollack and colleagues (19). Cells from specific prostate tissue foci were isolated by laser capture microdissection and the samples belonged to various groups that include: benign prostatic hyperplasia (BPH), stromal, atrophic epithelia, proliferative inflammatory atrophy (PIA) (20), postatrophic hyperplasia (PAH) (21), prostatic intraepithelial neoplasia (PIN), clinically localized prostate cancer (PCA) low-grade, L-PCA (Gleason pattern 3); foamy, F-PCA; high-grade, H-PCA (Gleason pattern 4); and metastatic prostate cancer (MET). Our group has recently noted the enrichment of various molecular concepts in gene expression signature of prostate cancer progression using this sample set (22). In the present study, we defined the minimal common regions (MCRs) corresponding to various sample groups and identified novel regions of aberrations and candidate genes that lie within. Our study also provides information on the occurrence of these MCRs in successive stages of cancer progression.

Results and Discussion:

Chromosomal aberrations in prostate cancer

We used laser capture microdissection (LCM) to isolate 62 specific cell populations from 38 patients representing a histopathologic spectrum of prostate cancer progression to perform aCGH analysis (**Table 2.1**). This thesis work describes the results from the aCGH (in this chapter) and its integrative analysis with gene expression data in next chapter, which has been reported previously (22). Analysis of the array CGH data using CGH-Miner software identified significantly altered contiguous chromosomal regions (q-value<0.01) within each sample (23) (and Materials and Methods). The total number of genes located within these altered regions was highest (average n=800) in metastatic samples (**Fig. 2.1** and **Table 2.1**) while it was lowest in the benign samples (average n=61). On average, high-grade PIN (HGPIN) samples had 194 gene alterations, while 151 and 360 altered genes were found in low- and high- grade localized prostate cancer samples, respectively. PAH and atrophy had 74 alterations on average, and the range of the number of observed chromosomal aberrations in these samples was not significantly different (p-value>0.05) from benign/normal samples. However in both HGPIN and cancer cases, the variability in the number of alterations compared to that of benign/normal samples was significantly different (HGPIN p-value<0.0005, H-PCA p-value<0.0003, and MET p-value<0.0001).

To calibrate the resolution of our array CGH technique, we used genomic DNA samples with varying copies (1 to 5) of chromosome X hybridized against normal human male genomic DNA obtained from a commercial source. The gain in copy number was evident with increasing signal for all genes derived from chromosome X (**Fig. 2.2**). Using this data, when the mean fluorescence ratios of the genes located in X chromosome from each experiment was plotted, the slope was 0.2228 with $R^2 = 0.9998$ (data not shown). In addition, genomic DNA from the epithelium and stroma of normal prostate tissue from a cadaveric donor, previously diagnosed with Down's syndrome, showed a single copy gain in chromosome 21 (**Fig. 2.2**). Resolution of the X chromosome copy number changes, as well as the detection of a single copy gain in chromosome 21, validated the performance of our arrays and the amplification technique employed with laser captured specimens.

The significantly altered genes from various prostate samples identified in the above analysis (**Fig. 2.1**) were ordered according to human genome map position from chromosome 1 to Y and the regions of gains and losses were displayed as a heatmap (**Fig. 2.2**). Benign samples, as expected, showed no significant regions of alteration (excluding the two samples with chromosome 21 amplification), while metastatic samples displayed the most alterations. Among the metastatic prostate cancer samples, frequent amplifications were observed in chromosomal arms 2p, 3q, 7, 8q, 9q, 16p, and 20q, and deletions in 6q, 8p, 10q, 13q, and 18q (**Fig. 2.2**). When the size of the altered span was considered, the H-PCA (Gleason Pattern 4) samples showed more resemblance to the MET samples than to L-PCA (Gleason Pattern 3), and the alteration sites tended to extend further, even encompassing entire chromosomal arms, for example in 3q, 8q, and 13 (**Fig. 2.2**). Next we calculated the percent alteration in each group, and the alteration frequency is displayed for all the chromosomes (**Fig. 2.3**). The aberrations observed in PIN, L-PCA, H-PCA, and MET groups in chromosomes 8, 10 and 13 revealed a distinct overlap (**Fig. 2.4).** Several known alterations, including amplifications in *TPD52* and *MYC* (8q21.13 and 8q24.22) and deletion in *PTEN* (10q23.3) are located within the highly altered regions on these chromosomes. For 8q24.22 and 10q23.3, the percent alteration in MET and H-PCA ranged from 30 to 50%, while 8q21.13 had a 40% alteration in MET samples. Among the PIN samples, the whole-arm amplification in 8q, which is often observed in the advanced form of the disease, was present in at least one sample. None of our other PIN samples had chromosomal-arm spanning alterations, although there were a number of smaller altered sites throughout the entire chromosome. Besheshti *et al.* have previously reported that more extensive aberrations are observed in

PCA than in PIN, and in contrast to 8q gain that was consistently observed from different tumor foci, 8q gain in PIN is not a common event (17). A recent prostate cancer aCGH study by Hughes *et al.* on a BAC platform consisting of 2,400 clones identified both 8q21.11-qter gain and 8p11.23-p23.3 loss in PIN and PCA samples, with gain in more than 37.5 and 50% and loss in more than 50% of PIN and PCA samples, respectively (sample size n = 7 for PIN and 8 for PCA) (18). Aberrations in chromosome 8 were not very common in our PIN sample cohort, with only 1 case of 8p loss and 1 case of 8q gain. However gain of 8q and loss of 8p were frequent events (40% and 30%, respectively) in the tumor samples.

Minimal common regions in prostate cancer

In order to refine the regions of alteration obtained from CGH-Miner output, we used MCR identification as described in the Materials and Methods. The automated algorithm we used has been previously applied to define overlapping regions of amplification, deletion, and focal regions of recurred alterations in myeloma, pancreatic, and lung cancer aCGH studies (9-10, 24). By using this method we identified minimal common regions (MCRs) in PCA and MET sample groups in our prostate cancer dataset (**Table 2.2**). More detailed information on all MCRs, including precise aberration sites, cytogenetic bands, number of samples altered and the gene names within the predicted regions that meet the cutoff threshold for amplification and deletion, along with the over- and under-expressed genes from matching mRNA profiles is provided (see **APPENDIX** to download).

The region 8q24.22 (132.98-134.64 Mb) had the highest percentage of alterations (50%) among the METs. Other frequently amplified regions with more than 40% of MET samples showing the alterations are, 2p23.3 (24.06-24.34 Mb), 7p15.1 (29.48-31.5 Mb), 8q21.13 (31.6-81.62 Mb), 8q22.2 (99.21-99.65 Mb), 8q22.2-q22.3 (101.23-101.99 Mb), 8q22.3 (102.57-103.29 Mb). Regions 10q23.31 (90.96-91.08 Mb), 18q21.31 (53.37-54.17 Mb), 18q21.32 (54.68-55.25 Mb), and 18q22.1 (63.32-64.54 Mb) were among the frequently deleted regions exhibited by 40% of MET samples. The region 18q21.2 which harbors *SMAD4* is deleted in 30% of MET samples. SMAD4, a TGF beta superfamily signaling molecule is significantly under-expressed in prostate cancer (25).

In PCA samples, regions 5q32 with 30% amplification and 13q14.12 with 30% deletion were the most frequently altered sites. Other alterations which include amplification in 5q32 and 8q24.11 and deletion in 5q13.3, 6q14.3-q15, 10q23.2-q24.1, 13q14.12-q14.2, 16p12.2, and 17q21.31 were observed with a frequency greater than 20%. In H-PCA samples, the deletion in 13q14.2 (47.55-48.9 Mb) was observed in more than 40% of the cases. Previously, van Dekken *et al.* compared Gleason pattern 3 and 4 tumors obtained from same cases by 2,400-element BAC array, and a 34% overlap in genomic aberrations, mainly in deleted regions was reported (26). The defined MCRs for H-PCA and L-PCA samples in our cohort show overlap in deleted regions 5q13.3, 6q14.2-q21, and 16p12.2; however, there were no overlapping regions of amplification between the two groups. This could be due to the identification of only four altered sites by automated MCR definition in L-PCA.

The shared regions of alterations among PIN, L-PCA, H-PCA, and MET samples were further investigated by the following method. If a given cytogenetic band shows 20%

recurrent aberration within a sample type, it is mapped to the same cytogenetic band region in other sample types to assess percent recurrent aberration in that region. Identifying these shared regions is of great interest as it might shed light on the mechanism of tumor progression, especially when the alteration is detected in the precursor lesions such as PIN and is preserved or becomes more frequent in other progressive stages of the cancer. The cytogenetic bands harboring the shared amplicons in all four groups are 1q21.1-q21.3, 1q24.1-q24.2, 3q21.1-q21.3, 3q29, 6p21.33-p21.1, 7p15.1-p14.3, 8q22.2-q24.12, 11p15.4-p15.1, 11q13.1, and 12p13.32-p13.2. In most of the PCA and PIN samples, the aberration within designated cytogenetic region was seen in less than 20% cases. Gain in 1q or 6p denotes poor outcome in melanomas (27), suggesting that alterations in those region may be associated with poor prognosis and lower survival rate in prostate cancer as well. All groups shared deletion at 6q16.2-q22.31, 13q12.12-q32.1, 17p12-p11.2, and 18q21.1-q23, whereas the deleted region in chromosome 13 is a frequent event in both MET and H-PCA cases (>30%). Among 13 defined MCRs in PIN, 9 are shared with PCA. These regions include amplifications in 3q29, 5q31.3-q32, 5q32, 6q27, and 8q24.3 and deletions in 6q22.31, 16p12.2, 17q21.2, and 17q21.31. 17q21.31 is known to be completely lost in the PC3 cell line (28). However, among 3 defined MCRs in PIA samples, none are shared with PIN or PCA. Among the putative precursor lesions, PIN, but not PIA, bears a closer resemblance to prostate cancer in copy-number alterations.

Overall, the most frequent amplifications are observed in 2p, 3q, 5q, 7p, 8q, 9q, 16p, and 20q for MET and 3q, 5q, 7q, and 8q for PCA. Deletion-prone sites in MET are

6q, 8p, 10q, 13q, and 18q, and they are 1q, 3q, 5q, 6q, 10q, 13q, 16p, 17q, and Xp in PCA, where most of the alterations are obtained from H-PCA samples.

Chromosomal aberrations in distant metastases

Several metastatic specimens from a single patient (case 34) that include lymph node, lung, liver, soft tissue around bone, other soft tissue, and tissue from the residual prostate gland were profiled in this study (**Table 2.1**). CGH-Miner output of significant alterations (q-value<0.01) for each sample was displayed for comparison (**Fig. 2.5**). Common alterations that occur in at least 3 samples are detected in 3q, 5q, 11q, 12p, 13q, 16q23-q24, 17p, 18q, and 22q (**Fig. 2.5**). The majority of the alteration sites are overlapping; however, there are also unshared altered sites present, likely accounted for by the heterogeneity in each clonal group as well as due to the alterations that occur after metastasis. The number of genes within the identified aberrant sites ranges from 583 to 1723, and the residual carcinoma of prostate gland and lymph node MET displayed lower number of altered genes compared to the other sites. Interestingly, in an earlier integrative mRNA and copy-number study on chromosome 16q, deletion in 16q23.1 to 16qter (a region harboring many candidate tumor suppressor genes) was reported in more than 50% of prostate cancer samples examined (29). Many other genes known to be deleted in prostate cancer such as, *CYP1B1*, *TNFRSF10B*, *ATAD1*, and *PTEN* are located within the commonly deleted regions in these distant metastasis samples.

In conclusion, aCGH analysis of laser-capture microdissected prostate cancer samples detected a multitude of chromosomally altered regions through the various stages of prostate cancer progression. Samples like PAH and PIA, were characterized for the first time by aCGH in this study. Minimal common regions were identified and the percentage of alterations in various prostate cancer stages that reflect the entire spectrum of the disease progression was calculated. This generated a list of altered regions and candidate genes that might play a role in cancer progression. The prostate cancer precursor lesion PIN resembled PCA in its genetic alterations. The complete prognosis information for this patient cohort is still being collected and is incomplete at the moment. Extensive analysis will be performed when this information is made available at a future date.

Materials and Methods:

Tissue specimen and genomic DNA isolation

Tissues were obtained from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program, which are both part of University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core. All samples were collected with informed consent of the patients and prior institutional review board approval. The prostate cancer samples obtained from a total of 38 patients/organ donors include 7 normal/benign prostatic hyperplasia, BPH; 8 stromal, S; 5 postatrophic hyperplasia, PAH (2 atrophic epithelium, ATR; 3 proliferative inflammatory atrophy, PIA), 7 prostatic intraepithelial neoplasia, PIN; 18 localized prostate cancer, PCA (8 low-grade PCA, L-PCA; 1 foamy PCA, F-PCA; 9 high-grade PCA, H-PCA) and 17 metastatic prostate cancer, MET (**Table 2.1**). A precision cut using laser-capture microdissection was performed on frozen tissue sections (6 microns) containing a minimum of 10,000 cells placed on specially manufactured membrane slides (MMI, Knoxville, TN) with the SL Microtest device (MMI, Knoxville, TN) using u(micro)CUT software (MMI, Knoxville, TN) (22). Genomic DNA was isolated from the cells using QIAamp DNA Mini Kit (Qiagen, Valencia, CA) and DNA concentration was determined using Quant-iT DNA Assay Kit, High Sensitivity (Invitrogen, Carlsbad, CA). For the threshold analysis, human genomic DNA samples having varying copies (1 to 5) of the X chromosome were purchased from the NIGMS Human Genetic Mutant cell repository (http://www.nigms.nih.gov/Initiatives/HGCR/). Normal human male and female genomic DNA was purchased from Promega Inc.

Array-CGH on cDNA microarrays

In-house cDNA microarrays containing 20,000 spotted elements representing ~13,000 different UniGene clusters used in our previous gene expression profiling studies (22, 30) were used for the aCGH studies. One hundred nanograms of genomic DNA was amplified using OmniPlex Whole Genome Amplification (WGA) kit (Sigma-Aldrich, St. Louis, MO) following the manufacturer's protocol. Amplified normal human male genomic DNA was used as reference for all the hybridizations. The amplified DNA was quantified by Quant-iT DNA Assay Kit, High Sensitivity (Invitrogen, Carlsbad, CA) and 4 µg of DNA from each sample was labeled using BioPrime® Array CGH Genomic Labeling System (Invitrogen, Carlsbad, CA). Two color hybridizations were performed as described earlier by Pollack *et al.* (19) The use of whole genome amplification was previously evaluated on frozen and formalin-fixed, paraffin-embedded (FFPE) Wilm's tumor specimens on aCGH platform by Little *et al* (31).

Data collection and analysis

cDNA microarray slides were scanned using an Axon GenePix 4000B dual-laser scanner, and its fluorescence signal was quantified with GenePix Pro 6.0 software (Axon Instruments, Union City, CA). Bad spots were flagged out, and data was Lowess normalized (32). Genes with multiple representations were averaged using GEPAS software (33), and $\log_2$-transformed. A total of 9,550 unique genes were analyzed as follows. Cutoff values were set at $\log_2$ ratio ≥ 0.22 for amplification and ≤ -0.22 for deletion (97% and 3% quantiles, respectively); the cutoff values for complementary expression profiling data were set at $\log_2$ ratio of ≥ 0.4 for over-expression of genes and ≤ -0.4 for under-expression (± 4 standard deviations of the middle 50% quantile of data) (9-10, 24). For the genome-wide integrative analyses, the data from aCGH and gene expression microarrays was moving averaged (symmetric 5-nearest neighbors) using CGH-Miner software (23). The CGH-Miner output for 9,550 unique genes was ordered according to the genome map positions from chromosome 1 to Y, and the moving averaged (symmetric 5-nearest neighbors) fluorescence ratios were depicted using $\log_2$-based pseudocolor scale.

Percentage of Alterations in Prostate Cancer

Prostate cancer data represented in global view analysis was classified into 4 different sample groups: PIN, L-PCA (Gleason Pattern 3), H-PCA (Gleason pattern 4), and MET. From the region of chromosomal aberrations, the amplified and deleted genes were selected using the thresholds set above. Among the samples profiled in the metastatic group, data from six hormone refractory tumors obtained from different metastatic sites from a single patient (case number 34) was averaged to reduce sample bias. Three additional samples (2 MET and one PIN) excluded in the global view analysis (as transcript profiling was not available) were included here, taking the total number of samples in the MET group to twelve and PIN group to seven. The percentage of alterations for the selected genes was calculated for each group. The gene list was ordered according to the chromosomal location of each gene and was moving averaged (n=5) for graphical representation. The residual prostate carcinoma obtained from case 34 was included in the metastatic group in all of our analyses as the gene expression analysis clustered this sample in the metastatic group. This sample had 851 alterations, which is in the range of alterations observed in metastatic samples from case 34 (456-1588 altered genes).

Minimal Common Region (MCR) characterization

MCR characterization was performed as described earlier (9-10, 24), with some modifications. A Perl-based algorithm was applied to the normalized data. Genes with $\log_2$ ratios greater or less than the predefined cutoff values (as described above in Data collection and analysis), within the significantly altered regions identified by CGH miner, were considered as altered. The CGH-miner output arranges genes according to their chromosomal location. Samples were grouped into 6 categories (PIA, PIN, L-PCA, H-PCA, PCA, MET). To identify most commonly amplified or deleted genes, a score was given to each gene based on the number of samples with alteration. We then scanned the scores to identify contiguous spans of altered genes having at least 75% of the peak alteration percentage to denote the MCRs.

Notes

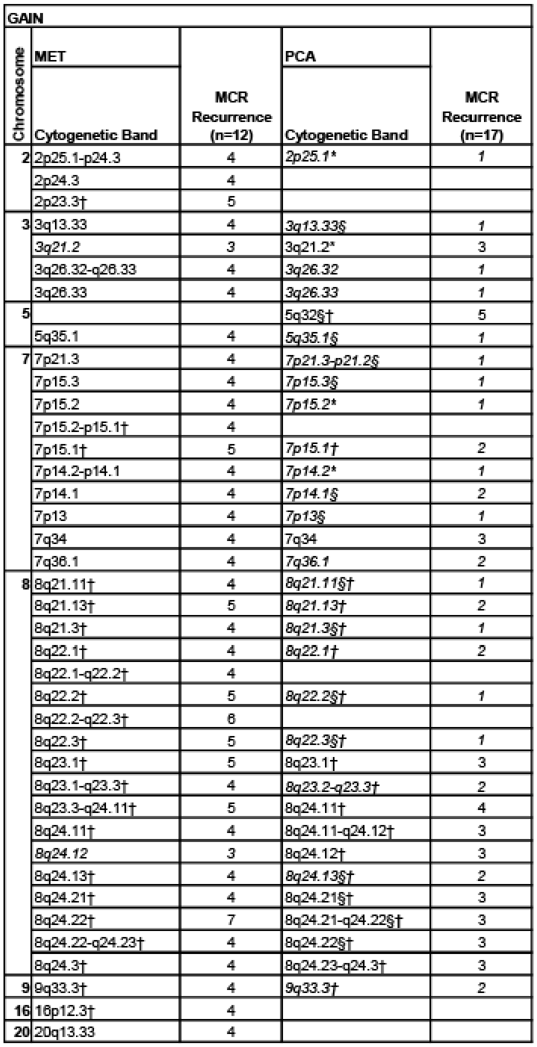
This chapter has been previously published: Kim JH, *et al.* (2007) Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res* 67(17):8229-8239. All rights reserved.

Table 2.1 Clinical Information of LCM samples profiled by aCGH**.** AA = African American, C = Caucasian, UO = Unknown/Others

Hybridization	Class	Ages	Race	Sample Gleason	LCM Gleason	# of Alterations	Case Number	Comments
***NORMAL**	Normal Epithelium - Organ Donor	46	C			124	29	Down's Syndr; Intracranial hemorrhage
NORMAL	Normal Epithelium - Organ Donor	21	AA			29	30	GSW
NORMAL	Normal Epithelium - Organ Donor	22	C			81	31	Subarrachn hemorrhage
NORMAL	Normal Prostate Glands	64	UO			71	13	
BPH	BPH Epithelium	64	UO			90	13	
BPH	BPH Epithelium	65	C			98	14	
BPH	BPH Epithelium					31	23	
Atrophy	Atrophic Epithelium - Simple Cystic Atrophy	57	C			94	11	
Atrophy	Atrophic Epithelium - Simple Atrophy	73	C			52	4	
PIA	Proliferative inflammatory atrophy (PIA)	41	C			99	7	
PIA	Proliferative inflammatory atrophy (PIA)	55	C			79	26	
PIA	Proliferative inflammatory atrophy (PIA)	61	C			45	27	
***NORMAL-Stroma**	Normal Stroma - Organ Donor	46	C			128	29	Down's Syndr; Intracranial hemorrhage
NORMAL-Stroma	Normal Stroma - Organ Donor	21	AA			50	30	GSW
NORMAL-Stroma	Normal Stroma - Organ Donor	22	C			105	31	Subarrachn hemorrhage
BPH-Stroma	Stromal BPH nodule	63	C			75	6	
BPH-Stroma	Stromal BPH nodule	52	UO			58	8	
BPH-Stroma	Stromal BPH nodule	70	C			35	24	
BPH-Stroma	Stroma - Epithelial BPH					39	23	
	Normal Stroma - Adjacent Tumor (NAS)	64	C			26	17	Stroma of stroma tumor interface
PIN	Prostatic Intraepithelial Neoplasia	76	C			232	5	Distant to PCA (Gleason 3+3)
PIN	Prostatic Intraepithelial Neoplasia	64	UO			76	13	
PIN	Prostatic Intraepithelial Neoplasia	65	C			72	14	
PIN	Prostatic Intraepithelial Neoplasia	72	AA			52	18	
PIN	Prostatic Intraepithelial Neoplasia	68	C			126	21	
PIN	Prostatic Intraepithelial Neoplasia	55	C			603	26	
PIN	Prostate Carcinoma	69	C				9	
†F-PCA	Prostate Carcinoma	41	C			49	7	Foamy PCA
†L-PCA	Prostate Carcinoma	55	C	3+3	3	178	2	
†L-PCA	Prostate Carcinoma	53	C	3+3	3	56	3	
†L-PCA	Prostate Carcinoma	76	C	3+3	3	87	5	
L-PCA	Prostate Carcinoma	65	C	3+3	3	200	10	
†L-PCA	Prostate Carcinoma	51	C	3+3	3	118	20	
L-PCA	Prostate Carcinoma	72	C	3+4	3	198	16	
L-PCA	Prostate Carcinoma			3+4	3+4	170	15	
L-PCA	Prostate Carcinoma	69	C	4+3	3	199	9	
H-PCA	Prostate Carcinoma	72	C	3+4	4	623	16	
H-PCA	Prostate Carcinoma	69	C	4+3	4	413	9	
‡H-PCA	Prostate Carcinoma	55	C	4+4	4	145	12	
‡H-PCA	Prostate Carcinoma	64	C	4+4	4	548	22	Gleason 4+4 with tertiary 5
†H-PCA	Prostate Carcinoma	65	C	4+4	4	271	14	
†H-PCA	Prostate Carcinoma	65	C	4+5	4	80	14	
†H-PCA	Prostate Carcinoma	65	C	4+5	4+5	60	14	
†H-PCA	Prostate Carcinoma	67	C	4+5	4+5	559	19	
H-PCA	Metastatic Prostate Carcinoma	56	C	4+5	4+5	542	25	
†MET	Metastatic Prostate Carcinoma					569	1	Lymph node
MET	Metastatic Prostate Carcinoma					852	37	Lymph node
MET	Metastatic Prostate Carcinoma	65	C			896	32	Liver
MET	Metastatic Prostate Carcinoma	76	C			918	33	Liver
‡MET	Metastatic Prostate Carcinoma					519	34	Lung
‡MET	Metastatic Prostate Carcinoma					456	34	Lymph node
‡MET	Metastatic Prostate Carcinoma					1588	34	Soft tissue
‡MET	Metastatic Prostate Carcinoma					1220	34	Soft tissue around bone
‡MET	Metastatic Prostate Carcinoma					851	34	Residual carcinoma in the prostate
‡MET	Metastatic Prostate Carcinoma					1080	34	Liver
MET	Metastatic Prostate Carcinoma	53	C			190	35	Soft tissue
MET	Metastatic Prostate Carcinoma	53	C			869	35	Soft tissue
MET	Metastatic Prostate Carcinoma	53	C			961	35	Soft tissue
†MET	Metastatic Prostate Carcinoma	71	C			479	36	Liver
†MET	Metastatic Prostate Carcinoma	71	C			1112	36	Soft tissue
MET	*Metastatic Prostate Carcinoma*	*66*	*C*				*38*	*Liver*
MET	*Metastatic Prostate Carcinoma*	*66*	*C*				*28*	*Soft tissue*

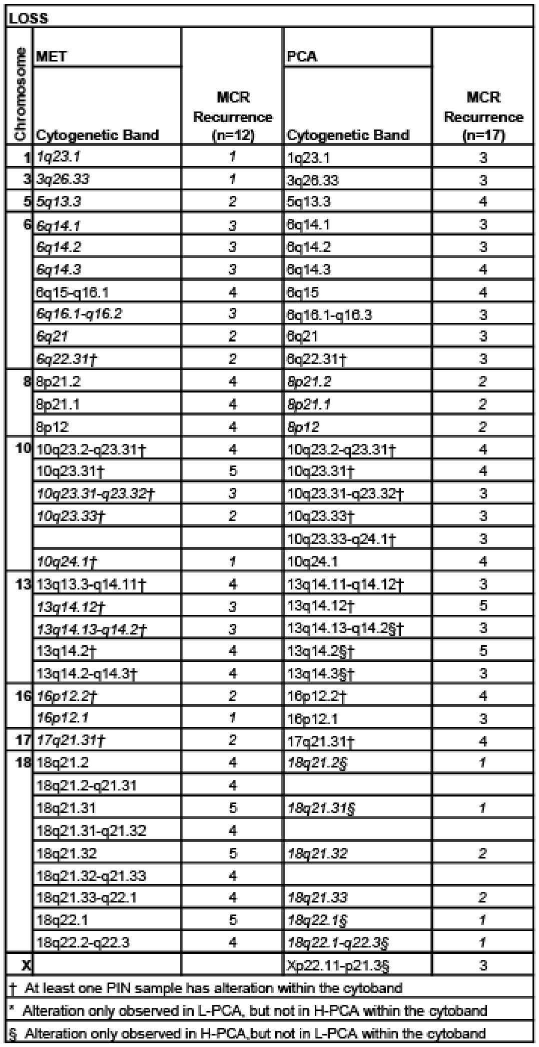
*Italic: Samples without transcript data *Trisomy in Chromosome 21*
†ERG-Over-expression Samples ‡ETV1 Over-expression Samples

Table 2.2 The most frequently observed chromosomal alteration sites in prostate cancer progression. Using automated locus definition (Materials and Methods in CHAPTER2), the minimal common regions in localized and metastasized prostate cancer samples are defined. The cytogenetic bands where the located genes showing frequent amplification and deletion in MET and PCA are listed.



GAIN

Chromosome	MET	MCR Recurrence (n=12)	PCA	MCR Recurrence (n=17)
	Cytogenetic Band		Cytogenetic Band	
2	2p25.1-p24.3	4	_2p25.1*_	_1_
	2p24.3	4		
	2p23.3†	5		
3	3q13.33	4	_3q13.33§_	_1_
	_3q21.2_	_3_	3q21.2*	3
	3q26.32-q26.33	4	_3q26.32_	_1_
	3q26.33	4	_3q26.33_	_1_
5			5q32§†	5
	5q35.1	4	_5q35.1§_	_1_
7	7p21.3	4	_7p21.3-p21.2§_	_1_
	7p15.3	4	_7p15.3§_	_1_
	7p15.2	4	_7p15.2*_	_1_
	7p15.2-p15.1†	4		
	7p15.1†	5	_7p15.1†_	_2_
	7p14.2-p14.1	4	_7p14.2*_	_1_
	7p14.1	4	_7p14.1§_	_2_
	7p13	4	_7p13§_	_1_
	7q34	4	7q34	3
	7q36.1	4	_7q36.1_	_2_
8	8q21.11†	4	_8q21.11§†_	_1_
	8q21.13†	5	_8q21.13†_	_2_
	8q21.3†	4	_8q21.3§†_	_1_
	8q22.1†	4	_8q22.1†_	_2_
	8q22.1-q22.2†	4		
	8q22.2†	5	_8q22.2§†_	_1_
	8q22.2-q22.3†	6		
	8q22.3†	5	_8q22.3§†_	_1_
	8q23.1†	5	8q23.1†	3
	8q23.1-q23.3†	4	_8q23.2-q23.3†_	_2_
	8q23.3-q24.11†	5	8q24.11†	4
	8q24.11†	4	8q24.11-q24.12†	3
	_8q24.12_	_3_	8q24.12†	3
	8q24.13†	4	_8q24.13§†_	_2_
	8q24.21†	4	8q24.21§†	3
	8q24.22†	7	8q24.21-q24.22§†	3
	8q24.22-q24.23†	4	8q24.22§†	3
	8q24.3†	4	8q24.23-q24.3†	3
9	9q33.3†	4	_9q33.3†_	_2_
16	16p12.3†	4		
20	20q13.33	4		



LOSS

Chromosome	MET	MCR Recurrence (n=12)	PCA	MCR Recurrence (n=17)
	Cytogenetic Band		Cytogenetic Band	
1	_1q23.1_	_1_	1q23.1	3
3	_3q26.33_	_1_	3q26.33	3
5	_5q13.3_	_2_	5q13.3	4
6	_6q14.1_	_3_	6q14.1	3
	_6q14.2_	_3_	6q14.2	3
	_6q14.3_	_3_	6q14.3	4
	6q15-q16.1	4	6q15	4
	_6q16.1-q16.2_	_3_	6q16.1-q16.3	3
	_6q21_	_2_	6q21	3
	_6q22.31†_	_2_	6q22.31†	3
8	8p21.2	4	_8p21.2_	_2_
	8p21.1	4	_8p21.1_	_2_
	8p12	4	_8p12_	_2_
10	10q23.2-q23.31†	4	10q23.2-q23.31†	4
	10q23.31†	5	10q23.31†	4
	_10q23.31-q23.32†_	_3_	10q23.31-q23.32†	3
	_10q23.33†_	_2_	10q23.33†	3
			10q23.33-q24.1†	3
	_10q24.1†_	_1_	10q24.1	4
13	13q13.3-q14.11†	4	13q14.11-q14.12†	3
	_13q14.12†_	_3_	13q14.12†	5
	_13q14.13-q14.2†_	_3_	13q14.13-q14.2§†	3
	13q14.2†	4	13q14.2§†	5
	13q14.2-q14.3†	4	13q14.3§†	3
16	_16p12.2†_	_2_	16p12.2†	4
	_16p12.1_	_1_	16p12.1	3
17	_17q21.31†_	_2_	17q21.31†	4
18	18q21.2	4	_18q21.2§_	_1_
	18q21.2-q21.31	4		
	18q21.31	5	_18q21.31§_	_1_
	18q21.31-q21.32	4		
	18q21.32	5	_18q21.32_	_2_
	18q21.32-q21.33	4		
	18q21.33-q22.1	4	_18q21.33_	_2_
	18q22.1	5	_18q22.1§_	_1_
	18q22.2-q22.3	4	_18q22.1-q22.3§_	_1_
X			Xp22.11-p21.3§	3

† At least one PIN sample has alteration within the cytoband

* Alteration only observed in L-PCA, but not in H-PCA within the cytoband

§ Alteration only observed in H-PCA,but not in L-PCA within the cytoband

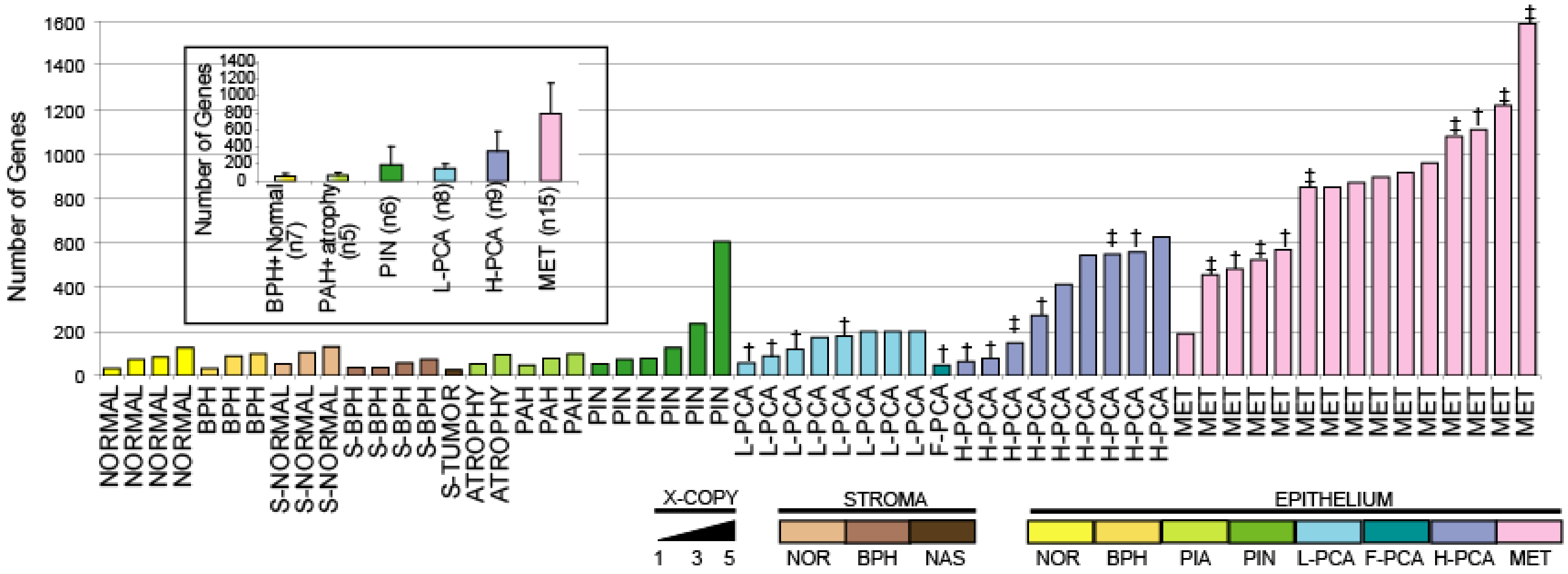


Figure 2.1 Number of altered genes in prostate cancer progression. The number of altered genes from significantly altered regions (q-value<0.01) in each sample is represented in a bar graph. Upper left inset represents averages of the data presented. † indicates ERG-overexpressing samples. ‡ indicates ETV1-overexpressing samples.

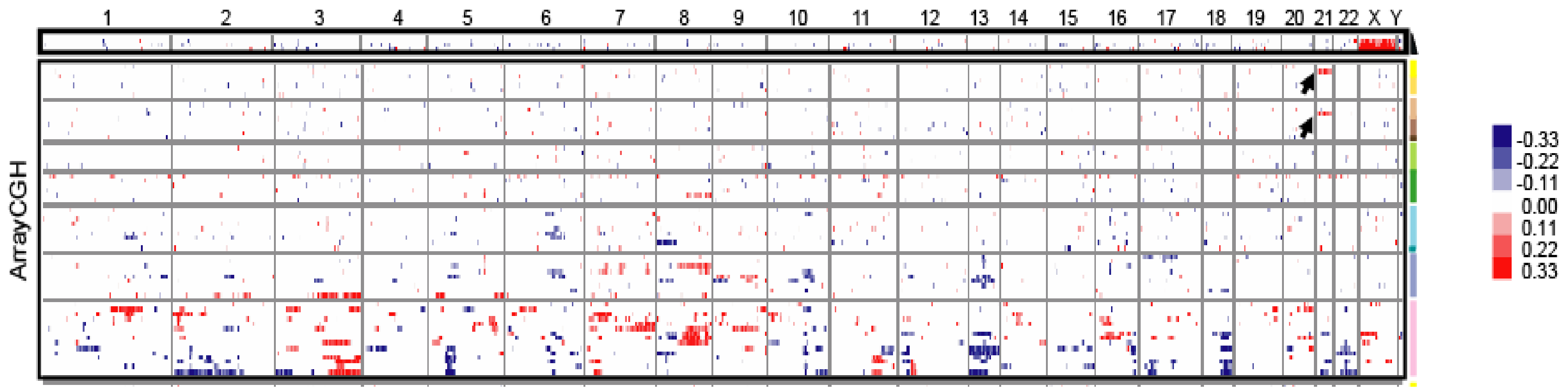
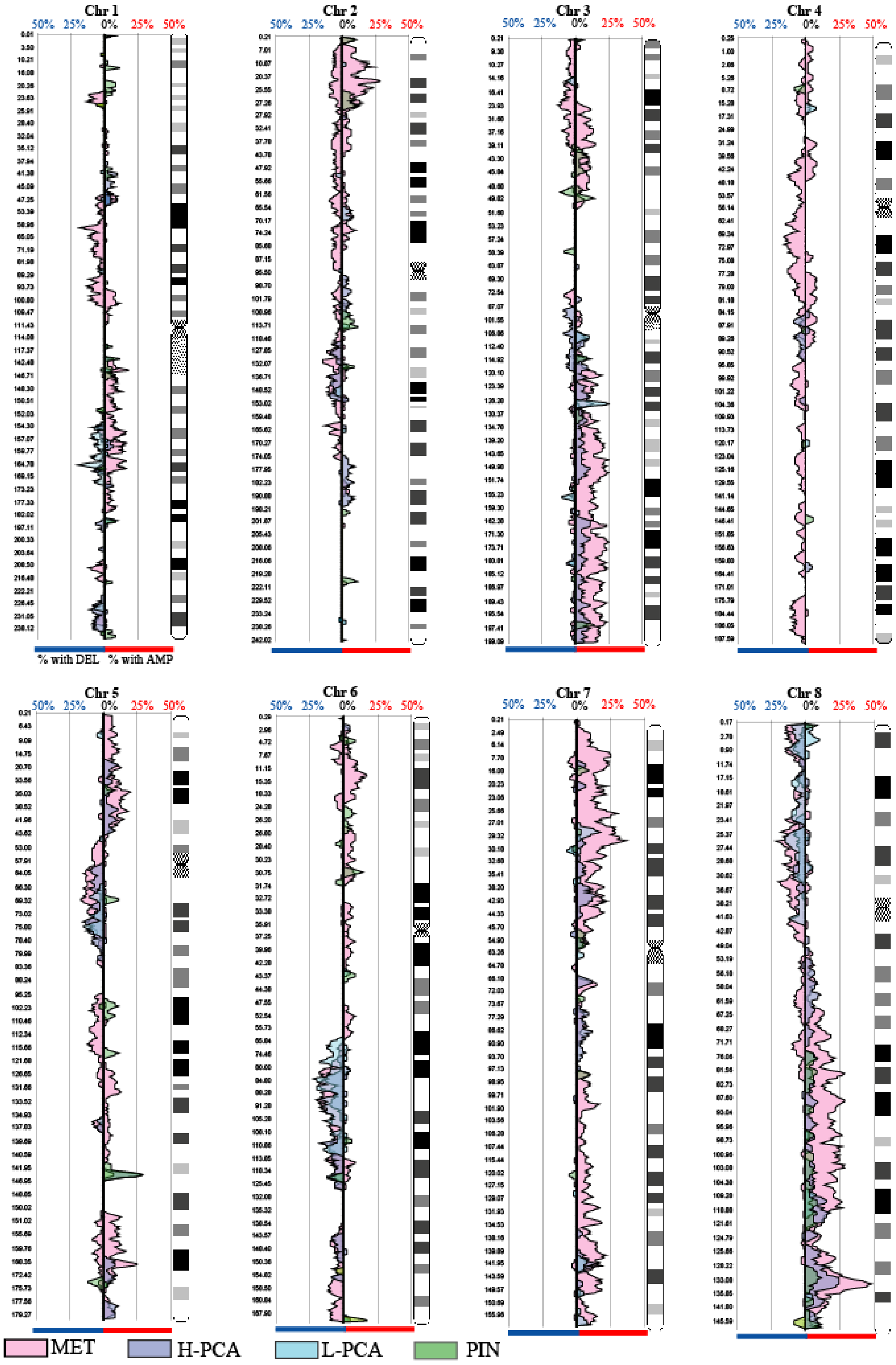
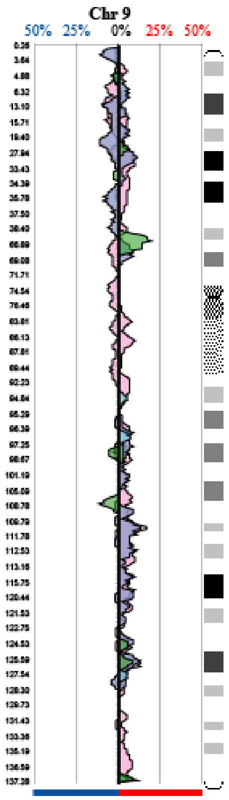
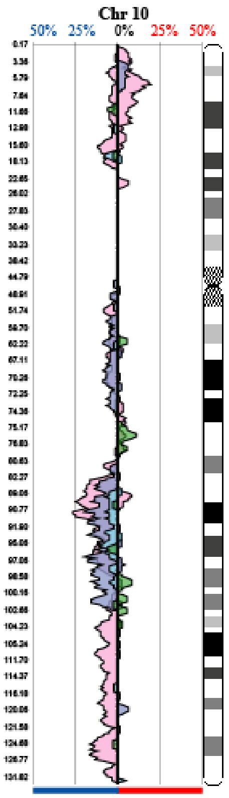
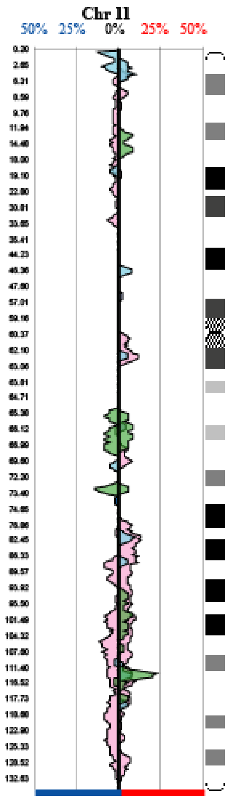


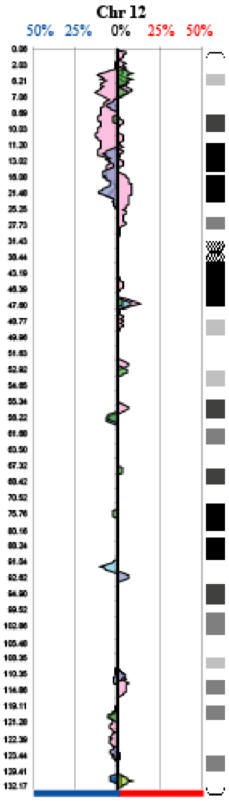
Figure 2.2 Genome-wide view of chromosomal alterations in prostate cancer progression. Top: Profiles are depicted for cell lines containing different numbers of X chromosomes. Genomic DNA isolated from laser captured cells from various prostate tissue sections were profiled for DNA copy number changes. Each row represents a tumor, benign prostate or cell line and each column represents one of 9,550 unique genes, ordered by genome map position from chromosome 1 to Y (red reflects fold-amplification, blue reflects fold-deletion, and white indicates no change). NOR- normal prostate from organ donors and patient; BPH- benign prostatic hyperplasia; S- adjacent stroma; Atrophy- atrophic epithelium; PAH- postatrophic hyperplasia; PIN- prostatic intraepithelial neoplasia; L-PCA- low-grade localized prostate cancer (Gleason Pattern 3); F-PCA- foamy localized prostate cancer; H-PCA- high-grade localized prostate cancer (Gleason Pattern 4); MET- metastatic prostate cancer. Arrows indicate single copy gain in chromosome 21 in a Down's syndrome patient.

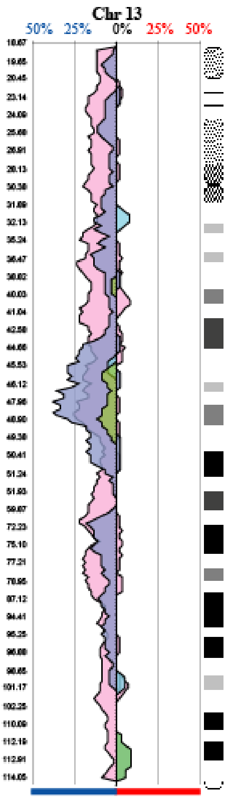

Chr 1
50% 25% 0% 25% 50%
% with DEL % with AMP
Chr 2
50% 25% 0% 25% 50%
Chr 3
50% 25% 0% 25% 50%
Chr 4
50% 25% 0% 25% 50%
Chr 5
50% 25% 0% 25% 50%
Chr 6
50% 25% 0% 25% 50%
Chr 7
50% 25% 0% 25% 50%
Chr 8
50% 25% 0% 25% 50%
MET
H-PCA
L-PCA
PIN

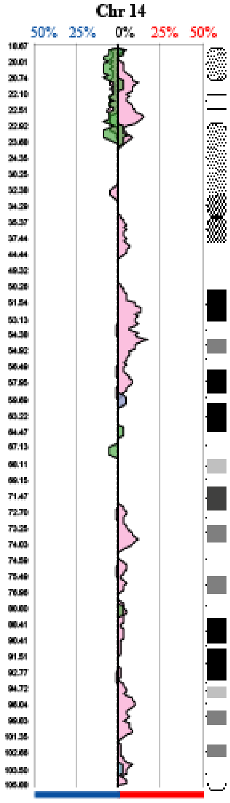

Chr 9
50%
25%
0%
25%
50%

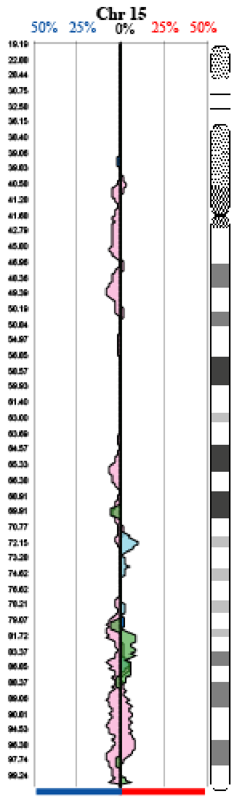

Chr 10
50%
25%
0%
25%
50%

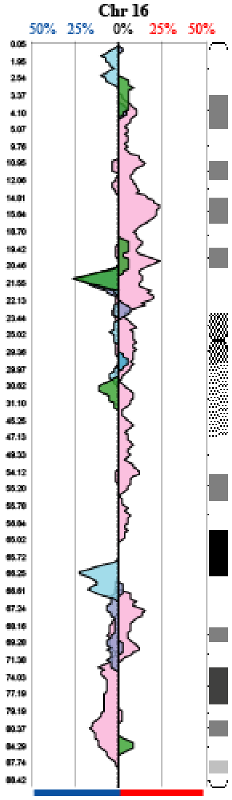

Chr 11
50%
25%
0%
25%
50%


Chr 12
50%
25%
0%
25%
50%


Chr 13
50%
25%
0%
25%
50%


Chr 14
50%
25%
0%
25%
50%


Chr 15
50%
25%
0%
25%
50%


Chr 16
50%
25%
0%
25%
50%

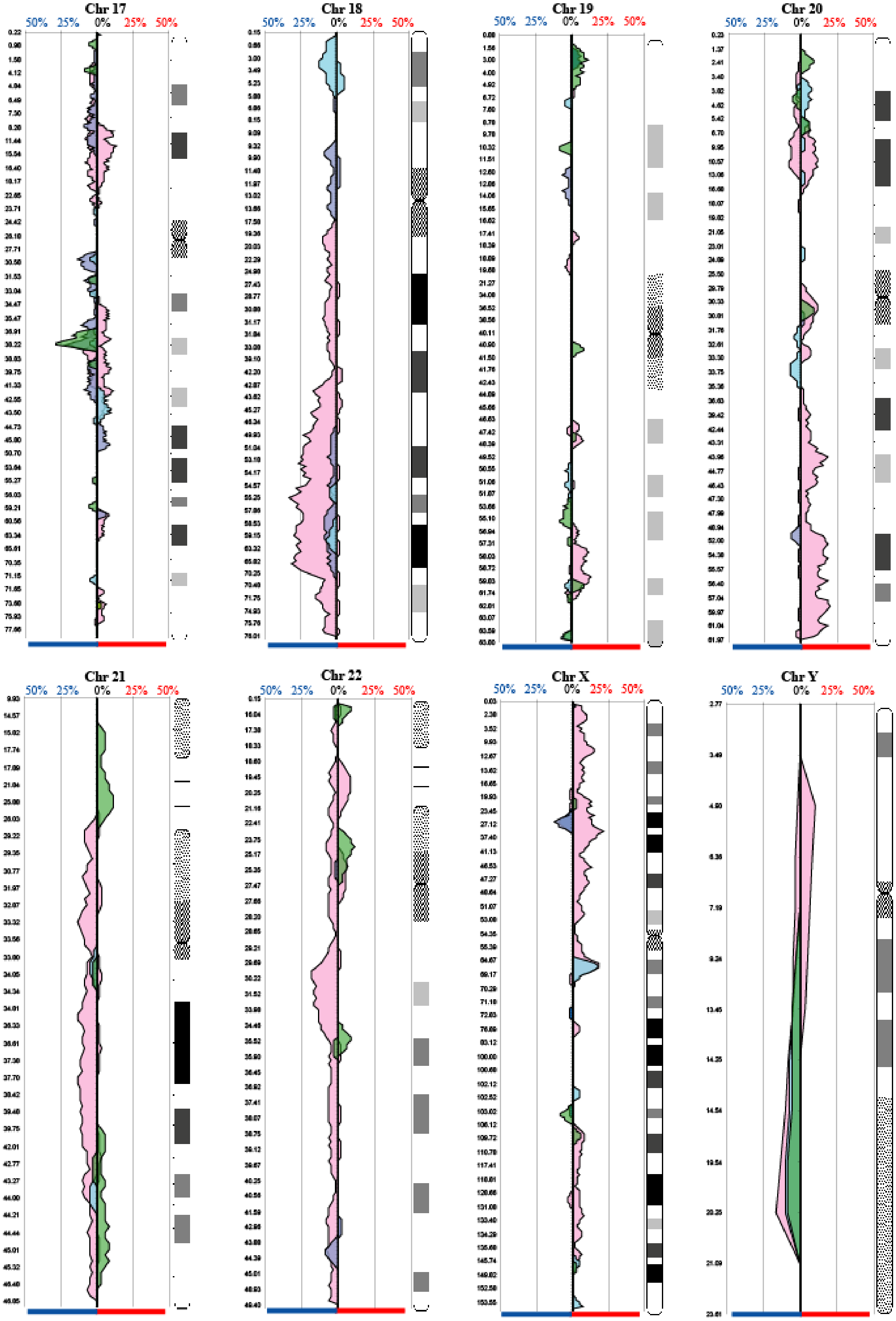


Figure 2.3 Recurrence in PIN, L-PCA, H-PCA, and MET. Recurred regions of amplification and deletion are examined.

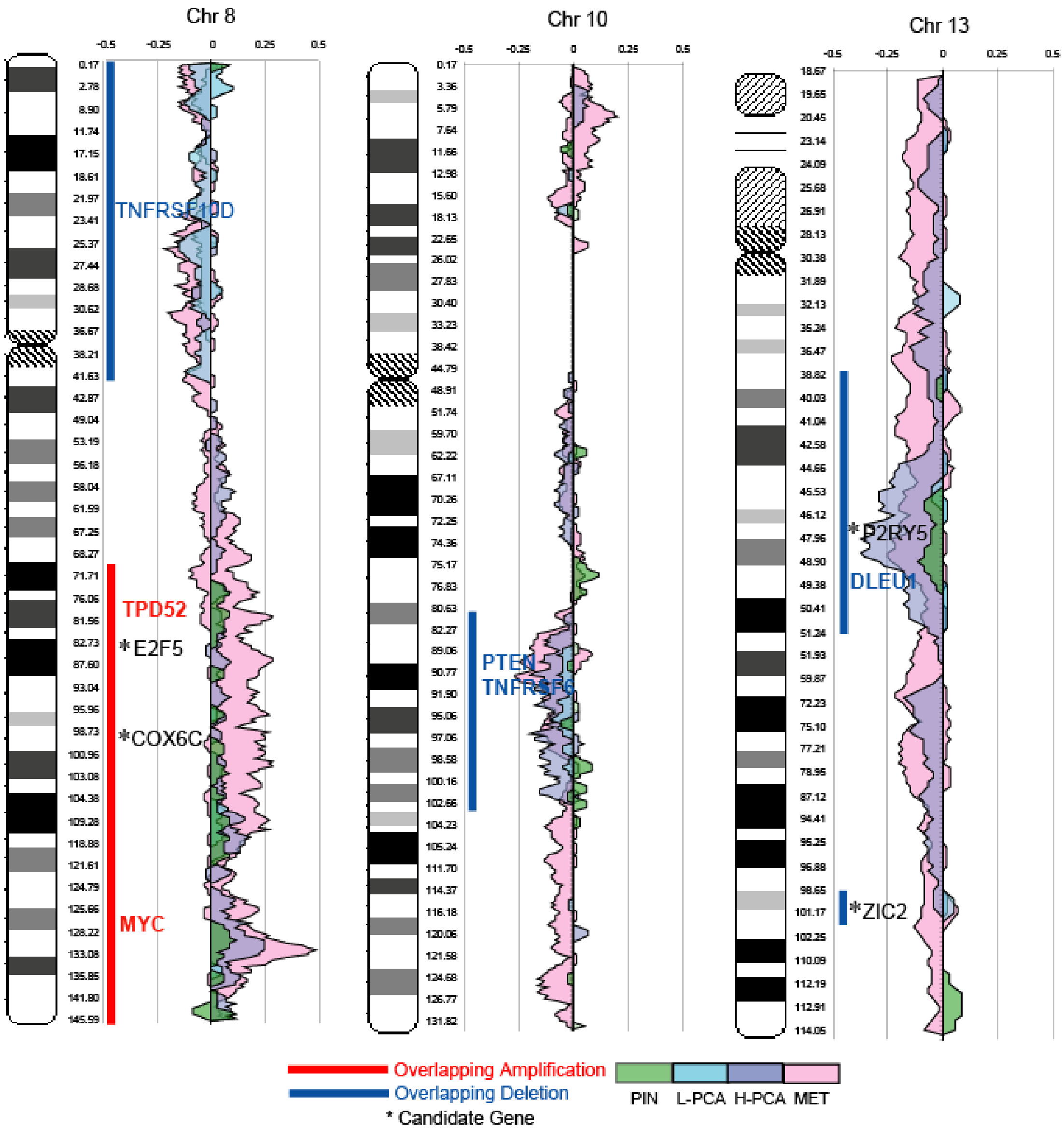


Figure 2.4 Representative chromosomal alterations in prostate cancer. The chromosomal aberrations observed in chromosomes 8, 10, and 13 of PIN, L-PCA, H-PCA, and MET samples are depicted. The peaks moving to the right indicate amplification, and to the left indicate deletion. The amplified genes, *MYC* and *TPD52*, and deleted tumor-suppressor *PTEN* are located within these frequently amplified and deleted regions, respectively. Regions harboring genes including *E2F5*, *COX6C*, *P2RY5*, *MYC*, and *ZIC2* are altered in specimens from all stages of prostate cancer.

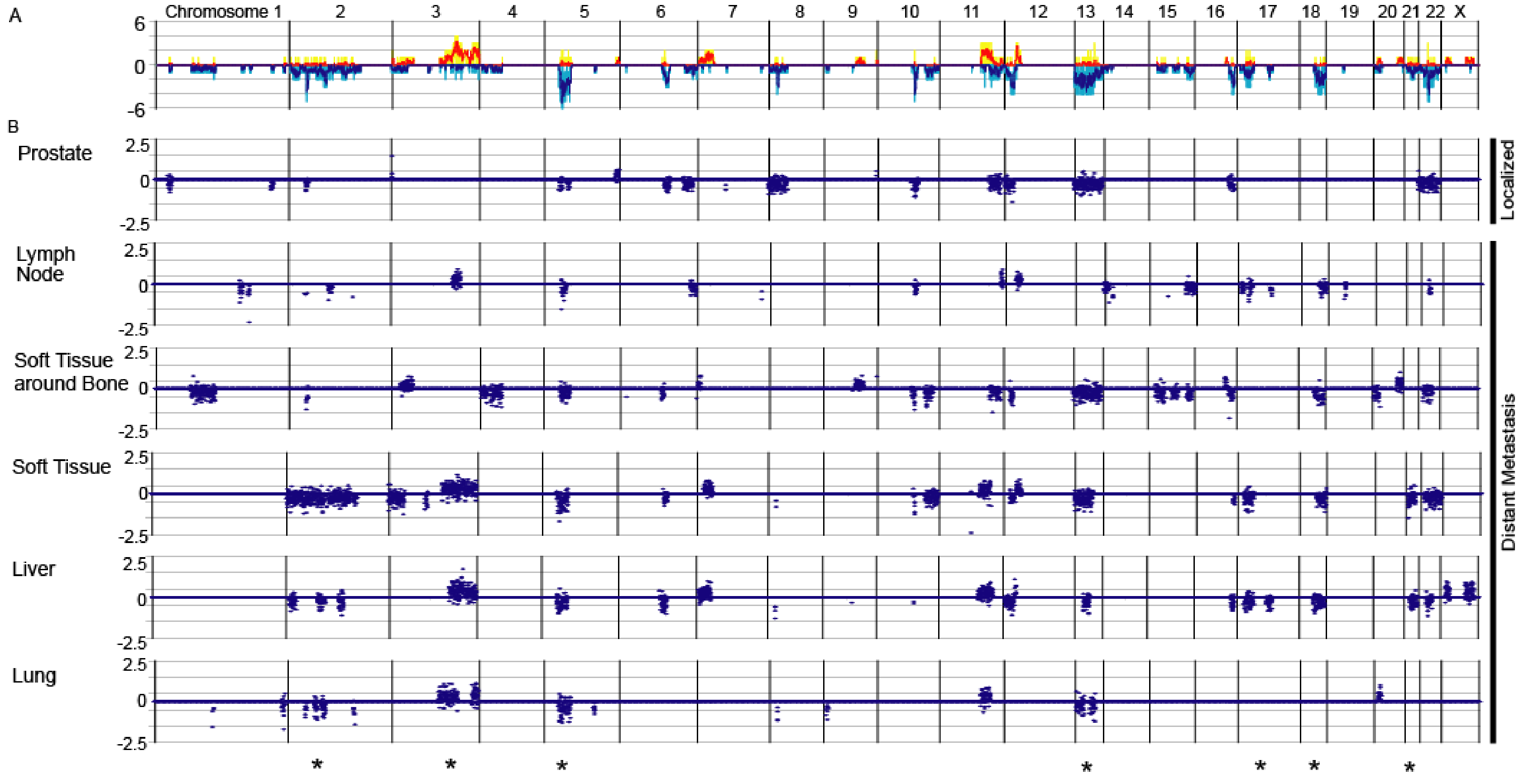


Figure 2.5 Chromosomal aberrations across multiple metastatic sites in a patient with lethal prostate cancer. Distribution of chromosomal alterations in tumor samples collected from 5 metastatic sites in addition to the residual carcinoma of prostate from a single patient. (A) The percentage of alterations observed across all the samples (B) Alterations observed in each individual sample. * indicates shared regions of alterations.

References

1. Albertson DG & Pinkel D (2003) Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet* 12 Spec No 2:R145-152.
2. Cahill DP, Kinzler KW, Vogelstein B, & Lengauer C (1999) Genetic instability and darwinian selection in tumours. *Trends Cell Biol* 9(12):M57-60 .
3. Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, & Meijer GA (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* 34(2):445-450.
4. Pinkel D & Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. *Nat Genet* 37 Suppl:S11-17.
5. Cho EK, *et al.* (2006) Array-based comparative genomic hybridization and copy number variation in cancer research. *Cytogenet Genome Res* 115(3-4):262-272.
6. Tsafrir D, *et al.* (2006) Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res* 66(4):2129-2137.
7. Fritz B, *et al.* (2002) Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. *Cancer Res* 62(11):2993-2998.
8. Pollack JR, *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 99(20):12963-12968.
9. Tonon G, *et al.* (2005) High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A* 102(27):9625-9630.
10. Aguirre AJ, *et al.* (2004) High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A* 101(24):9067-9072.
11. Chaudhary J & Schmidt M (2006) The impact of genomic alterations on the transcriptome: a prostate cancer cell line case study. *Chromosome Res* 14(5):567-586 .
12. Zhao H, *et al.* (2005) Genome-wide characterization of gene expression variations and DNA copy number changes in prostate cancer cell lines. *Prostate* 63(2):187-197.
13. Wolf M, *et al.* (2004) High-resolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. *Neoplasia* 6(3):240-247 .
14. Hermans KG, *et al.* (2006) TMPRSS2:ERG fusion by translocation or interstitial deletion is highly relevant in androgen-dependent prostate cancer, but is bypassed in late-stage androgen receptor-negative prostate cancer. *Cancer Res* 66(22):10658-10663 .
15. Saramaki OR, Porkka KP, Vessella RL, & Visakorpi T (2006) Genetic aberrations in prostate cancer by microarray analysis. *Int J Cancer* 119(6):1322-1329.
16. Paris PL, *et al.* (2004) Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Hum Mol Genet* 13(13):1303-1313.
17. Beheshti B, Vukovic B, Marrano P, Squire JA, & Park PC (2002) Resolution of genotypic heterogeneity in prostate tumors using polymerase chain reaction and comparative genomic hybridization on microdissected carcinoma and prostatic intraepithelial neoplasia foci. *Cancer Genet Cytogenet* 137(1):15-22.

18. Hughes S, *et al.* (2006) The use of whole genome amplification to study chromosomal changes in prostate cancer: insights into genome-wide signature of preneoplasia associated with cancer progression. *BMC Genomics* 7:65.
19. Pollack JR, *et al.* (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23(1):41-46.
20. De Marzo AM, Marchi VL, Epstein JI, & Nelson WG (1999) Proliferative inflammatory atrophy of the prostate: implications for prostatic carcinogenesis. *Am J Pathol* 155(6):1985-1992.
21. Shah R, Mucci NR, Amin A, Macoska JA, & Rubin MA (2001) Postatrophic hyperplasia of the prostate gland: neoplastic precursor or innocent bystander? *Am J Pathol* 158(5):1767-1773.
22. Tomlins SA, *et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39(1):41-51.
23. Wang P, Kim Y, Pollack J, Narasimhan B, & Tibshirani R (2005) A method for calling gains and losses in array CGH data. *Biostatistics* 6(1):45-58.
24. Carrasco DR, *et al.* (2006) High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. *Cancer Cell* 9(4):313-325.
25. Horvath LG, *et al.* (2004) Loss of BMP2, Smad8, and Smad4 expression in prostate cancer progression. *Prostate* 59(3):234-242.
26. van Dekken H, *et al.* (2004) Evaluation of genetic patterns in different tumor areas of intermediate-grade prostatic adenocarcinomas by high-resolution genomic array analysis. *Genes Chromosomes Cancer* 39(3):249-256.
27. Namiki T, *et al.* (2005) Genomic alterations in primary cutaneous melanomas detected by metaphase comparative genomic hybridization with laser capture or manual microdissection: 6p gains may predict poor outcome. *Cancer Genet Cytogenet* 157(1):1-11.
28. Clark J, *et al.* (2003) Genome-wide screening for complete genetic loss in prostate cancer by comparative hybridization onto cDNA microarrays. *Oncogene* 22(8):1247-1252.
29. Watson JE, *et al.* (2004) 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(19):3487-3494.
30. Dhanasekaran SM, *et al.* (2005) Molecular profiling of human prostate tissues: insights into gene expression patterns of prostate development during puberty. *Faseb J* 19(2):243-245.
31. Little SE, *et al.* (2006) Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. *Genomics* 87(2):298-306.
32. Yang YH, *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30(4):e15.
33. Herrero J, *et al.* (2003) GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res* 31(13):3461-3467.

CHAPTER 3

GENOMIC ABERRATION PLAYS A ROLE IN TRANSCRIPTIONAL CHANGE IN PROSTATE CANCER PROGRESSION

The role of copy number changes on transcriptional regulation has been reported in several studies including breast, pancreatic, lung among other cancers. Using integrative analysis, several groups have identified the regions of differentially expressed genes that accompany copy number change. For example, in lung cancer, the genome-wide aCGH and expression array profiling of adenocarcinoma (AC) and squamous-cell carcinoma (SCC) tumor types identified 3q26-29 as the key region of genomic difference, where both copy number gain and corresponding increase in gene expression is observed in SCC samples, but not in AC samples (1). The concordance between the amplification and corresponding gene-upregulation is reported to be around 40-60%, recognizing genomic aberration as a strong driving mechanism for gene expression changes. In breast cancer, an aCGH study coupled with gene expression on cDNA microarray platform estimated a 62% (representing 54 unique genes) association between 117 highly amplified genes and transcript over-expression (2). Other similar study also on breast cancer reported a 44% concordance between amplification and over-expression, and identified gene amplification as the underlying mechanism in 10.5% of highly over-expressed genes (3). In pancreatic cancer, as many as 60% of the genes located within highly amplified regions were reported over-expressed (4). Similar high level concordance between DNA copy number change and mRNA expression level was also observed in our prostate study, supporting the role of gene copy number in transcriptional up-regulation. In this dissertation study, we integrated aCGH and corresponding gene expression data (5) obtained from matched samples to evaluate genomic aberrations accompanying gene expression changes during prostate cancer progression. Recent finding in prostate cancer include the recurrent gene fusion events involving ETS transcriptional family. This event plays a crucial role in prostate cancer development, and these gene fusions appear to be one of the earliest events involving prostate cancer (6). We hypothesized the samples with ETS gene fusion might use an alternate path for tumorigenesis compared with non-ETS samples. The copy number differences between ETS and non-ETS prostate samples are addressed through Molecular Concept Map (MCM) analysis (5) and identified various chromosomal regions including 6q21 that distinguish ETS overexpressing samples from others.

Results and Discussion:

Analysis of array CGH data and mRNA expression profiles on matched LCM prostate specimens

The aim of our integrative analysis was to identify candidate regions with genetic alterations that accompany corresponding transcriptomic changes. Transcript expression patterns of genes located in the chromosomal regions with significant aberrations in identical samples were compared (**Fig. 3.1**). An association between mRNA over-expression and chromosomal gain was observed, such that among highly amplified genes

($\log_2$ ratio≥0.5), there was a 26% and 20% concordance in high level mRNA expression ($\log_2$ ratio≥1) for MET and PCA samples respectively, while at moderate level mRNA expression ($\log_2$ ratio≥0.4), a 42% and 22% concordance was observed, for MET and PCA tissues. Among all the amplified genes ($\log_2$ ratio≥0.22), a moderate level over-expression ($\log_2$ ratio≥0.4) was observed in 38% of MET and 20% of PCA cases, and a high level over-expression ($\log_2$ ratio≥1) was observed in 23% MET and 15% of PCA cases. A previous breast cancer aCGH and coupled gene expression study estimated a 62% (representing 54 unique genes) association between 117 highly amplified genes and transcript over-expression (2). Hyman *et al*. have reported 44% highly amplified genes associated with over-expression, and 10.5% of highly over-expressed genes to be amplified in breast cancer (3). In pancreatic cancer, as many as 60% of the genes located within highly amplified regions were reported over-expressed (4). Similar high level concordance between DNA copy number change and mRNA expression level was also observed in our prostate study, in which the increased dosage in the gene copy number most likely plays a major role in their transcriptional up-regulation. The MCM analysis (5) from the Oncomine database revealed the enrichment of over-expressed genes in chromosomal arms 8q, 1q, 7p, 9q, 16p, 10p, and 3q (p-value<0.05), where 8q, 7p, 9q, 16p, and 3q are among the top chromosomal alteration sites in our MCR analysis from Chapter 2 (**Table 2.2**). Integrative analysis of our aCGH and gene expression data allows a direct comparison between the change in copy number and transcript expression levels, and genes within regions of significant genomic alterations show concordance at the mRNA expression level.

Integrative analysis of genomic and transcriptomic profiles associated with prostate cancer progression

To identify the top altered genes that are associated with a change in expression level, we selected the candidate genes based on three criteria mentioned under integrative analysis of copy-number-based differential gene expression section in Materials and Methods. These significantly altered genes are located within the commonly observed regions of chromosomal aberrations and are accompanied with the altered mRNA expression in a correlated manner. The chosen genes from PCA and MET samples are likely to play a role in mRNA expression level, and they are cross indexed with three independent gene expression datasets available in public domain (**Fig. 3.2**). The gene expression data was obtained from grossly dissected localized and metastatic prostate tumor tissues from previously reported studies from our group (7-8) and others (9). The amplified gene section was enriched with transcript overexpression and the deleted section with mRNA down regulation. These differentially expressed genes are located in either PCA and MET or both. Well-known amplified genes such as *MYC* and *TPD52(10-13)* (10-13) are among the top genes that show over-expression pattern in PCA and MET samples in various datasets. The tumor suppressor *PTEN* and suppressor of cytokine signaling, *SOCS6* that are known to be deleted in various cancers are also seen as under-expressed. Some of the other previously described gains are *PTK2, KIAA0196, PVT1, NSE2*, and *RAB25*, and some of the earlier reported losses include *SPTA1, NEFL, FVT1,* *TNFRSF6*, *EDNRB*, *C13ORF1*, *LCP1*, *BMPR1A*, and *CDH19*. Some of the novel amplified and deleted genes identified in this study include *DDEF1*, *LCHN, F5, DDX56, P2RY5, ATAD1, ZNF532, RAB27B, and PPIL6* which merit further characterization. In addition, a progressive gene signature was identified by the transcriptome analysis performed on identical samples studied here. These genes showed a robust progression signature whose expression increased or decreased during the progression from benign epithelium to PIN to PCA to MET (5). We used the aCGH data for the corresponding samples to look for possible underlying genetic alterations involving the proposed candidate genes. The data is presented in **Table 3.1**.

Genetic alterations in ETS versus non-ETS samples

ETS transcription factors that include ERG, ETV1, and ETV4 were identified as outliers in prostate cancer gene expression data set and are demonstrated to be involved in recurrent gene fusion (14). Two recent studies, one using SNP arrays on human prostate cancer tissues and the other using BAC arrays on prostate cancer xenografts (15) propose interstitial deletions as a mechanism of *TMPRSS2:ERG* gene fusion on chromosome 21 (14). These gene fusions appear to be one of the earliest events involving prostate cancer and lead to the over expression of the fused ETS gene in an androgen-regulated manner (6). We previously characterized ETS expression in the cancer samples used in this study (5). Accordingly the localized and metastatic prostate cancer samples were either grouped into ETS (ERG or ETV1 overexpressing) or non-ETS samples (**Fig. 3.3A**). A significance test was performed to identify regions that distinguish between these two sample groups from the CGH miner output (q-value<0.01). A total of 50 genes passed the cutoff of p-value<0.05, and this list was analyzed using MCM. MCM identified chromosome sub-region concepts like 1q23 (p-value<4.1e-4), 6q16 (p-value<1.4e-9), 6q21 (p-value<1.5e-5), 10q23 (p-value<7.5e-7), and 10q24 (p-value<2.1e-4). Importantly, various oncomine gene expression signature concepts that define ETS positive versus non-ETS samples from the matched mRNA dataset (5), and other independent dataset like Lapointe *et al.* (9) and Glinsky *et al.* (16) were enriched in this analysis (**Fig. 3.3B**). The aberration summary and accompanying gene expression pattern in these chromosomal subregions are presented as a heatmap (**Fig. 3.3C**). Gene expression analysis by Tomlins *et al.* showed differential enrichment in chromosome subarm 6q21 between ETS and non-ETS samples. We speculated either amplification of 6q21 in ETS or loss in non-ETS tumors. Our aCGH data showed several non-ETS samples with loss of 6q21 region (>45%), suggesting that under-expression of genes from this region could be due to deletions in a subset of non-ETS samples. Several groups have previously identified loss of 6q21 in localized prostate cancers (17), and here we demonstrate this phenomenon to be mainly associated with non-ETS samples. FOXO3A (18) and CCNC (17) that have been proposed to participate in prostate carcinogenesis are located in this region. These alterations on 6q21 and others identified in this analysis may collectively play a role in tumor development in the non-ETS group, and further molecular characterization of these alterations is required to understand its importance in prostate cancer. This observation was also validated in an independent grossly dissected prostate cancer aCGH dataset (data not shown).

Integrative analysis of copy-number-based differential gene expression

The lists of genes that are candidates for the copy-number-based differential expression were selected on the basis of three criteria: 1. The percent alteration among samples (described above), 2. The correlation between mRNA expression and aCGH, and 3. The significance of copy number change. Genes were ranked by their correlation with mRNA expression data and the degree of copy number change in either direction. The genes that ranked among top 500 in either category were selected individually from MET and PCA groups. Among these, the genes that show aberration in more than 5 samples in MET and PCA groups (n=364), were mapped and compared (n=210) to the other prostate cancer gene expression datasets from Dhanasekaran *et al.* (7), LaPointe *et al.* (9), and Varambally *et al.* (19) studies, obtained from the Oncomine database (www.oncomine.org). For candidate gene progression analysis, a given gene must have chromosomal alteration in all stages of prostate cancer progression starting from the precursor lesions. A total of 504 (over-expressed and amplified) and 241 (under-expressed and deleted) filtered genes were mapped to the mRNA expression progression list (p-value<0.05) from our matched study available in the Oncomine database (5). Genes that are ranked among the top/bottom 100 genes in mRNA expression progression list are reported in **Table 3.1**.

In conclusion, the direct relationship between copy-number change and mRNA expression levels were investigated utilizing a parallel transcriptomic study, where more than 40% of the highly altered genes were associated with elevated mRNA expression level. This study also has identified some novel regions of aberrations and candidate genes in prostate cancer. Lastly, MCM analysis of the cancer specimens identified chromosomal regions including 6q21 and gene expression concepts that distinguish ETS overexpressing samples from non-ETS samples.

Materials and Methods:

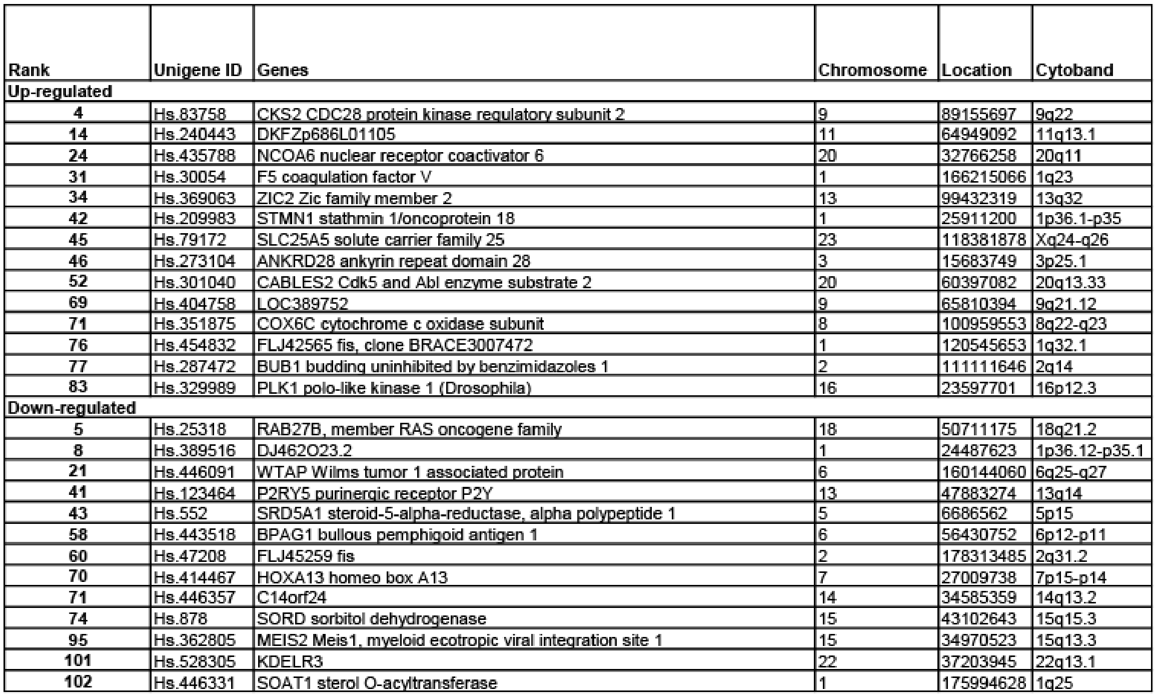
Integrative analysis of copy-number-based differential gene expression

The lists of genes that are candidates for the copy-number-based differential expression were selected on the basis of three criteria: 1. The percent alteration among samples (described above), 2. The correlation between mRNA expression and aCGH, and 3. The significance of copy number change. Genes were ranked by their correlation with mRNA expression data and the degree of copy number change in either direction. The genes that ranked among top 500 in either category were selected individually from MET and PCA groups. Among these, the genes that show aberration in more than 5 samples in MET and PCA groups (n=364), were mapped and compared (n=210) to the other prostate cancer gene expression datasets from Dhanasekaran *et al.* (7), LaPointe *et al.* (9), and Varambally *et al.* (19) studies, obtained from the Oncomine database (www.oncomine.org). For candidate gene progression analysis, a given gene must have chromosomal alteration in all stages of prostate cancer progression starting from the precursor lesions. A total of 504 (over-expressed and amplified) and 241 (under-expressed and deleted) filtered genes were mapped to the mRNA expression progression list (p-value<0.05) from our matched study available in the Oncomine database (5). Genes that are ranked among the top/bottom 100 genes in mRNA expression progression list are reported in **Table 3.1**.

Statistical Analysis

Molecular Concepts Map analysis. Molecular Concepts Map (MCM) analysis (20) is a bioinformatic tool offered through the Oncomine database that enables integration of molecular concepts, pathways, and networks previously defined in the literature or other datasets. In brief, MCM analysis uses Fisher's exact test to find various significantly enriched concepts in an uploaded gene list and provides visual interaction networks. Querying a user-defined uploaded gene set against the MCM database allows for an integrative analysis of the strength of overlap between the user-defined gene set and all MCM gene sets (which represent the previously defined molecular concepts, pathways, and networks). The results are then visualized as a series of nodes and lines, where lines represent significant overlap between gene sets (shown as nodes).

Table 3.1 Progression gene list. Genes that are ranked in the top 100 of a parallel transcriptome study and whose chromosomal alterations are observed during the early onset of cancer progression are shown along with chromosomal location information as well as the rank from the progression analysis.



Rank	Unigene ID	Genes	Chromosome	Location	Cytoband
Up-regulated					
4	Hs.83758	CKS2 CDC28 protein kinase regulatory subunit 2	9	89155697	9q22
14	Hs.240443	DKFZp686L01105	11	64949092	11q13.1
24	Hs.435788	NCOA6 nuclear receptor coactivator 6	20	32766258	20q11
31	Hs.30054	F5 coagulation factor V	1	166215066	1q23
34	Hs.369063	ZIC2 Zic family member 2	13	99432319	13q32
42	Hs.209983	STMN1 stathmin 1/oncoprotein 18	1	25911200	1p36.1-p35
45	Hs.79172	SLC25A5 solute carrier family 25	23	118381878	Xq24-q26
46	Hs.273104	ANKRD28 ankyrin repeat domain 28	3	15683749	3p25.1
52	Hs.301040	CABLES2 Cdk5 and Abl enzyme substrate 2	20	60397082	20q13.33
69	Hs.404758	LOC389752	9	65810394	9q21.12
71	Hs.351875	COX6C cytochrome c oxidase subunit	8	100959553	8q22-q23
76	Hs.454832	FLJ42565 fis, clone BRACE3007472	1	120545653	1q32.1
77	Hs.287472	BUB1 budding uninhibited by benzimidazoles 1	2	111111646	2q14
83	Hs.329989	PLK1 polo-like kinase 1 (Drosophila)	16	23597701	16p12.3
Down-regulated					
5	Hs.25318	RAB27B, member RAS oncogene family	18	50711175	18q21.2
8	Hs.389516	DJ462O23.2	1	24487623	1p36.12-p35.1
21	Hs.446091	WTAP Wilms tumor 1 associated protein	6	160144060	6q25-q27
41	Hs.123464	P2RY5 purinergic receptor P2Y	13	47883274	13q14
43	Hs.552	SRD5A1 steroid-5-alpha-reductase, alpha polypeptide 1	5	6686562	5p15
58	Hs.443518	BPAG1 bullous pemphigoid antigen 1	6	56430752	6p12-p11
60	Hs.47208	FLJ45259 fis	2	178313485	2q31.2
70	Hs.414467	HOXA13 homeo box A13	7	27009738	7p15-p14
71	Hs.446357	C14orf24	14	34585359	14q13.2
74	Hs.878	SORD sorbitol dehydrogenase	15	43102643	15q15.3
95	Hs.362805	MEIS2 Meis1, myeloid ecotropic viral integration site 1	15	34970523	15q13.3
101	Hs.528305	KDELR3	22	37203945	22q13.1
102	Hs.446331	SOAT1 sterol O-acyltransferase	1	175994628	1q25

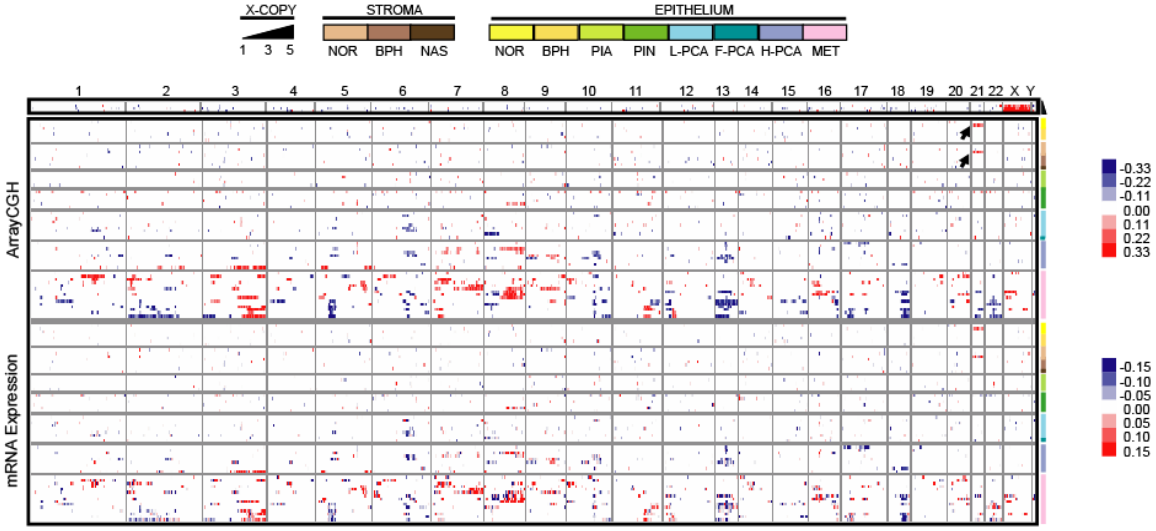


Figure 3.1 Genome-wide view of integrative analysis of chromosomal alterations and change in gene expression in prostate cancer. Top: Chromosomal aberrations are depicted for cell lines containing different numbers of X chromosomes, followed by benign and tumor prostate samples. Each column represents one of 9,550 unique genes, ordered by genome map position from chromosome 1 to Y (red reflects fold-amplification, blue reflects fold-deletion, and white indicates no change). Lower panel displays the mRNA expression of matched samples within regions of significant genomic alteration. NOR- normal prostate from organ donors and patient; BPH- benign prostatic hyperplasia; S- adjacent stroma; Atrophy- atrophic epithelium; PAH- postatrophic hyperplasia; PIN- prostatic intraepithelial neoplasia; L-PCA- low-grade localized prostate cancer (Gleason Pattern 3); F-PCA- foamy localized prostate cancer; H-PCA- high-grade localized prostate cancer (Gleason Pattern 4); MET- metastatic prostate cancer. Arrows indicate single copy gain in chromosome 21 in a Down's syndrome patient. † indicates ERG-overexpressing samples. ‡ indicates ETV1-overexpressing samples.

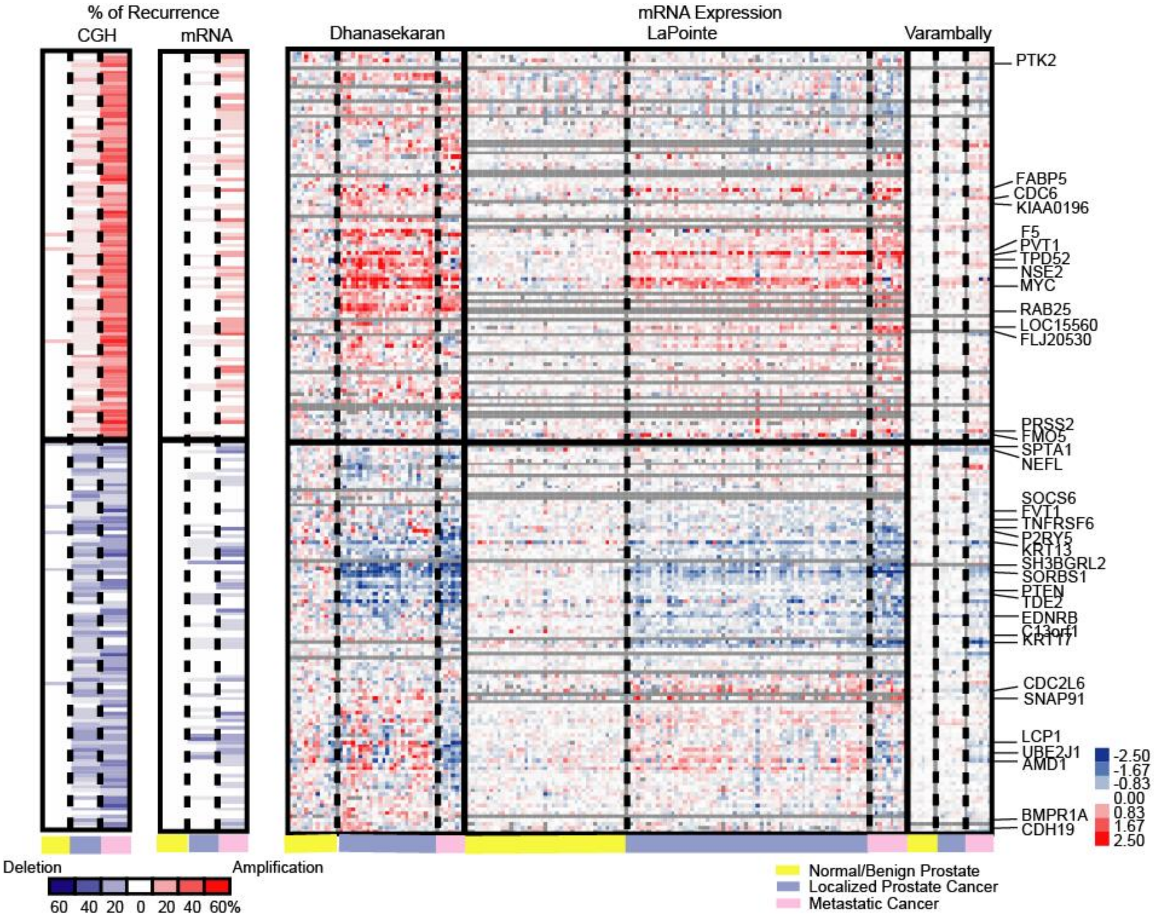


Figure 3.2 Concordantly altered candidate genes in various prostate cancer studies. The proposed candidate amplified/deleted genes that are correlated with matched mRNA expression data with high percentage of alterations are mapped to Dhanasekaran *et al.*, Lapointe *et al.*, and Varambally *et al.* datasets available from www.oncomine.org and are displayed. Left panel indicates the percentage of alterations in CGH (blue indicates the percentage of deleted samples, and red indicates the percentage of amplified samples) and the percentage of samples with over-expressed/under-expressed genes in matched mRNA expression data.

Kim et al. Fig5

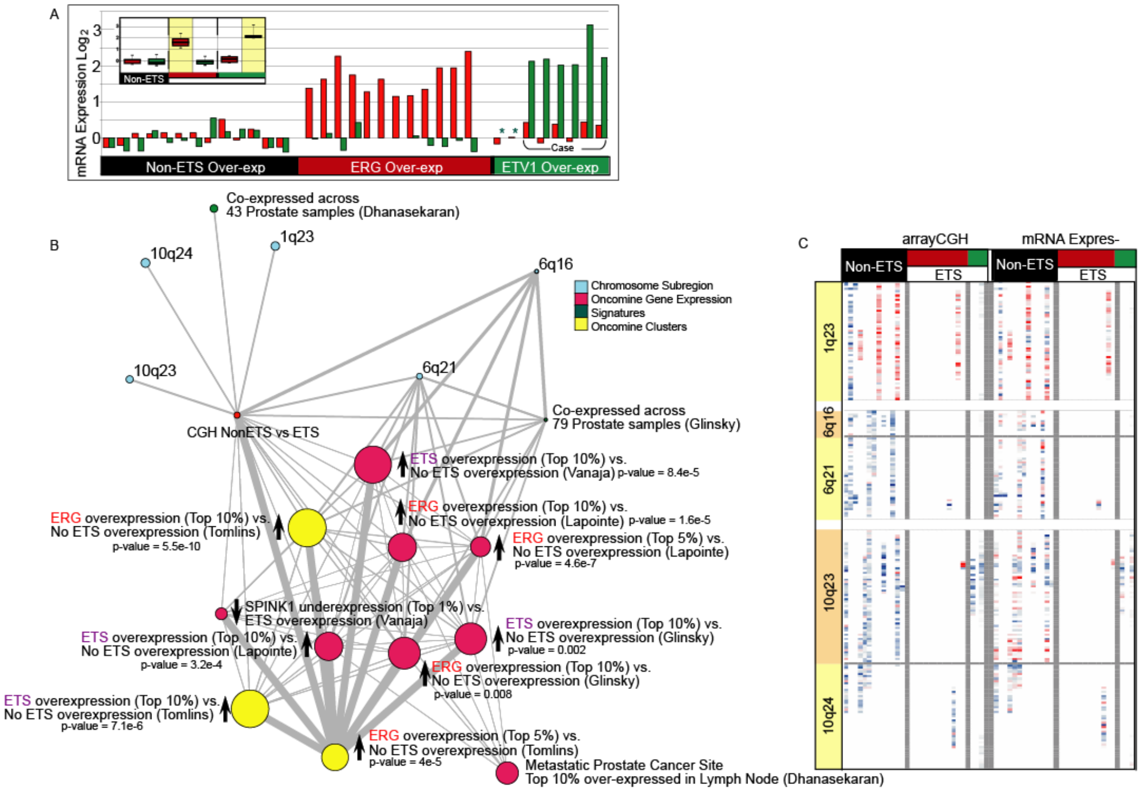


Figure 3.3 Genetic alterations in non-ETS vs. ETS samples in prostate cancer. The genomic aberration differences as well as the enriched concepts of the genes located within differentially altered regions in non-ETS and ETS samples were analyzed using the Molecular Concept Map (MCM) (5) (**A**) mRNA expression of non-ETS and ETS over-expressing prostate cancer samples (red bar for ERG and green bar for ETV1 expression values). Upper left inset represents an average of the data presented. (**B**) Network map showing enrichment in chromosomal sub-regions and gene expression signatures that define non-ETS and ETS samples. “CGH non-ETS vs. ETS” represents the data gathered from this study. (**C**) Heatmap of differentially aberrant genomic regions between non-ETS and ETS samples *Data not available (ETV1 expression is confirmed from an independent sample obtained from the same case)

Notes

This chapter has been previously published: Kim JH, *et al.* (2007) Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res* 67(17):8229-8239. All rights reserved.

References

1. Tonon G, *et al.* (2005) High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci U S A* 102(27):9625-9630.
2. Pollack JR, *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 99(20):12963-12968.
3. Hyman E, *et al.* (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res* 62(21):6240-6245.
4. Heidenblad M, *et al.* (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene* 24(10):1794-1801.
5. Tomlins SA, *et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39(1):41-51.
6. Rubin MA & Chinnaiyan AM (2006) Bioinformatics approach leads to the discovery of the TMPRSS2:ETS gene fusion in prostate cancer. *Lab Invest* 86(11):1099-1102.
7. Dhanasekaran SM, *et al.* (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature* 412(6849):822-826.
8. Varambally S, *et al.* (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* 419(6907):624-629.
9. Lapointe J, *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* 101(3):811-816.
10. Boutros R, Fanayan S, Shehata M, & Byrne JA (2004) The tumor protein D52 family: many pieces, many puzzles. *Biochem Biophys Res Commun* 325(4):1115-1121.
11. Byrne JA, *et al.* (2005) Tumor protein D52 (TPD52) is overexpressed and a gene amplification target in ovarian cancer. *Int J Cancer* 117(6):1049-1054.
12. Rubin MA, *et al.* (2004) Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res* 64(11):3814-3822.
13. Wang R, *et al.* (2004) PrLZ, a novel prostate-specific and androgen-responsive gene of the TPD52 family, amplified in chromosome 8q21.1 and overexpressed in human prostate cancer. *Cancer Res* 64(5):1589-1594.
14. Tomlins SA, *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 310(5748):644-648.
15. Hermans KG, *et al.* (2006) TMPRSS2:ERG fusion by translocation or interstitial deletion is highly relevant in androgen-dependent prostate cancer, but is bypassed in late-stage androgen receptor-negative prostate cancer. *Cancer Res* 66(22):10658-10663 .
16. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, & Gerald WL (2004) Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* 113(6):913-923.
17. Konishi N, Shimada K, Ishida E, & Nakamura M (2005) Molecular pathology of prostate cancer. *Pathol Int* 55(9):531-539.

18. Trotman LC, *et al.* (2006) Identification of a tumour suppressor network opposing nuclear Akt function. *Nature* 441(7092):523-527.

19. Varambally S, *et al.* (2005) Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* 8(5):393-406.

20. Rhodes DR, *et al.* (2007) Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia* 9(5):443-454 .

CHAPTER 4

CHARACTERIZATION OF PROSTATE CANCER CELL LINE DNA METHYLOME BY METHYLPLEX-NEXT GENERATION SEQUENCING

CpG residues, the targets of DNA methylation, have an asymmetric distribution in mammalian genomes and are often found in small clusters termed CpG islands(1). In total, CpG islands populate the promoter regions of 60-70% of all human genes(2), and in cancer, hypermethylation of gene promoters commonly marks disease progression, silencing putative tumor suppressor genes. In prostate cancer, aberrant DNA methylation is established in at least two waves(3), with promoter methylation of GSTP1, APC, COX2 occurring earlier in localized disease and promoter methylation for ER and p14/INK4a detected only during metastatic stages(4). Previously, DNA methylation studies in prostate cancer have employed methodologies of variable scale, focusing on either a few promoters(5) or several thousand genomic regions with a CpG island array(6). Alternatively, functional approaches that monitored gene expression changes after treatment with the demethylating agent 5-Aza-2'-deoxycytidine (5-Aza) have also been utilized(4, 7). However, to date only 85 genes are listed in the Pubmeth database (www.pubmeth.org) as known targets of methylation in prostate cancer(8). The advent of next generation sequencing (NGS) now presents a novel approach to assess genome-wide epigenetic changes without the limitations of probe-based microarray platforms.

Recently, several groups employed NGS with bisulfite converted genomic DNA for plant(9) and mammalian genomes(10). Yet, to achieve better representation of the methylome, enrichment of methylated regions based on methylation-specific antibodies(11-12), binding proteins(13), restriction enzymes(14) or capture technology(12, 15) has also been employed. In this chapter, we have described the first example of NGS with a restriction enzyme-based enrichment method coupled with amplification to characterize the DNA methylome of the prostate cancer cell line LNCaP and primary benign prostate epithelial cells PrEC.

Results and Discussion:

Characterization of DNA Methylation in Prostate Cancer

To prepare LNCaP and PrEC methylome samples, we employed the M-NGS workflow to enrich regions with DNA methylation and generate next generation sequencing libraries. Fifty nanograms of genomic DNA from LNCaP and PrEC cells were used as input DNA to create Methylplex libraries. Methylplex libraries were constructed by digesting input DNA with methylation-sensitive restriction enzymes, followed by adaptor ligation and subsequent PCR amplification with universal primers. A second round of enzymatic treatment depleted non-GC rich sequences, followed by an additional amplification to ensure enrichment of highly methylated DNA fragments. The amplification adaptors were enzymatically removed prior to NGS library preparation. The schematic of sample preparation is presented in **Figure 4.1**. The Methyplex libraries described above were constructed through the commercial service option provided by Rubicon Genomics Inc, Ann Arbor, MI as a part of early access to the technology. While many NGS methodologies to date require several micrograms of input genomic DNA(11-12, 14), the very low input requirement in this protocol facilitates the use of finite samples of limited availability. This technology if further validated, opens up the possibility of carrying out differential global DNA methylation analyses with similar ease as is currently carried out with gene expression technologies.

For this study, we used two different concentrations (1 and 5 ug) of each Methylplex library as starting material to obtain single-read sequencing on the Illumina Genome Analyzer II (see Methods for protocol details). For each cell type (LNCaP and PrEC) a total of 4 sequencing libraries (200bp-1, 200bp-5, 400bp-1 and 400bp-5) were prepared corresponding to 200- and 400-bp size selections of 1 ug and 5 ug of Methylplex product. We obtained an average of 5 million mappable reads per M-NGS sample (see **Table 4.1**). CG dinucleotides were enriched by the Methylplex procedure up to three-fold in mapped reads from M-NGS as compared to previously obtained pan-histone Chip-Seq data (**Table 4.1**). A Hidden Markov Model (HMM)-based algorithm described in the Methods section (http://www.sph.umich.edu/csg/qin/HPeak) was used to detect enriched regions from mapped reads obtained in each sequencing run (**Table 4.1**). While direct counting of the sequencing reads would have achieved the same purpose, this application highlighted significantly enriched genomic regions by inferring true peaks over background signal. The H-peak results from two independent LNCaP experiments done in two separate batch of samples showed over 80% overlap. Moreover, the HPA II methylation-sensitive restriction enzyme sites analyzed in our sample is around 600,000, which is 5-fold higher than 113,000 sites analyzed by Brunner et al. (14).

To demonstrate experimental consistency, a comparative analysis of data from 1 and 5ug Methylplex DNA exhibited high correlation both for reads mapping to chromosome 21 and for reads mapping to all CpG islands (see **Fig. 4.2** and **Materials and Methods**). Hybridizing the Methylplex product to an Agilent CpG island array showed maximum overlap (70%) with the 400bp-5ug runs and therefore data from these runs were selected for further analysis (**Fig. 4.3**). Globally, we found a 70% overlap in methylated genomic regions between LNCaP (56,727 regions) and PrEC cells (61,615 regions) (**Fig. 4.4A**). While the overall number of methylated regions between two cell lines were comparable (**Fig. 4.4B**), 8 out of 10 genes known to have CpG island hypermethylation in prostate cancer tissues(5) were methylated in LNCaP cells and not in PrEC, as expected.

Global Differences in CpG Methylation

Because aberrant hypermethylation in CpG rich promoters is a common feature of tumorigenesis(16) and thought to contribute to the repression of tumor-suppressor genes(17), we assessed the extent of CpG island methylation in our samples. Of the 71,120 (56Mb) CpG islands identified by using Takai Jones criteria(18) in the human genome, 6,865 (7.6Mb) and 5,767 (6.1Mb) CpG islands were methylated in LNCaP and PrEC respectively. An overall 1.7-fold difference in uniquely methylated CpG islands observed between LNCaP and PrEC increased to ~7-fold specifically in CpG islands associated within gene promoters and not located elsewhere (**Fig. 4.4C**).

In total, 3,496 Ref-Seq gene promoters (±1,500bps flanking transcription start site) contained methylation in at least one sample (**Fig. 4.4D**). Visualization of these in the context of promoter CpG islands revealed the presence of distinct methylation patterns on gene promoters (**Fig. 4.4D**). Broadly, the promoters fell into 2 groups based on the presence or absence of a CpG island within this specified region. Interestingly, 35% of promoters (n=1,232) lacked CpG islands but nevertheless exhibited methylation around the transcription start site (TSS) (**Fig. 4.4D**, X to XII). The remaining 65% (n=2,264) (I to IX) had CpG islands spanning the TSS and 3 distinct methylation patterns were observed in this group: (1) Methylation was mostly confined (39.6%, n=1,383) to the island (I to III), and interestingly with much higher frequency (greater than 6 fold difference) in LNCaP (n=952) compared to PrEC (n=147) cells (**Fig. 4.4D**). (2) Methylation was positioned 5' to the CpG island (11.8%, n=412, IV to VI); and (3) methylation was positioned 3' to the CpG island (13.4%, n=469, VII to IX). In total, methylation flanking the 5' or 3' of promoter CpG islands accounted for 25.2% of all promoter methylation observed (n=881), which corroborates the methylation of "shores" as observed by Irizarry et al.(19) up to 2kb away from CpG islands. Among the differentially methylated regions (DMRs), LNCaP showed comparable DNA methylation in shores and CpG islands, whereas PrEC cells exhibited 2-fold more methylation in shores, using the Irizarry et al. criteria (data not shown). In total, this promoter analysis identified 813 unique gene promoters methylated only in LNCaP, which were then considered for further analysis.

Independent Validation of DMRs

We next characterized DMRs identified by M-NGS using 3 independent approaches including "Methyl-Profiler" qPCR (SABiosciences, Frederick, MD), bisulfite sequencing (see **Table 4.2** for primers), and a functional strategy using gene expression arrays. First, bisulfite sequencing and methyl-profiler qPCR, which also provides a quantitative estimate of percent methylation, validated the methylation status of 15/15 genes (100%) tested for LNCaP-specific and PrEC-specific methylation, as well as an unmethylated control region in the MYC promoter (**Figs. 4.5A** and **B**). Finally, following treatment of LNCaP cells with the demethylating agent 5-Aza, we observed upregulation of 246 out of 973 methylated genes identified by M-NGS by using Agilent 244K gene expression arrays (Significance Analysis of Microarray, 5% false discovery rate) (**Fig. 4.5C**). This approach provided a functional validation for DNA methylation observed in these target regions**.**

Molecular Concepts Map Analysis

To identify molecular concepts, pathways or networks enriched in the 813 promoter regions with LNCaP-specific methylation, we analyzed our dataset using the Molecular Concept Map (MCM) derived from the Oncomine database(20-21). Out of 813 genes, 789 mapped to the Oncomine database, and MCM analysis of these genes revealed preferential enrichment with methylated and under-expressed genes signatures from localized and metastatic PCa samples (lowest p-value<1.90E-14) from several independent studies, as well as datasets representing genes with higher expression in benign prostate tissue relative to cancer. (**Fig. 4.6A,** and **Materials and Methods**). Of note, we observed concepts such as, "genes previously known to be methylated in prostate cancer" (p-value<1.40E-06), "gene ontology-tumor suppressor genes" (p-value<0.009), and "hypermethylated cancer genes" (p-value<4.10E-07) (**Fig. 4.6A**).

Importantly, we observed the enrichment of DMRs identified by an independent Differential Methylation Hybridization (DMH) profiling of 29 tissue samples from various stages of prostate cancer progression (**Table 4.3**). Forty percent of mappable probes between the 2 platforms (134/309) had methylation in at least one PCa tissue sample (**Fig. 4.7**), which corroborates with the presence of a DMH-Tissue ("Methylated in PCa") MCM concept demonstrating methylation-mediated repression in PCa. By contrast, PrEC cells did not share this enrichment, and MCM analysis of PrEC-only methylated regions showed minimal overlap with LNCaP MCM analysis (**Fig. 4.6B**). Indeed, only MCM concepts relating to histone modification are common to both PrEC and LNCaP MCM analysis. Taken together, these findings indicate that the targets of DNA methylation in LNCaP have significant overlap with genes that are methylated and/or repressed in prostate cancer tissues.

Promoter Methylation and Transcriptional Repression

This association of promoter methylation and gene repression was further confirmed by applying Gene Set Enrichment Analysis (GSEA) to transcriptome NGS data from LNCaP and PrEC cells to separately examine the expression levels of all respective genes identified as methylated by M-NGS in these individual samples (**Materials and Methods)**. As expected, LNCaP and PrEC cells showed significant enrichment of gene repression only among genes with methylated promoters in LNCaP (p <0.0013) and PrEC (p<0.03), and not gene body methylation (**Figs. 4.8A and B**). Methylation-mediation repression of 3 novel targets in LNCaP (KCTD1, TACSTD2, and

CALML3; see **Figs. 4.5A, 4.5B and 4.9A**), PrEC (SPON2, GAGE genes) and in both cell lines (HIC1; see **Fig. 4.9A**) highlight this phenomenon.

Next, we explored the association between methylation and repression among the various methylation categories presented in **Figure 4.4D**. In LNCaP cells, we observed a strong association between gene repression and promoter methylation regardless of whether the promoters contain CpG islands ($p < 0.0039$ and $p < 0.0015$, respectively) (**Fig. 4.8B**), which is consistent with a recent report showing repression of Oncostatin M (OSM) by methylation despite the absence of CpG island in its promoter(22). However, we failed to find a significant association among genes that displayed methylation in regions flanking promoter CpG islands. These results suggest that methylation-mediated gene repression does not require a CpG island-containing promoter, but rather all gene promoters may be functionally regulated by DNA methylation.

Finally, to identify putative methylation targets with low expression in PCa tissues we performed a meta-analysis using the Oncomine database. Among 783 genes with greater than 5 fold repression in LNCaP RNA-seq data compared to PrEC, 81 genes were both represented in Oncomine and contained LNCaP-specific promoter methylation. Scoring these genes across 13 PCa datasets for gene repression validated GSTP1 as the top-ranked gene (**Fig. 4.9B**), and overall 44 out of 81 genes are in the top 25% repressed genes in at least one study. Among the top 5 candidates from meta-analysis, the first four(5, 23-24) have been previously demonstrated to be methylated in prostate cancer. We validated the 5th ranked gene WFDC2 as well as TACSTD2 (48th), by qPCR on a small PCa tissue cohort and cell line panel (**Fig. 4.10A and 4.10B**). Both WFDC2 and TACSTD2 exhibited cancer-specific methylation (both methylated in 4/4 PCa samples but 0/6 benign tissues); but interestingly only WFDC2 was observed methylated in metastatic samples. As a positive control, we validated GSTP1 methylation for this cohort as well (**Fig. 4.10C**).

Transcript Isoform Regulation by DNA Methylation

During our M-NGS data analysis we observed that a subset of genes displayed selective promoter methylation in a transcript isoform specific manner, suggesting a novel mechanism for regulating isoform expression in cancer. A well-known example, RASSF1, which is frequently inactivated by epigenetic alteration in human cancers(25), is comprised of three distinct isoforms. In LNCaP, we observed silencing of the longer transcript of RASSF1, variant-1, by DNA methylation, while the smaller isoforms, variants-2 and -3, which code for N-terminal variant proteins expressed in multiple cancer cell lines and tissues including PCa(26-27)(**Fig. 4.11A**), retains high expression (**Fig. 4.11B**). Active transcription of variants-2 and -3 in LNCaP cells is supported by histone 3 lysine 4 trimethylation (H3K4me3) as observed in ChIP-Seq data, and 5’ Rapid Amplification of cDNA Ends (5’RACE) showed presence of shorter transcripts but not variant-1 in LNCaP (**Fig. 4.11A**). Isoform-specific methylation of variant-1 was confirmed by preferential re-expression of this transcript upon 5-Aza treatment of LNCaP cells (**Fig. 4.11B**). Interestingly, when we superimposed the promoter methylation (**Fig. 4.4D**) and H3K4me3 ChIP-seq data (**Fig. 4.12**) from LNCaP cells, we found that these epigenetic marks segregate into distinct genomic regions, and H3K4me3 binding was remarkably sparse among the promoters that lacked CpG islands compared to those with CpG islands. A recent publication from Adrian Bird’s lab demonstrated the role of protein Cfp1, a CpG binding protein in H3K4me3 modification of non-methylated CpG islands. They observed a 93% overlap in Chip-Seq data between regions occupied by Cfp1 protein and those modified by H3K4me3 mark in mouse brain tissue. Hence a low recruitment of Cfp1 could be the reason for sparse H3K4me3 modification in promoters without CpG islands (28).

While the published paper states that Cfp1-histone H3K4 methyltransferase Setd1 complex is recruited to unmethylated CpG islands, our observation supports an alternate notion. We see both DNA methylation and H3K4me3 modification at distinct and mutually exclusive region on the same CpG islands (**Fig. 4.12**). The boundaries between the two modifications is readily visible in promoters that exhibit CGI-shore DNA methylation. This alludes to the scenario where DNA methylation could act like a foot print to histone H3K4 modifcations on certain promoters and this could be due to stearic hindrance of the protein complexes that make these epigenetic marks. While integration of other epigenetic marks is necessary for a full analysis, these data further suggest that multiple epigenetic modifications may co-occur in distinct patterns to regulate transcript expression in cancer.

Since our M-NGS methodology accurately detected DNA methylation events of RASSF1, we queried our data for differential methylation of transcript variants compared to H3K4me3 marks, and identified 34 genes in LNCaP that exhibit isoform-specific promoter methylation (**Table 4.4**). We validated 2 genes from this list namely NDRG2 and APC (**Fig. 4.11C and 4.11E**). In both of these candidates, the transcript variants (variants 1-4 in NDRG2 and variant-2 and -3 in APC) showing DNA methylation were confirmed to be under-expressed in LNCaP as compared to PrEC by qRT-PCR and

5'RACE (**Fig. 4.11C-F**). Furthermore, these variants were preferentially re-expressed upon 5-Aza treatment of LNCaP cells. To determine whether patient tissues demonstrated similar isoform-specific expression patterns, we tested for NDRG2 isoforms in tissue samples by qRT-PCR. Similar to LNCaP cells, variants 1-4 were significantly under-expressed as compared to variants 5-8 in localized PCa (p-value=0.034) and adjacent benign prostate (p-value=0.012), but not in normal (non-prostate cancer) samples (**Fig. 4.13**). These results indicate the presence of a global cancer-specific DNA methylation pattern that regulates the use of alternative TSS. This mechanism may play a role in the differential expression of transcript variants between normal and cancerous cells, leading to functionally important transcriptomic differences in tumorigenesis.

In summary, we used a high throughput M-NGS strategy to characterize the DNA methylome map of LNCaP and PrEC cells using a minimal amount of input DNA. We observed distinct patterns of DNA methylation around TSSs that frequently occur on promoters either containing or lacking a CpG island. We also found evidence that selective regional DNA methylation regulates expression of specific transcript isoforms between normal and cancer cells.

Materials and Methods:

Reagents, cell lines and prostate tissue samples

Human primary prostate epithelial cells (PrEC) were purchased from Lonza Inc. (Mapleton IL.), and the prostate cancer cell line LNCaP was obtained from ATCC

(Manassas, VA). The LNCaP cells were grown RPMI 1640 containing 10% FBS (Invitrogen, Carlsbad, CA). The PrEC cells were cultured in PrEGM media (Lonza Inc. Mapleton, IL) and cells from passage 7 and 8 were used in this study. Grossly dissected human prostate tissue samples from University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core (SPORE) were collected with informed consent of the patients and prior institutional review board approval. A total of 6 normal, 12 localized cancer and 11 metastatic prostate samples (n=29) were characterized by DMH (**Table 4.3**). Twelve thousand element CpG island microarray was purchased from University Health Network Microarray Center (Toronto, Ontario, Canada) and used for DMH analysis. Agilent Human CpG Island (244K) microarray (Agilent Technologies, Santa Clara, CA) was used to hybridize methylplex products for validation. Genomic DNA was isolated from cultured cells and tissue using DNeasy Blood and tissue kit (Qiagen Inc, Valencia, CA) according to manufacturer's instructions. 5-Aza-2´-deoxycytidine (5-Aza) was purchased from Sigma-Aldrich Co (St. Louis, MO) and used at 6 uM final concentration dissolved in DMSO.

M-NGS library generation

Early access to the Methylplex library synthesis and GC-enrichment was obtained through a commercial service provided by Rubicon Genomics Inc. Ann Arbor, MI. A kit version of this protocol is also under development by Rubicon. Briefly, fifty nanograms of gDNA were digested with a proprietary cocktail of methylation-sensitive restriction enzymes (Rubicon Genomics) and then amplified by PCR with universal primers to create a MethylPlex library that is enriched for methylated DNA. MethylPlex DNA was then subjected to additional enzymatic treatment to deplete all non-GC-rich DNA sequences, purified and amplified in a second round of PCR. This created a highly enriched library of fully methylated GC-rich regions of the human genome, representing about 1% of total DNA. After purification the amplification adaptors were removed by a restriction enzyme digestion. One and five micrograms of the purified products from each cell line were directly incorporated into the genomic DNA sequencing sample preparation kit procedure of Illumina (Illumina Inc, San Diego, CA) at the end repair step, skipping the nebulization process. An adenine base was then added to the purified end repaired products using Klenow exo (3' to 5' exo minus) enzyme. The reaction product was purified, ligated to Illumina adaptors with DNA ligase and resolved on a 2% agarose gel. Gel pieces were excised at 200 and 400 base pair positions and the DNA was extracted using Qiagen gel extraction kit (Qiagen Inc, Valencia, CA). Four sequencing libraries (200bp-1, 200bp-5, 400bp-1 and 400bp-5) corresponding to 200- and 400-bp size selections of 1 ug and 5 ug of Methylplex product were prepared for each cell line. One microliter of this eluate was used as a template in a PCR amplification reaction with Phusion DNA polymerase (Finnzymes, INC., Woburn, MA) to enrich the adapter modified DNA fragments. The PCR product was purified and analyzed by Bioanalyzer (Agilent Technologies, San Diego, CA) before using it for flow cell generation, where 10nM of library was used to prepare flowcells with approximately 30,000 clusters per lane. The raw sequencing image data were analyzed by the Illumina analysis pipeline, aligned to the unmasked human reference genome (NCBI v36, hg18) using the ELAND software (Illumina) to generate sequence reads of 25-32 bps. This data will be deposited to NCBI Sequence Read Archive (SRA) upon acceptance.

Choice of 400bp-5 Methylplex Libraries for Global Analysis

We prepared 4 next generation sequencing (NGS) libraries for each Methylplex DNA sample (LNCaP and PrEC), which corresponded to two different size selections for NGS libraries prepared with 1ug and 5ug of Methylplex product (200bp-1, 200bp-5, 400bp-1, 400bp-5). All NGS libraries were sequenced in one lane on a Illumina Genome Analyzer II flowcell, using single read technology. Each library generated an average of 5 million mappable reads. A regression analysis of the mappable reads revealed high correlation between the corresponding 400bp-1 and 400bp-5 samples for PrEC and LNCaP. First, the 400bp-1 and 400bp-5 samples were compared for all reads that map to chromosome 21 (R^2 = 0.9508 for LNCaP and R^2 = 0.8556 for PrEC, **Fig. 4.2A**). Next, these samples were compared for all reads that mapped to a CpG island, with slightly higher concordance observed (R^2 = 0.9644 for LNCaP and R^2 = 0.9819 for PrEC, **Fig. 4.2B**). When we looked at all regions obtained from the LNCaP 400bp-1 and 400bp-5 sequencing runs (covering 22.15 Mb and 24.74 Mb, respectively), a significant overlap with the coverage of 20.38 Mb was observed. Similar results were seen in PrEC, where we observed a coverage overlap of 21.19 Mb (overlapping between 22.71 Mb for 400bp-1 and 26.54 Mb for 400bp-5). We chose the LNCaP and PrEC 400bp-5 samples for analysis in this study because data from these sequencing runs showed slightly higher enrichment of CG-rich sequences (**Table 4.1**) and showed maximum overlap (70%) with methylation identified independently by hybridizing the Methylplex product to an Agilent CpG island array.

Methylplex Library Hybridization

Two micrograms of the purified products from each PrEC and LNCaP MethylPlex DNA were labeled and hybridized to an Agilent Human CpG 244K array (G4492A, Santa Clara, CA), where LNCaP sample was coupled with Cy5 and PrEC to Cy3. A dye-flip experiment was also performed. The samples were labeled and hybridized according to the manufacturer's protocol (Agilent). The scanned images were analyzed and extracted using Agilent Feature Extraction Software 9.1.3.1.

RNA Seq library preparation

Poly-A RNA from LNCaP and PrEC cells (200 ng) was isolated from total RNA using SeraMag Magnetic Oligo(dT) Beads (Thermo Fisher Scientific, Waltham, MA). RNA was fragmented at 70 °C for 5 min in a fragmentation buffer (Ambion, Austin, TX), and converted to first-strand cDNA using SuperscriptII (Invitrogen, Carlsbad, CA). Second-strand cDNA synthesis was performed with *Escherichia coli* DNA pol I (Invitrogen, Carlsbad, CA). The double-stranded cDNA library was further processed following Illumina Genomic DNA sample preparation protocol which involved end repair using T4 DNA polymerase, Klenow DNA polymerase and T4 Polynucleotide kinase followed by a single 'A' base addition using Klenow 3' to 5' exo^{-} polymerase. Illumina's adaptor oligo was ligated using T4 DNA ligase. The adaptor-ligated library was size selected by separating on a 4% agarose gel and cutting out the library smear at 200 base pairs. The library was PCR amplified by Phusion polymerase (Finnzymes, INC., Woburn, MA), and purified by PCR purification kit (Qiagen, Valencia, CA). The library was quantified with Quant-iT picogreen dsDNA assay kit (Invitrogen, Carlsbad, CA) on a

Modulus single tube luminometer (Turner Biosystems Inc, Sunnyvale, CA) following the manufacturer's instructions. The library (10 nM) was used to prepare flowcells with approximately 30,000 clusters per lane.

Statistical Analysis

HMM analysis of M-NGS data. Hidden Markov Model (HMM) based next generation sequencing analysis is conducted in a two-step process that takes in raw reads and outputs refined boundaries of enriched chromosomal regions (29). The first step includes the formation of hypothetical DNA fragments (HDFs) from uniquely mapped reads, where the coverage of HDFs is determined by the specified DNA fragment size and overlapped HDFs are merged to represent one consecutive genomic region. The second step is designed to refine the boundaries of enriched region using HMM with bin size of 25bp (by default). Under null hypothesis, raw reads are assumed to land on the genome following a Poisson distribution with the background rate of r^0, and enriched regions are expected to have more HDFs with statistical significance. The rate of the Poisson distributions in a given sample is assumed to be r^1, and the transition probabilities are estimated empirically, based on inferred enriched regions defined in the first step. The output from HMM is selected based on the posterior probability of being in the enriched regions, and then further filtered using maximum read counts. The threshold for maximum read counts is determined from Bonferroni corrected p-value of 0.001 calculated using a Poisson distribution with background rate r^0. The output is provided in BED format as well as Wiggle format for UCSC genome browser visualization. The ouput file annotation field contains information such as enriched genomic position and length, max height, GC content, repeated sequencing genomic position and length, mean and standard deviation of conservative scores for enriched region, relationship with nearest genes including whether the enriched region is located within the gene or between genes, gene name, GB accession number, strand, distance to gene transcription start site.

Molecular Concepts Map analysis. Molecular Concepts Map (MCM) analysis (30) is a bioinformatic tool offered through the Oncomine database that enables integration of molecular concepts, pathways, and networks previously defined in the literature or other datasets. In brief, MCM analysis uses Fisher's exact test to find various significantly enriched concepts in an uploaded gene list and provides visual interaction networks. Querying a user-defined uploaded gene set against the MCM database allows for an integrative analysis of the strength of overlap between the user-defined gene set and all MCM gene sets (which represent the previously defined molecular concepts, pathways, and networks). The results are then visualized as a series of nodes and lines, where lines represent significant overlap between gene sets (shown as nodes). In addition to over 15,000 biological concepts from Oncomine, which include manual curation of the literature, target gene sets from genome-scale regulatory motif analyses, and reference gene sets from several gene and protein annotation databases, we have uploaded a gene list from differentially methylated regions identified from an independent Differential Methylation Hybridization profiling (concept named "DMH-Tissue Methylated in PCa"), as well as known methylated genes in cancers provided from Pubmeth database.

Gene Set Enrichment Analysis (GSEA) Gene Set Enrichment Analysis (GSEA)(31-33) is a computational method that assesses whether a defined set of genes shows statistically significant, concordant differences between any two given conditions. The fold change between the raw counts from RNA-seq NGS data on LNCaP and PrEC (representing 24,167 unique genes) was calculated and genes were ranked by the order of expression in LNCaP. This list was uploaded as a pre-ranked gene list to GSEA v2.04 (Broad Institute, Cambridge, MA), and using respective gene lists of methylated targets in LNCaP and PrEC cell lines, GSEA was performed using a weighted enrichment statistic and default normalization mode.

Oncomine Meta-analysis. A complete description of meta-analysis performed in Oncomine is available(20). In brief, a genelist of interest is uploaded to the Oncomine database, and the built-in meta-analysis tool rank-orders the genelist by the p-value, which is determined by Student´s t-test for comparisons made within each available dataset (for example Cancer vs. Normal). The ranked genes were visualized with pink and green shades (top ranked ones with darker shades, pink for over-expression and green for repression) in heatmap format, with each row representing genes and each column representing the dataset. Final order of the genes is determined by averaging ranks across the datasets.

Calculating gene expression from RNA-Seq data. Gene expression levels of passing filter reads from RNA-Seq data that mapped by ELAND to exons (March 2006 assembly of UCSC KnownGene table) in LNCaP and PrEC cell lines are quantified as described(34).

Significance Analysis of Microarray (SAM). Significance analysis of microarray (SAM) (35) (http://www-stat.stanford.edu/~tibs/SAM/) was performed on the gene expression dataset obtained from 5-Aza and DMSO-treated LNCaP cells by selecting genes that were methylated in LNCaP. From 1,171 methylated genes from LNCaP M-NGS

(Supplementary Table 4), a total of 973 genes was mapped to Agilent expression profiling data. One-class SAM analysis was done using default settings, and significant genes were calculated with a false discovery rate (FDR) of 0.05.

Differential Methylation Hybridization (DMH)

Differential methylation hybridization was performed according to a previously published protocol(36). DNA samples were isolated from tissues using DNeasy Blood and Tissue kit (Qiagen Inc, Valencia, CA). DNA was digested with MseI and ligated to linkers, and while one half of the sample was digested with McrBC, the other half was mock digested. The purified DNA was used as template in PCR amplification. The amplified product was labeled with fluorescent dyes using indirect labeling method, where mock digested samples were coupled with Cy5 and McrBC digests were coupled to Cy3. The labeled pair from each sample was combined, denatured at 95ºC for 2 minutes and hybridized to the 12,000 element CpG array from University Health Network Microarray Center (Toronto, Ontario, Canada) overnight at 60ºC. Washed slides were scanned using 4000A scanner (Axon Instruments) and acquired images were analyzed with GenePix 6.0 software. All clone annotations can be obtained from the supplier's website (http://data.microarrays.ca/cpg/). The microarray data will be deposited in GEO for public access. Universally methylated (Zymo Research Corporation, Orange, CA) and unmethylated DNA (Millipore, Billerica, MA) was purchased for experimental controls.

Methyl-Profiler

Methyl-Profiler^{TM} (SABiosciences, Frederick, MD) is a restriction enzyme digestion based novel technology for CGI methylation profiling, requiring less than 500 ng input genomic DNA. The samples were first digested with methylation-sensitive (Ms) and/or methylation-dependent (Md) restriction enzymes along with mock digestion according to manufacturer's instruction. PCR reactions were performed with ABI StepOne qPCR machine (Applied Biosystems, Foster City, CA) with RT^2 SYBR Green/ROX qPCR Master Mix (SABiosciences, Frederick, MD) and primers targeting the region of interest. The PCR reactions were carried out with following conditions: 10 min at 95 C, followed by 40 cycles of 97 C for 15'', 72 C 1 min as described in manufacturer's protocol. Using delta-Ct values, the relative amounts of methylation are calculated using an automated Excel-based data analysis template provided by the manufacturer. The mock digested template is used for initial DNA input quantification, the Ms enzyme is used for hypermethylation quantification, and the Md enzyme is used for quantifying unmethylated DNA. A mixture of these 2 enzymes (Msd) is used to quantify the undigested amount of DNA. A methylation rate below 5 % is considered not significant. While the calculated methylation percentage between 10 and 60 is considered intermediate, the values above sixty are taken as heavy methylation.

Bisulfite Sequencing

Bisulfite conversion was carried out using EZ DNA methylation gold kit (Zymo Research Corporation, Orange, CA) according to manufacturer's instructions. Briefly 500ng of genomic DNA from either LNCaP or PrEC cells in 20ul volume was mixed with 130uls of CT conversion reagent and was initially incubated at 98ºC for 10 minutes followed by incubation at 64ºC for 2.5 hours. M-biding buffer (600ul) was added to the above reaction and DNA purified using a Zymo spin column. Sequential washes were performed with 100ul M-Wash buffer, 200ul M-sulphonation buffer and 200ul of M-wash buffer was carried out before eluting the DNA in 30ul of M-elution buffer. Purified DNA (2ul) was used as template for PCR reactions with primers (Integrated DNA Technologies Inc. San Diego, CA) and synthesized according to bisulfite converted DNA sequences for the regions of interest using the Methprimer software (37). The PCR product was gel purified and cloned into pCR4 TOPO TA sequencing vector (Invitrogen, Carlsbad, CA). Plasmid DNA isolated from 10 colonies from each sample was sequenced by conventional Sanger Sequencing (University of Michigan DNA Sequencing Core). The "BIQ Analyzer"(38) online tool was used to calculate the methylation percentage and to generate the bar graphs.

Gene expression profiling

For 5-Aza stimulation experiments, LNCaP cells cultured in RPMI 1640 were treated with vehicle, dimethyl sulfoxide (DMSO) or 6 uM 5-Aza for 4 or 6 days in duplicates. Total RNA was isolated with Trizol (Invitrogen) and further purified using RNAeasy Micro Kit (Qiagen) according to the manufacturer's instructions. Expression profiling was performed using the Agilent 44K expression array. One microgram of total RNA was converted to cRNA and then labeled according to the manufacturer's protocol (Agilent). Hybridizations were performed for 16 h at 65 °C. Scanned images from Agilent microarray scanner were analyzed and extracted using Agilent Feature Extraction Software 9.1.3.1, with linear and lowess normalization performed for each array. A total of 4 hybridizations were performed including two 4 day and two 6 day 5-Aza treated samples (Cy5) against control DMSO-treated samples (Cy3). Gene expression data will be submitted to GEO database.

5' RACE

5' RACE was performed as previously described(39). First-strand cDNA was amplified with gene-specific reverse primers RASSF1, APC, and NDRG2 (**Table 4.2**) and 5' GeneRacer primers (Invitrogen) using Platinum Taq High Fidelity enzyme (Invitrogen) after the touchdown PCR protocol according to manufacturer's instructions. PCR amplification products were cloned into pCR4-TOPO TA vector (Invitrogen) and sequenced bidirectionally using vector primers as described(40).

Total RNA isolation and Quantitative real time PCR (QPCR)

The total RNA was isolated from cells using RNeasy mini kit (Qiagen, Valencia, CA) according to manufacturer's instructions. A DNAseI treatment step was included during the total RNA isolation procedure to remove genomic DNA from the samples. One microgram of total RNA was used in cDNA synthesis using Superscript III reverse transcriptase (Invitrogen, Carlsbad,CA). Quantitative real time PCR (QPCR) was performed on prostate cell line cDNA samples using SYBR Green Mastermix (Applied Biosystems) on an Applied Biosystems 7900 Real Time PCR system as described(40). All oligonucleotide primers were synthesized by Integrated DNA Technologies and are listed in **Table 4.2**. GAPDH primers were as described(41). The amount of target transcript and GAPDH in each sample was normalized by standard ddCt methodology, and then to the reference PrEC or DMSO-treated LNCaP samples accordingly.

ChIP-Sequencing

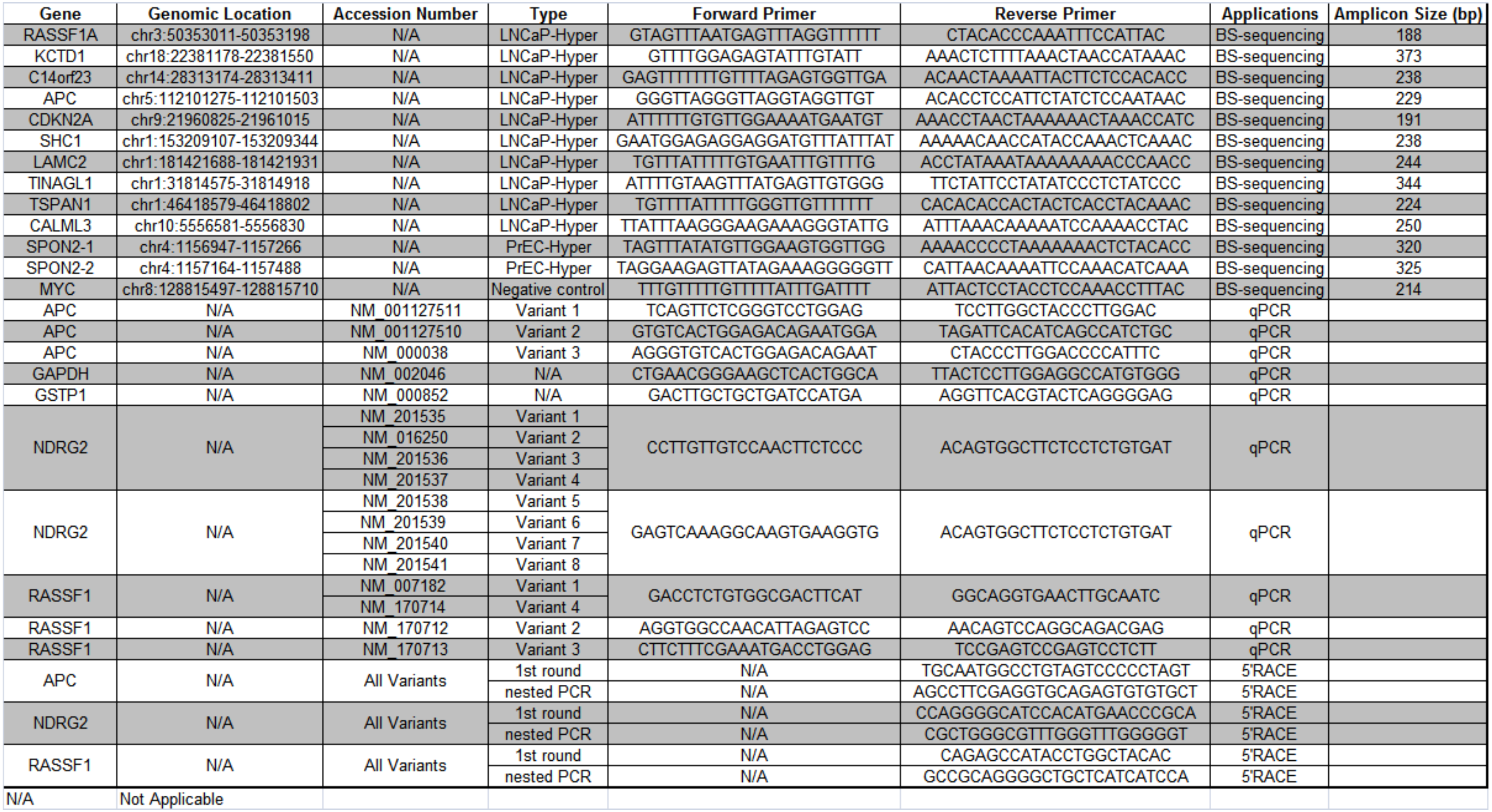
ChIP-Seq data obtained from LNCaP cells for H3K4me3 antibody (Abcam) and PanH3 (Abcam) is from the manuscript by Jindan Yu et al., currently under press (Cancer Cell 2010). ChIP samples were prepared for sequencing using the Genomic DNA sample prep kit (Illumina) following manufacturers protocols. To facilitate ChIP-Seq data analysis, a Hidden Markov Model (HMM)-based enriched region identifying algorithm (described in the Methods section under statistical analysis) was utilized.

Table 4.1 Summary of Methylplex sequencing results

FC ID		20FAMAAXX.s1	20FAMAAXX.s2	20FAMAAXX.s3	20FAMAAXX.s4	20FR7AAXX.s2	20FR7AAXX.s3	20FR7AAXX.s4	20FR7AAXX.s6	20E46AAXX.s6	20FFJAAXX.s4
Cell Line		PrEC		LNCaP		PrEC		LNCaP		VCaP	LNCaP
Gel Cut Position (bp)		200				400				200	
PCR starting amt (ul)		1	5	1	5	1	5	1	5	1	1
Method		Methyplex-seq (M-NGS)								CHIP-seq (PAN-H3)	
Total_Count		14811616	18256298	12544655	17288784	6804113	11052480	6474869	9662549	7436098	5316154
PF_Count		4593385	4781606	5946287	5503787	4570589	6102042	4595660	6085438	5297890	3675564
Adaptor_Count		37	23	38	24	511	1033	482	883	263	182
Repeat_Count		15910	17569	19935	37531	27783	25570	32352	56173	88596	118004
NM		347333	331616	385252	383185	406890	681012	399206	634534	119818	137737
QC		243	312	675	687	6657	884	561	808	5944	1384
Hs_Ref	total	4050482	4234706	5286063	4826830	3905822	5082676	4016058	5196774	4620638	3012625
	chr1	310363	324482	451974	420725	313855	410053	321201	416946	358460	272625
	chr2	311745	330800	415824	387064	253813	336497	276277	361285	369876	213058
	chr3	199561	216273	269374	266256	146130	194806	155453	201859	342989	215170
	chr4	198400	212602	225550	221752	158591	210419	141775	183841	311332	181116
	chr5	172485	186575	226120	222479	152374	204335	155676	209491	273496	181669
	chr6	176111	189339	178785	178834	116389	157148	102769	134491	239286	157810
	chr7	242843	251563	320079	288871	231368	302500	222371	289405	284486	170535
	chr8	189223	200082	227081	215654	193141	251305	164054	214599	300211	162011
	chr9	202518	207598	268544	237473	195851	250518	215493	275472	214887	131363
	chr10	170041	180512	217371	206164	176636	228940	172598	223687	200879	152130
	chr11	206608	212974	277225	246119	195114	249473	208985	267523	218344	137867
	chr12	158533	169341	196591	187579	131522	171671	134939	173276	237547	138510
	chr13	100165	107656	94887	91864	85664	113805	68337	89271	140915	65781
	chr14	121053	126388	169918	153646	118439	151131	123734	158738	123707	96879
	chr15	117916	123396	167275	153069	93976	123969	115883	149565	150170	98559
	chr16	229032	230250	320926	262305	267857	344587	303795	395459	80713	101161
	chr17	197729	200621	290706	240902	228699	293140	265230	338375	138395	107809
	chr18	95072	101013	103990	102843	88686	117420	73169	99545	111066	82529
	chr19	190341	186081	256916	200409	253705	318451	268521	333899	81074	69221
	chr20	116356	119033	166789	143921	121568	157601	136220	175842	116959	73370
	chr21	72000	73788	81457	70704	97923	124755	82870	106678	78949	34175
	chr22	144349	143708	206103	166253	189715	241792	224113	287195	54970	52737
	chrX	103535	113673	121050	128097	70861	96763	61063	81233	156140	92546
	chrY	24503	26958	31528	33847	23945	31597	21532	29099	35787	23994
chrM.fa		912	1117	1430	1595	235	369	224	339	407	545
humRibosomal.fa		11466	11805	4899	4923	14951	18207	1205	1586	2496	12413
newcontam.fa		5	1	3	4	14	39	11	25	20	24
Total	CG count	18791904	25068884	16207099	23856814	9646523	16984400	8987359	14171516	3134182	3100931
	# sequences	14811616	18256298	12544655	17288784	6804113	11052480	6474869	9662549	7436098	5316154
	CG rate (per kb)	35.24242955	38.14343351	35.8875708	38.33058923	39.38191094	42.68624679	38.55658873	40.740101	11.70783536	16.20287377
Mapped	CG count	7214851	7472873	8238755	8058058	6288173	9185179	6161898	8604773	1905496	1736594
	# sequence	7302423	8061039	7760391	8554604	4724503	6856012	4754005	6544390	5544993	3589393
	CG rate (per kb)	27.44466156	25.75099879	29.49004832	26.16543611	36.97139619	37.21461706	36.00413406	36.52310945	9.545628722	13.4392423
read length		36	36	36	36	36	36	36	36	36	36
HMM Analysis (# of Peaks)		43883	48298	46990	51215	56215	61615*	53085	56727*	1899	1398

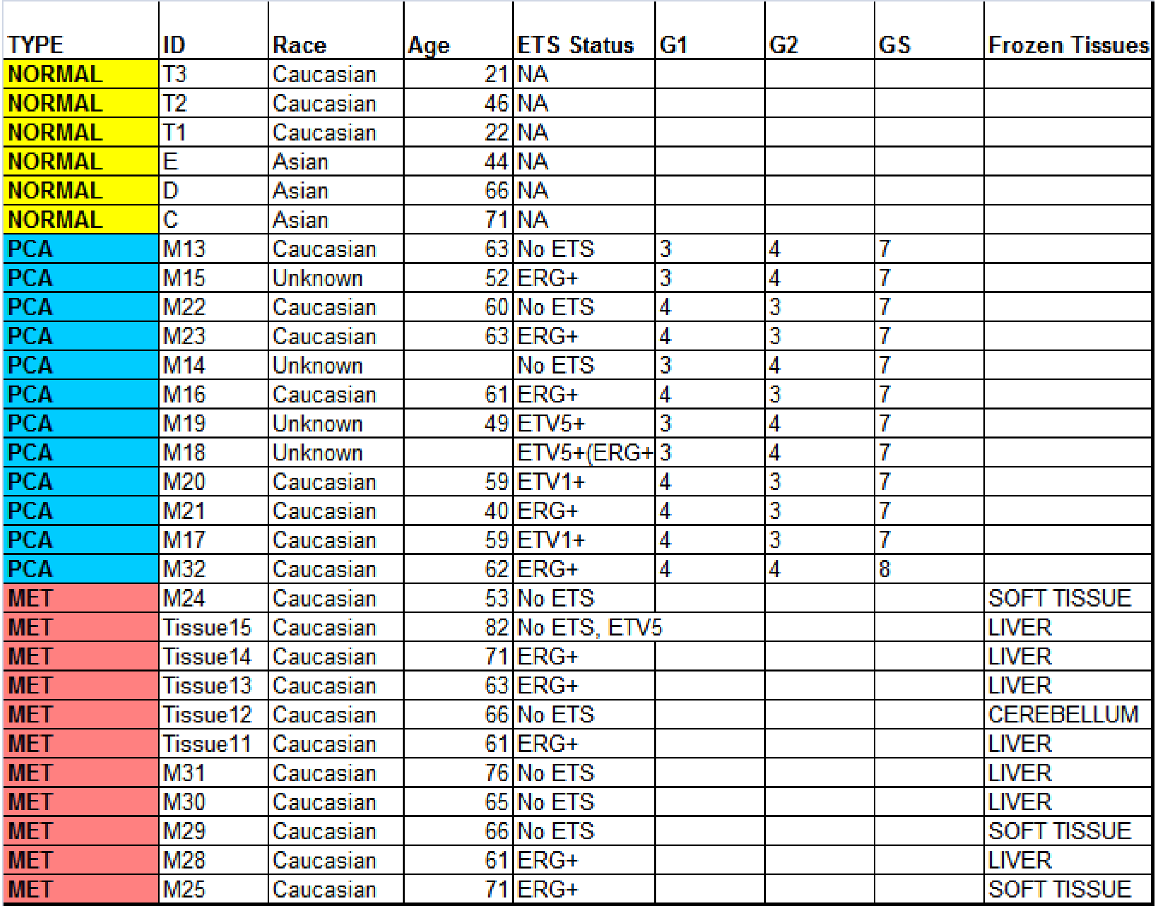
*These numbers are presented in Figure 4.4

Table 4.2 Oligonucleotide sequences



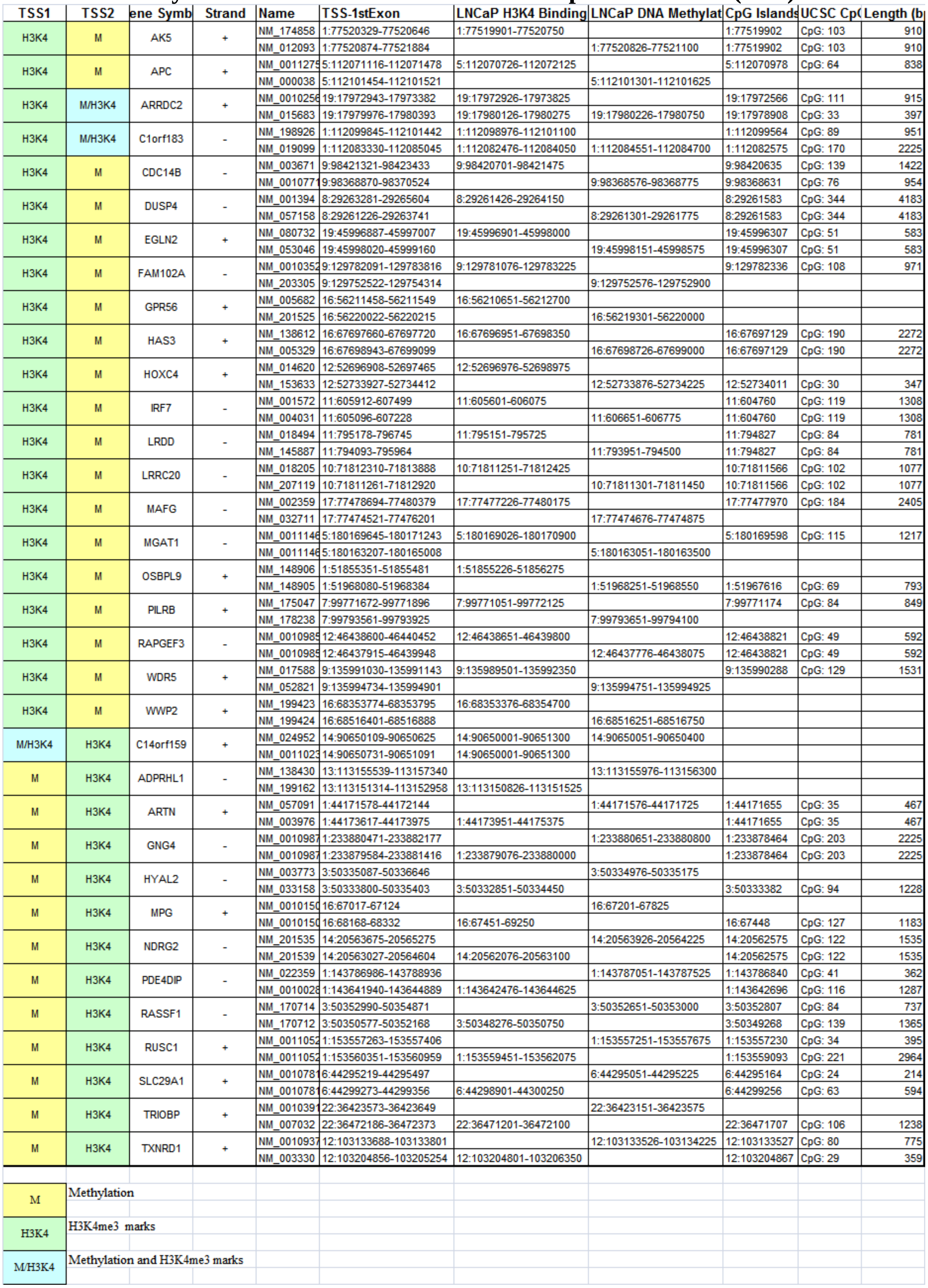
Gene	Genomic Location	Accession Number	Type	Forward Primer	Reverse Primer	Applications	Amplicon Size (bp)
RASSF1A	chr3:50353011-50353198	N/A	LNCaP-Hyper	GTAGTTTAATGAGTTTAGGTTTTTT	CTACACCCAAATTTCCATTAC	BS-sequencing	188
KCTD1	chr18:22381178-22381550	N/A	LNCaP-Hyper	GTTTTGGAGAGTATTTGTATT	AAACTCTTTTAAACTAACCATAAAC	BS-sequencing	373
C14orf23	chr14:28313174-28313411	N/A	LNCaP-Hyper	GAGTTTTTTTGTTTTAGAGTGGTTGA	ACAACTAAAATTACTTCTCCACACC	BS-sequencing	238
APC	chr5:112101275-112101503	N/A	LNCaP-Hyper	GGGTTAGGGTTAGGTAGGTTGT	ACACCTCCATTCTATCTCCAATAAC	BS-sequencing	229
CDKN2A	chr9:21960825-21961015	N/A	LNCaP-Hyper	ATTTTTTGTGTTGGAAAATGAATGT	AAACCTAACTAAAAAACTAAACCATC	BS-sequencing	191
SHC1	chr1:153209107-153209344	N/A	LNCaP-Hyper	GAATGGAGAGGAGGATGTTTATTTAT	AAAAACAACCATACCAAACTCAAAC	BS-sequencing	238
LAMC2	chr1:181421688-181421931	N/A	LNCaP-Hyper	TGTTTATTTTTGTGAATTTGTTTTG	ACCTATAAATAAAAAAAACCCAACC	BS-sequencing	244
TINAGL1	chr1:31814575-31814918	N/A	LNCaP-Hyper	ATTTTGTAAGTTTATGAGTTGTGGG	TTCTATTCCTATATCCCTCTATCCC	BS-sequencing	344
TSPAN1	chr1:46418579-46418802	N/A	LNCaP-Hyper	TGTTTTATTTTTGGGTTGTTTTTTT	CACACACCACTACTCACCTACAAAC	BS-sequencing	224
CALML3	chr10:5556581-5556830	N/A	LNCaP-Hyper	TTATTTAAGGGAAGAAAGGGTATTG	ATTTAAACAAAAATCCAAAACCTAC	BS-sequencing	250
SPON2-1	chr4:1156947-1157266	N/A	PrEC-Hyper	TAGTTTATATGTTGGAAGTGGTTGG	AAAACCCCTAAAAAAACTCTACACC	BS-sequencing	320
SPON2-2	chr4:1157164-1157488	N/A	PrEC-Hyper	TAGGAAGAGTTATAGAAAGGGGGTT	CATTAACAAAATTCCAAACATCAAA	BS-sequencing	325
MYC	chr8:128815497-128815710	N/A	Negative control	TTTGTTTTTGTTTTTATTTGATTTT	ATTACTCCTACCTCCAAACCTTTAC	BS-sequencing	214
APC	N/A	NM_001127511	Variant 1	TCAGTTCTCGGGTCCTGGAG	TCCTTGGCTACCCTTGGAC	qPCR	
APC	N/A	NM_001127510	Variant 2	GTGTCACTGGAGACAGAATGGA	TAGATTCACATCAGCCATCTGC	qPCR	
APC	N/A	NM_000038	Variant 3	AGGGTGTCACTGGAGACAGAAT	CTACCCTTGGACCCCATTTC	qPCR	
GAPDH	N/A	NM_002046	N/A	CTGAACGGGAAGCTCACTGGCA	TTACTCCTTGGAGGCCATGTGGG	qPCR	
GSTP1	N/A	NM_000852	N/A	GACTTGCTGCTGATCCATGA	AGGTTCACGTACTCAGGGGAG	qPCR	
NDRG2	N/A	NM_201535	Variant 1	CCTTGTTGTCCAACTTCTCCC	ACAGTGGCTTCTCCTCTGTGAT	qPCR	
		NM_016250	Variant 2				
		NM_201536	Variant 3				
		NM_201537	Variant 4				
NDRG2	N/A	NM_201538	Variant 5	GAGTCAAAGGCAAGTGAAGGTG	ACAGTGGCTTCTCCTCTGTGAT	qPCR	
		NM_201539	Variant 6				
		NM_201540	Variant 7				
		NM_201541	Variant 8				
RASSF1	N/A	NM_007182	Variant 1	GACCTCTGTGGCGACTTCAT	GGCAGGTGAACTTGCAATC	qPCR	
		NM_170714	Variant 4				
RASSF1	N/A	NM_170712	Variant 2	AGGTGGCCAACATTAGAGTCC	AACAGTCCAGGCAGACGAG	qPCR	
RASSF1	N/A	NM_170713	Variant 3	CTTCTTTCGAAATGACCTGGAG	TCCGAGTCCGAGTCCTCTT	qPCR	
APC	N/A	All Variants	1st round	N/A	TGCAATGGCCTGTAGTCCCCCTAGT	5'RACE	
			nested PCR	N/A	AGCCTTCGAGGTGCAGAGTGTGTGCT	5'RACE	
NDRG2	N/A	All Variants	1st round	N/A	CCAGGGGCATCCACATGAACCCGCA	5'RACE	
			nested PCR	N/A	CGCTGGGCGTTTGGGTTTGGGGGT	5'RACE	
RASSF1	N/A	All Variants	1st round	N/A	CAGAGCCATACCTGGCTACAC	5'RACE	
			nested PCR	N/A	GCCGCAGGGGCTGCTCATCATCCA	5'RACE	
N/A	Not Applicable						

Table 4.3 Clinical information of prostate tissue samples profiled by Differential Methylation Hybridization (DMH) Analysis



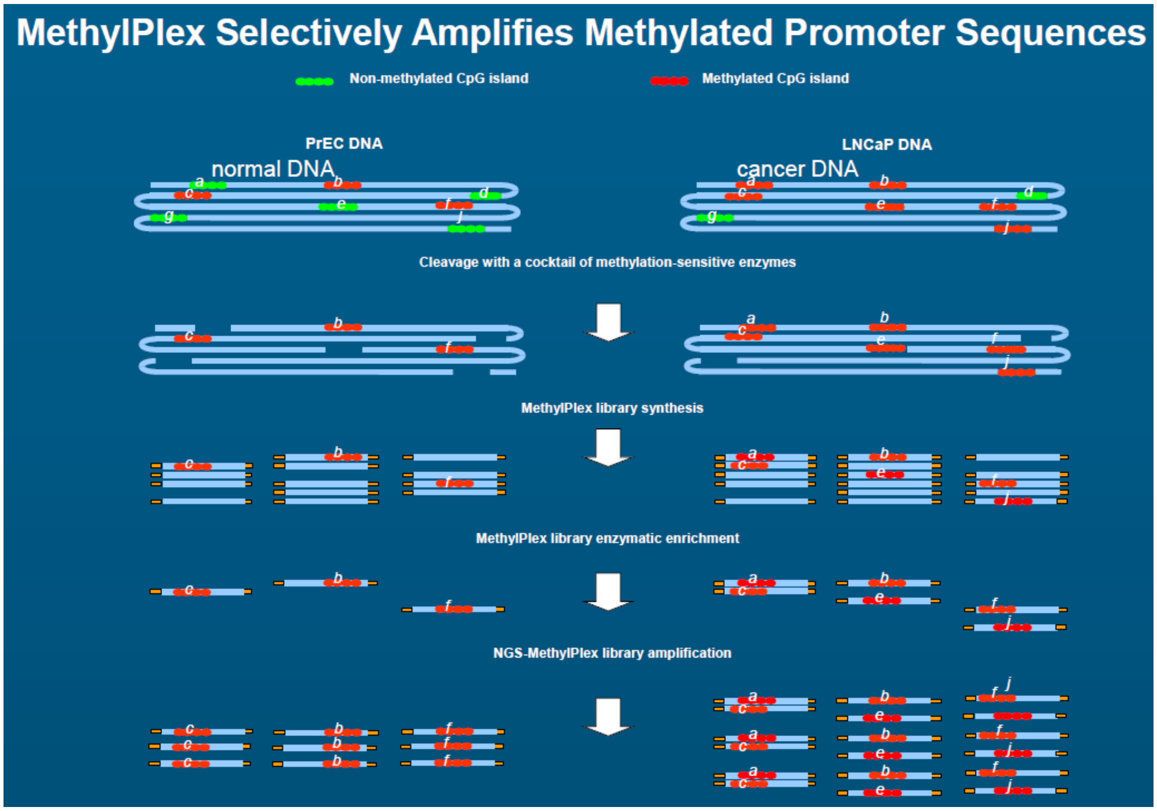
TYPE	ID	Race	Age	ETS Status	G1	G2	GS	Frozen Tissues
NORMAL	T3	Caucasian	21	NA				
NORMAL	T2	Caucasian	46	NA				
NORMAL	T1	Caucasian	22	NA				
NORMAL	E	Asian	44	NA				
NORMAL	D	Asian	66	NA				
NORMAL	C	Asian	71	NA				
PCA	M13	Caucasian	63	No ETS	3	4	7	
PCA	M15	Unknown	52	ERG+	3	4	7	
PCA	M22	Caucasian	60	No ETS	4	3	7	
PCA	M23	Caucasian	63	ERG+	4	3	7	
PCA	M14	Unknown		No ETS	3	4	7	
PCA	M16	Caucasian	61	ERG+	4	3	7	
PCA	M19	Unknown	49	ETV5+	3	4	7	
PCA	M18	Unknown		ETV5+(ERG+	3	4	7	
PCA	M20	Caucasian	59	ETV1+	4	3	7	
PCA	M21	Caucasian	40	ERG+	4	3	7	
PCA	M17	Caucasian	59	ETV1+	4	3	7	
PCA	M32	Caucasian	62	ERG+	4	4	8	
MET	M24	Caucasian	53	No ETS				SOFT TISSUE
MET	Tissue15	Caucasian	82	No ETS, ETV5				LIVER
MET	Tissue14	Caucasian	71	ERG+				LIVER
MET	Tissue13	Caucasian	63	ERG+				LIVER
MET	Tissue12	Caucasian	66	No ETS				CEREBELLUM
MET	Tissue11	Caucasian	61	ERG+				LIVER
MET	M31	Caucasian	76	No ETS				LIVER
MET	M30	Caucasian	65	No ETS				LIVER
MET	M29	Caucasian	66	No ETS				SOFT TISSUE
MET	M28	Caucasian	61	ERG+				LIVER
MET	M25	Caucasian	71	ERG+				SOFT TISSUE

Table 4.4 Methylation marks and alternate transcription start site (TSS)



TSS1	TSS2	ene Symb	Strand	Name	TSS-1stExon	LNCaP H3K4 Binding	LNCaP DNA Methylat	CpG Islands	UCSC Cp(	Length (b
H3K4	M	AK5	+	NM_174858	1:77520329-77520646	1:77519901-77520750		1:77519902	CpG: 103	910
				NM_012093	1:77520874-77521884		1:77520826-77521100	1:77519902	CpG: 103	910
H3K4	M	APC	+	NM_001127	5:112071116-112071478	5:112070726-112072125		5:112070978	CpG: 64	838
				NM_000038	5:112101454-112101521		5:112101301-112101625			
H3K4	M/H3K4	ARRDC2	+	NM_001025	19:17972943-17973382	19:17972926-17973825		19:17972566	CpG: 111	915
				NM_015683	19:17979976-17980393	19:17980126-17980275	19:17980226-17980750	19:17978908	CpG: 33	397
H3K4	M/H3K4	C1orf183	-	NM_198926	1:112099845-112101442	1:112098976-112101100		1:112099564	CpG: 89	951
				NM_019099	1:112083330-112085045	1:112082476-112084050	1:112084551-112084700	1:112082575	CpG: 170	2225
H3K4	M	CDC14B	-	NM_003671	9:98421321-98423433	9:98420701-98421475		9:98420635	CpG: 139	1422
				NM_001077	9:98368870-98370524		9:98368576-98368775	9:98368631	CpG: 76	954
H3K4	M	DUSP4	-	NM_001394	8:29263281-29265604	8:29261426-29264150		8:29261583	CpG: 344	4183
				NM_057158	8:29261226-29263741		8:29261301-29261775	8:29261583	CpG: 344	4183
H3K4	M	EGLN2	+	NM_080732	19:45996887-45997007	19:45996901-45998000		19:45996307	CpG: 51	583
				NM_053046	19:45998020-45999160		19:45998151-45998575	19:45996307	CpG: 51	583
H3K4	M	FAM102A	-	NM_001035	9:129782091-129783816	9:129781076-129783225		9:129782336	CpG: 108	971
				NM_203305	9:129752522-129754314		9:129752576-129752900			
H3K4	M	GPR56	+	NM_005682	16:56211458-56211549	16:56210651-56212700				
				NM_201525	16:56220022-56220215		16:56219301-56220000			
H3K4	M	HAS3	+	NM_138612	16:67697660-67697720	16:67696951-67698350		16:67697129	CpG: 190	2272
				NM_005329	16:67698943-67699099		16:67698726-67699000	16:67697129	CpG: 190	2272
H3K4	M	HOXC4	+	NM_014620	12:52696908-52697465	12:52696976-52698975				
				NM_153633	12:52733927-52734412		12:52733876-52734225	12:52734011	CpG: 30	347
H3K4	M	IRF7	-	NM_001572	11:605912-607499	11:605601-606075		11:604760	CpG: 119	1308
				NM_004031	11:605096-607228		11:606651-606775	11:604760	CpG: 119	1308
H3K4	M	LRDD	-	NM_018494	11:795178-796745	11:795151-795725		11:794827	CpG: 84	781
				NM_145887	11:794093-795964		11:793951-794500	11:794827	CpG: 84	781
H3K4	M	LRRC20	-	NM_018205	10:71812310-71813888	10:71811251-71812425		10:71811566	CpG: 102	1077
				NM_207119	10:71811261-71812920		10:71811301-71811450	10:71811566	CpG: 102	1077
H3K4	M	MAFG	-	NM_002359	17:77478694-77480379	17:77477226-77480175		17:77477970	CpG: 184	2405
				NM_032711	17:77474521-77476201		17:77474676-77474875			
H3K4	M	MGAT1	-	NM_001114	5:180169645-180171243	5:180169026-180170900		5:180169598	CpG: 115	1217
				NM_001114	5:180163207-180165008		5:180163051-180163500			
H3K4	M	OSBPL9	+	NM_148906	1:51855351-51855481	1:51855226-51856275				
				NM_148905	1:51968080-51968384		1:51968251-51968550	1:51967616	CpG: 69	793
H3K4	M	PILRB	+	NM_175047	7:99771672-99771896	7:99771051-99772125		7:99771174	CpG: 84	849
				NM_178238	7:99793561-99793925		7:99793651-99794100			
H3K4	M	RAPGEF3	-	NM_001098	12:46438600-46440452	12:46438651-46439800		12:46438821	CpG: 49	592
				NM_001098	12:46437915-46439948		12:46437776-46438075	12:46438821	CpG: 49	592
H3K4	M	WDR5	+	NM_017588	9:135991030-135991143	9:135989501-135992350		9:135990288	CpG: 129	1531
				NM_052821	9:135994734-135994901		9:135994751-135994925			
H3K4	M	WWP2	+	NM_199423	16:68353774-68353795	16:68353376-68354700				
				NM_199424	16:68516401-68516888		16:68516251-68516750			
M/H3K4	H3K4	C14orf159	+	NM_024952	14:90650109-90650625	14:90650001-90651300	14:90650051-90650400			
				NM_001102	14:90650731-90651091	14:90650001-90651300				
M	H3K4	ADPRHL1	-	NM_138430	13:113155539-113157340		13:113155976-113156300			
				NM_199162	13:113151314-113152958	13:113150826-113151525				
M	H3K4	ARTN	+	NM_057091	1:44171578-44172144		1:44171576-44171725	1:44171655	CpG: 35	467
				NM_003976	1:44173617-44173975	1:44173951-44175375		1:44171655	CpG: 35	467
M	H3K4	GNG4	-	NM_001098	1:233880471-233882177		1:233880651-233880800	1:233878464	CpG: 203	2225
				NM_001098	1:233879584-233881416	1:233879076-233880000		1:233878464	CpG: 203	2225
M	H3K4	HYAL2	-	NM_003773	3:50335087-50336646		3:50334976-50335175			
				NM_033158	3:50333800-50335403	3:50332851-50334450		3:50333382	CpG: 94	1228
M	H3K4	MPG	+	NM_001015	16:67017-67124		16:67201-67825			
				NM_001015	16:68168-68332	16:67451-69250		16:67448	CpG: 127	1183
M	H3K4	NDRG2	-	NM_201535	14:20563675-20565275		14:20563926-20564225	14:20562575	CpG: 122	1535
				NM_201539	14:20563027-20564604	14:20562076-20563100		14:20562575	CpG: 122	1535
M	H3K4	PDE4DIP	-	NM_022359	1:143786986-143788936		1:143787051-143787525	1:143786840	CpG: 41	362
				NM_001002	1:143641940-143644889	1:143642476-143644625		1:143642696	CpG: 116	1287
M	H3K4	RASSF1	-	NM_170714	3:50352990-50354871		3:50352651-50353000	3:50352807	CpG: 84	737
				NM_170712	3:50350577-50352168	3:50348276-50350750		3:50349268	CpG: 139	1365
M	H3K4	RUSC1	+	NM_001105	1:153557263-153557406		1:153557251-153557675	1:153557230	CpG: 34	395
				NM_001105	1:153560351-153560959	1:153559451-153562075		1:153559093	CpG: 221	2964
M	H3K4	SLC29A1	+	NM_001078	6:44295219-44295497		6:44295051-44295225	6:44295164	CpG: 24	214
				NM_001078	6:44299273-44299356	6:44298901-44300250		6:44299256	CpG: 63	594
M	H3K4	TRIOBP	+	NM_001039	22:36423573-36423649		22:36423151-36423575			
				NM_007032	22:36472186-36472373	22:36471201-36472100		22:36471707	CpG: 106	1238
M	H3K4	TXNRD1	+	NM_001093	12:103133688-103133801		12:103133526-103134225	12:103133527	CpG: 80	775
				NM_003330	12:103204856-103205254	12:103204801-103206350		12:103204867	CpG: 29	359

M	Methylation
H3K4	H3K4me3 marks
M/H3K4	Methylation and H3K4me3 marks



MethylPlex Selectively Amplifies Methylated Promoter Sequences

🟩 Non-methylated CpG island

🟥 Methylated CpG island

PrEC DNA

normal DNA

LNCaP DNA

cancer DNA

Cleavage with a cocktail of methylation-sensitive enzymes

MethylPlex library synthesis

MethylPlex library enzymatic enrichment

NGS-MethylPlex library amplification

Figure 4.1 The schematic of Methylplex library sample preparation.

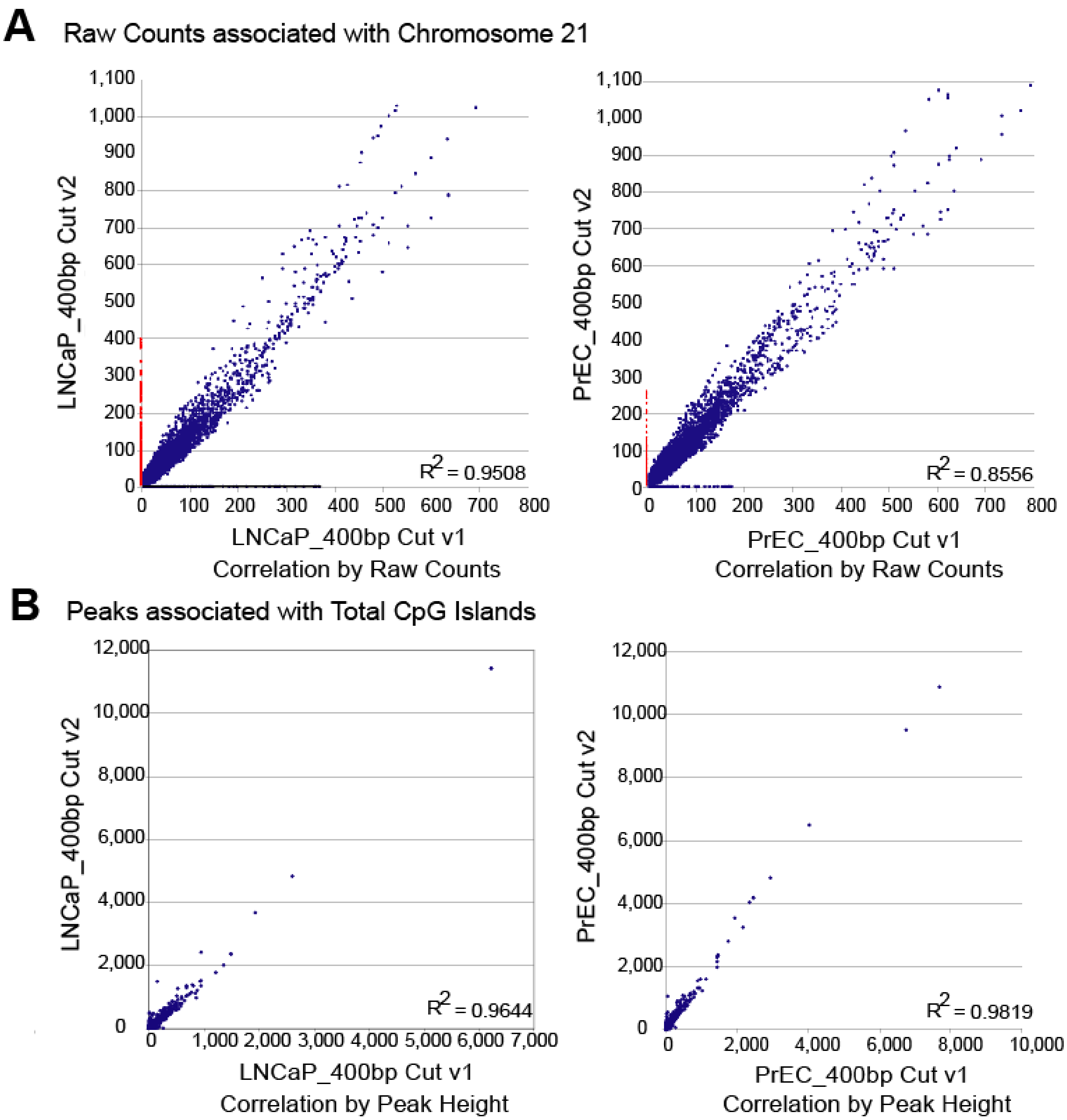


Figure 4.2 Regression analysis of M-NGS mapped reads and HMM output. (A) Reads that mapped to chromosome 21 in LNCaP400bp-1 and -5, and PrEC400bp-1 and -5 runs were compared using the window size of 25bp. In LNCaP samples, a total of 33,627 reads were present at 25 bp windows with R^2 value of 0.9508, and in PrEC, 37,406 reads with R^2 value of 0.8556 was observed. (B) Linear regression analysis of all DNA methylation that occurred on CGIs showed high correlation (R^2 value = 0.9398 and 0.9819, n=5,734 and 4,966, respectively).

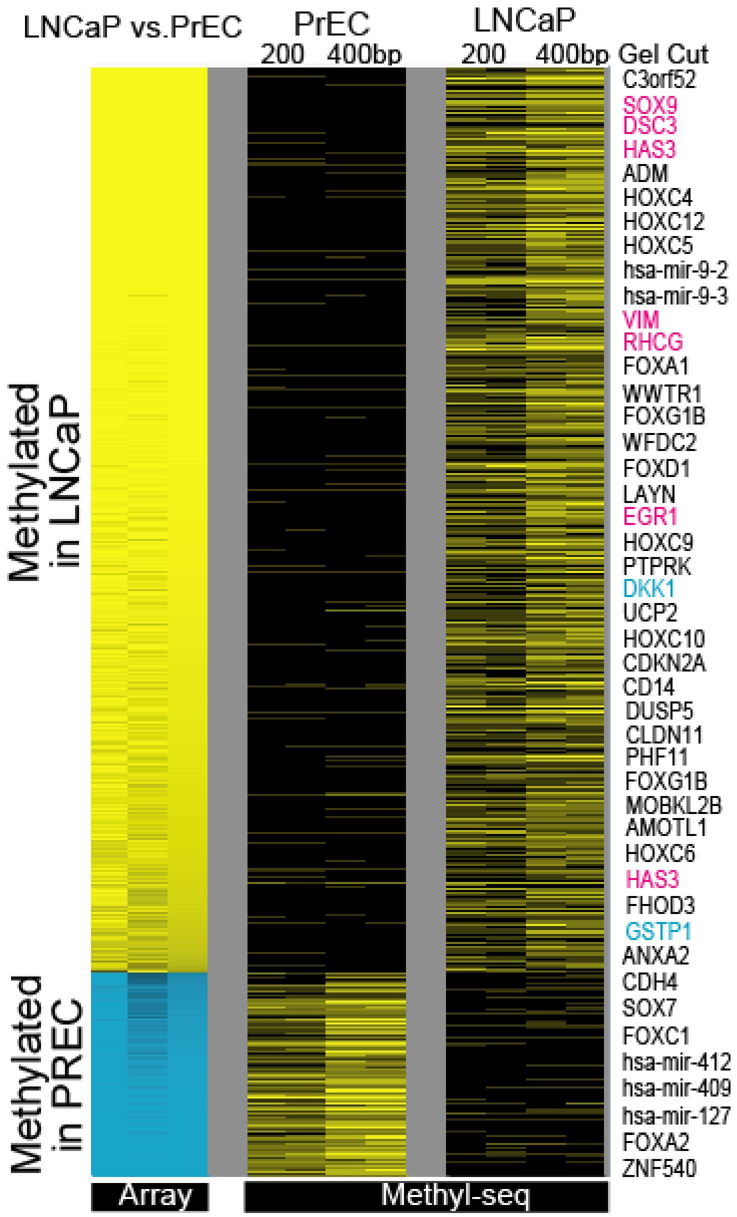


Figure 4.3 High correlation between Methylplex-array and M-NGS results. Methylplex-array libraries made from LNCaP (Cy5) and PrEC (Cy3) cells were hybridized to Agilent human CGI microarray. Array results are displayed on the left in heatmap form (*Yellow*: hypermethylated in LNCaP; *Blue:* hypermethylated in PrEC), and were compared to M-NGS results (*Yellow:* methylated regions) on right. PrEC and LNCaP 200/400, indicates data obtained by M-NGS from size-selected bands excised at 200 and 400 bp during sample preparation. Gene names are displayed on the right and genes previously identified as methylated in prostate cancer are indicated in blue, with genes known to be methylated in other cancers indicated in pink.

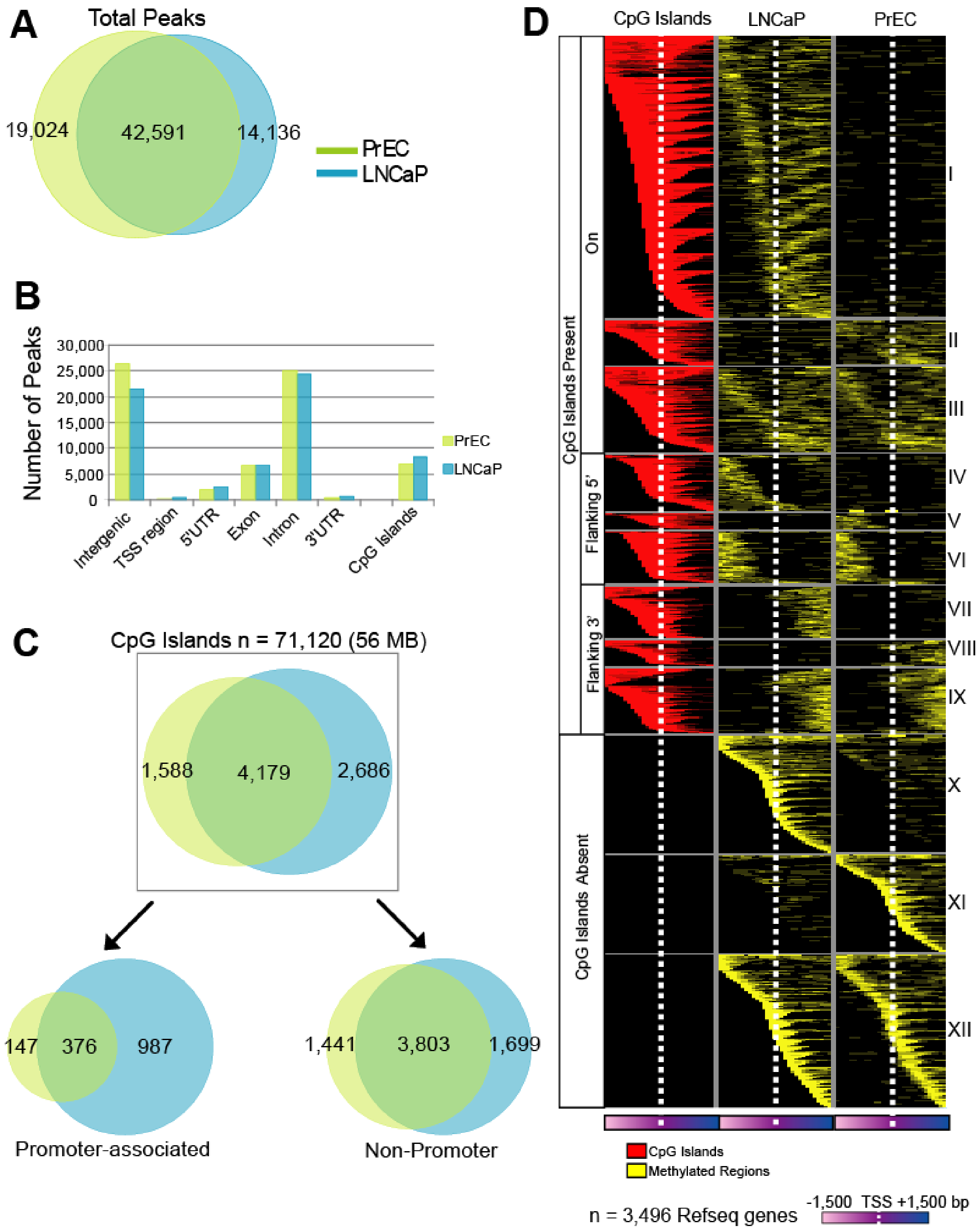


Figure 4.4 Distinct patterns of promoter methylation revealed by Methylplex-Next generation sequencing (M-NGS) of prostate cells. (A) Venn diagram representing a 70% overlap between the regions methylated in LNCaP (blue) and PrEC cells (green). (B) DNA methylation in intergenic and intronic regions represents a large proportion of all DNA methylation and had similar number of total methylated regions in both LNCaP

(blue bars) and PrEC (green bars) cells. (C) While there is a 7 fold increase in the methylation of promoter associated CpG islands in LNCaP (blue) compared to PrEC (green) cells, this difference was not seen in non-promoter associated CpG islands. (D) Analysis of gene promoter regions (±1,500 bp from TSS) identified methylation in 3,496 Refseq genes in either LNCaP and PrEC cells or both. Each row represents a unique promoter region, ±1,500 bp from the transcription start site (white dotted line) at 100 bp window size. CpG island location is indicated in red in the first column. The methylation (yellow) observed in the corresponding location in each cell line is indicated (LNCaP second column and PrEC third column). Among promoters represented, 3 distinct patterns of methylation were found. Methylation occurred either (1) on CpG islands, (2) in regions flanking the island (5' or 3') or (3) in promoters without any CpG island.

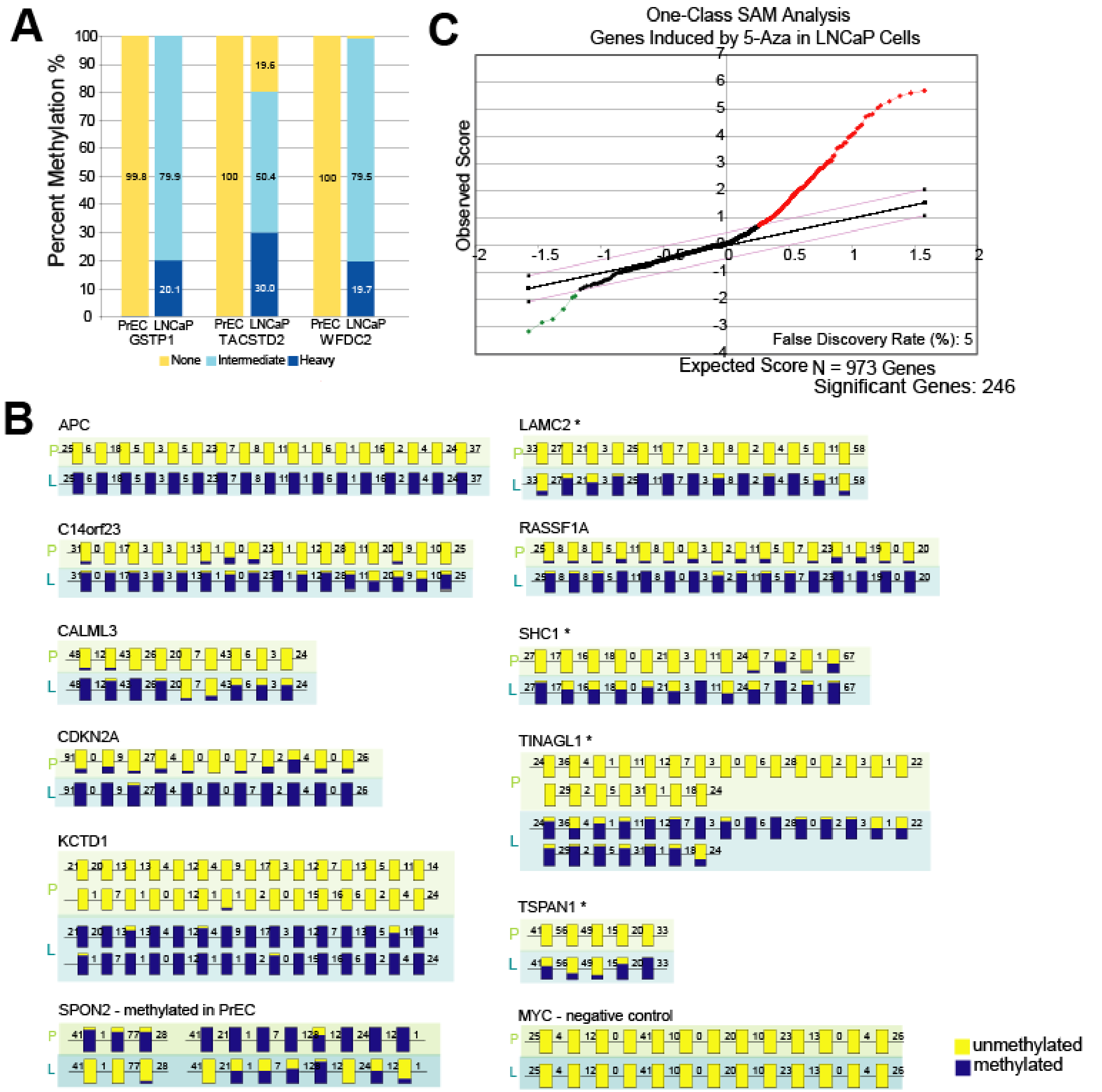


Figure 4.5 Validation of differentially methylated regions predicted by M-NGS. (A) The Methyl-Profiler qPCR assay validation of the methylation of GSTP1, TACSTD2 and WFDC2 gene promoters in LNCaP cells. (B) Bisulfite sequencing validation of methylation of APC, C14orf23, CALML3, CDKN2A, KCTD1, LAMC2, RASSF1A, SHC1, TINAGL1 and TSPAN1 gene promoters in LNCaP cells and SPON2 in PrEC cells. Methylation status of each CG residue was analyzed using the BIQ Analyzer (38) program, where the height of the blue bar indicates percent methylation at a given position, yellow indicates no methylation. The number between each bar indicates the distance between each CG residue. * CpG islands were absent in these promoters. (C) Methylated target genes (n=973) that contained promoter methylation in LNCaP (depicted in Fig. 4.4D) are re-expressed after 5-Aza treatment (red points, n=246).

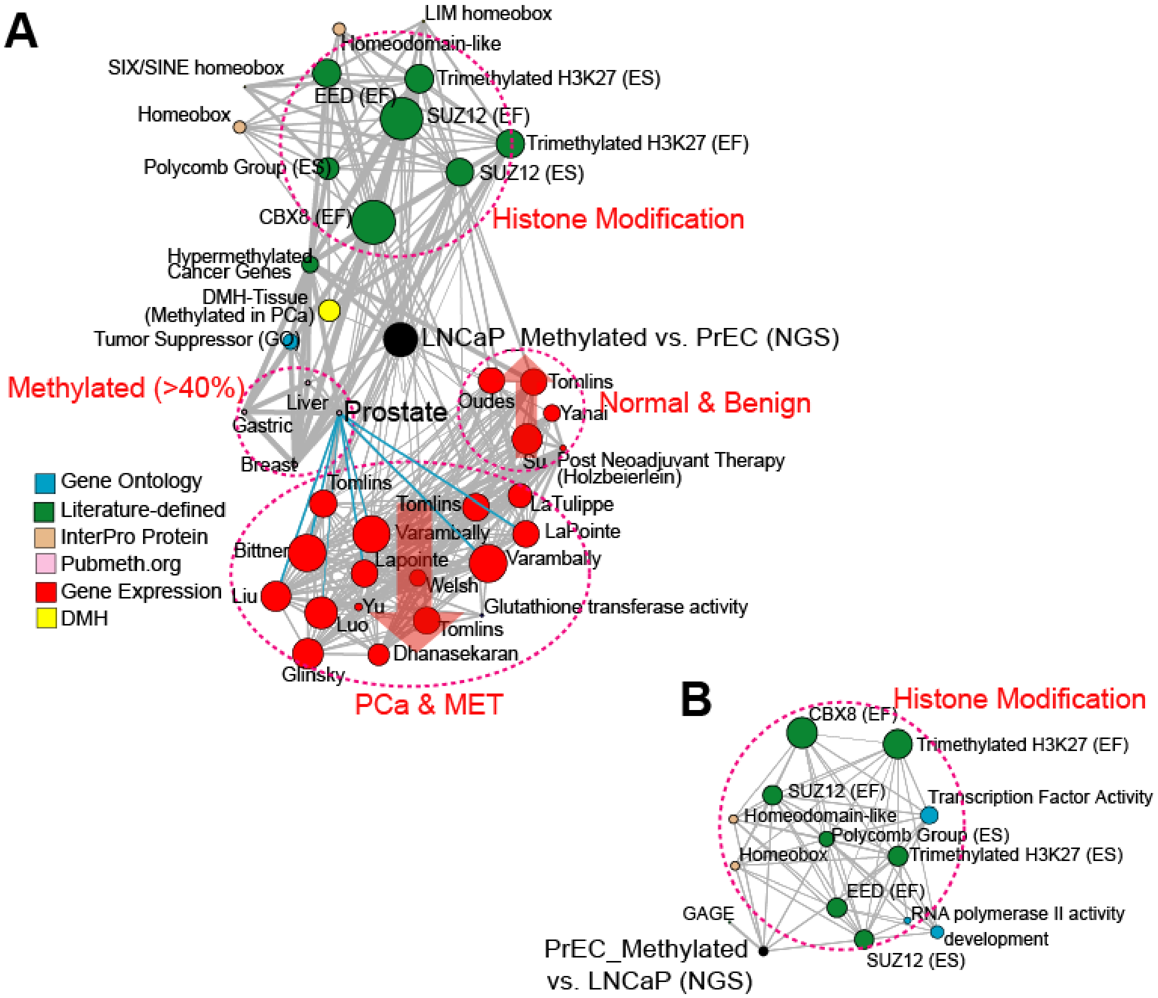


Figure 4.6 Molecular Concept Map (MCM) analysis of LNCaP and PrEC methylated genes. (A) MCM analysis of LNCaP methylated genes (black) revealed enrichment of gene signatures (red) repressed in prostate cancer and over-expressed in benign prostate. Histone modification concepts (green), gene ontology concepts such as tumor suppressor genes (blue), genes previously known to be methylated from the Pubmeth database(8) (pink) and genes methylated in prostate cancer by differential methylation hybridization (DMH, yellow) EF = embryonic fibroblast, ES = embryonic stem cells. (B) MCM analysis of PrEC methylated genes (black) shows the enrichment only for histone modification concept.

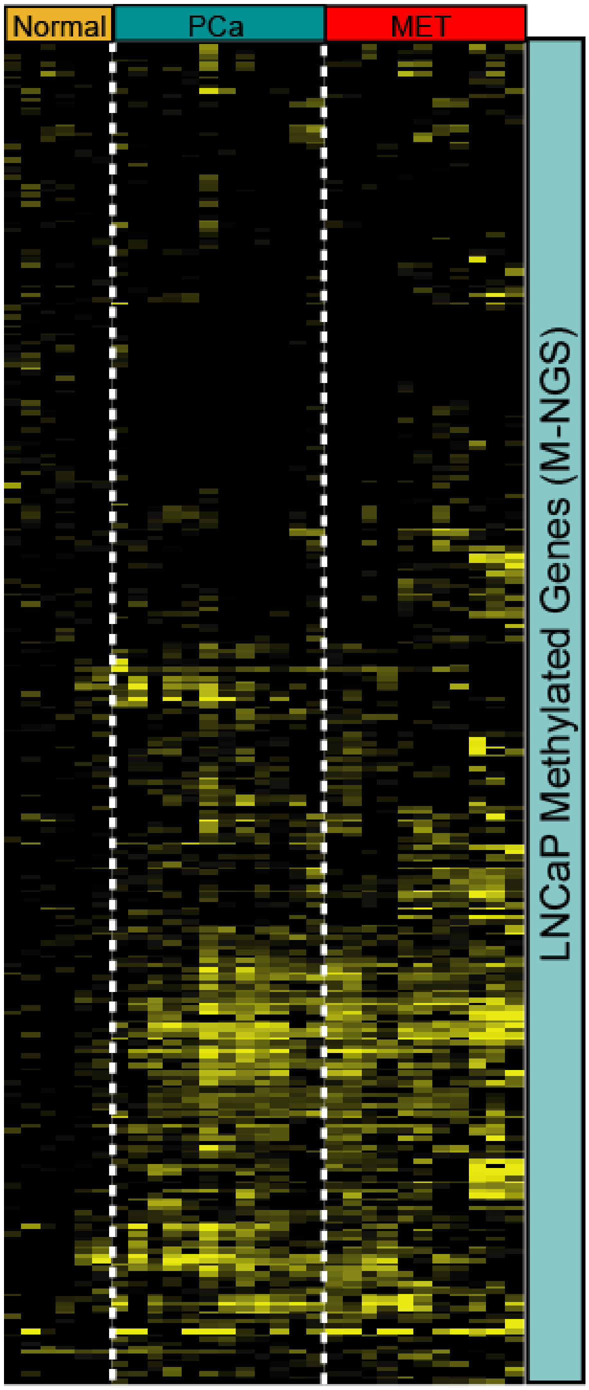


Figure 4.7 The overlap between differentially methylated regions identified by DMH and M-NGS. The heatmap depicts methylation status of probes in a 12K CpG island array that overlap with differentially methylated regions identified by LNCaP M-NGS. Forty percent (134/309) of the probes were methylated in at least one prostate cancer tissue analyzed (yellow). PCa, clinically localized prostate cancer; MET, metastatic prostate cancer.

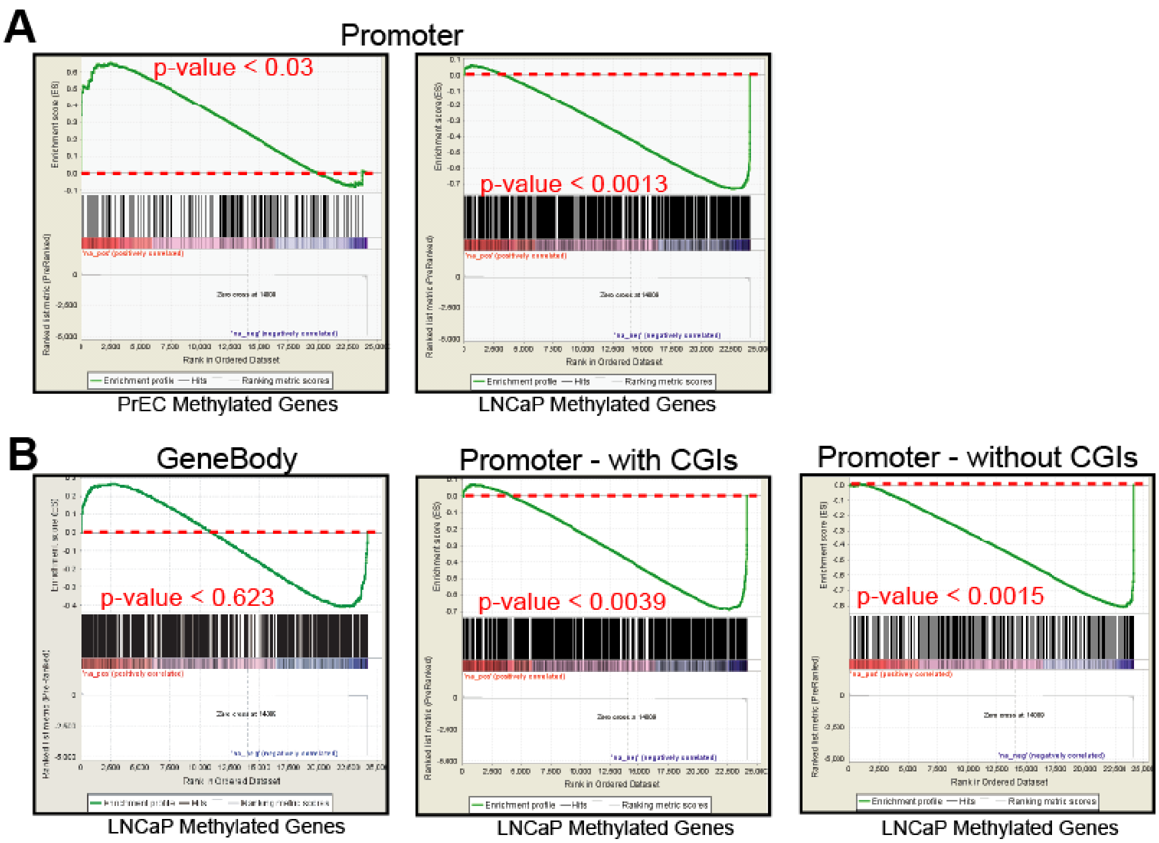


Figure 4.8 The association between gene repression and promoter methylation.via Gene Set Enrichment Analysis (GSEA) (A) Methylated genes from LNCaP and PrEC cells were tested for their corresponding ranked gene expression in next generation transcriptomic sequencing. Both LNCaP and PrEC methylated promoters show enrichment with gene repression in LNCaP (p-value<0.0013) and PrEC (p-value<0.03) respectively. (B) While no significant association was observed between gene body methylation and gene expression in LNCaP (p-value<0.623), candidates with gene promoter methylation with and without the presence of CGIs in LNCaP cells are enriched with under-expressed genes (p-value<0.0039 and 0.0015, respectively).

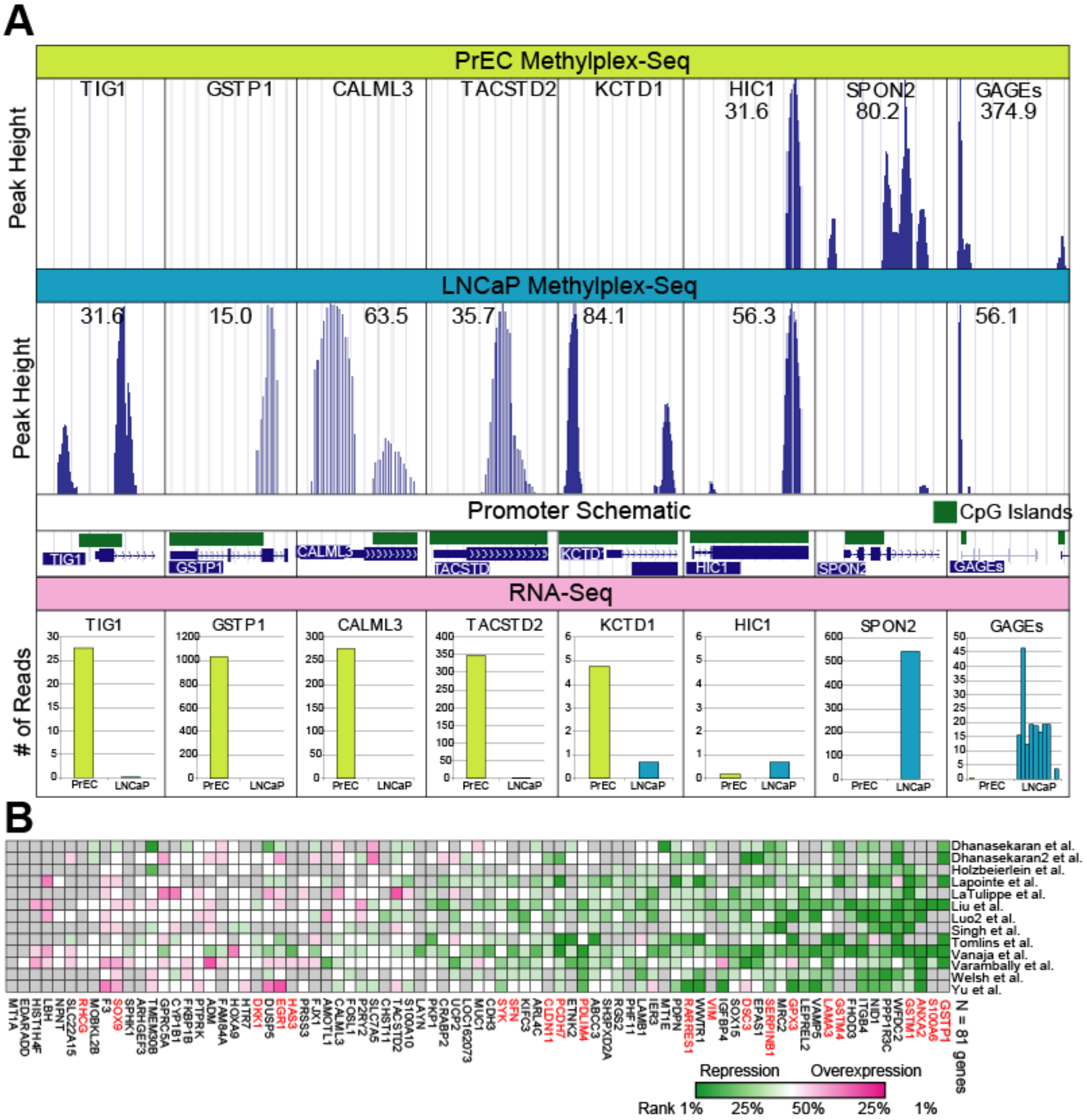


Figure 4.9 The association between DNA methylation and gene repression. Methylation status as determined by M-NGS of a gene was correlated to transcript expression assessed by RNA-Seq. (A) Genes with promoter methylation in LNCaP cells (TIG1, GSTP1, CALML3, TACSTD2, KCTD1) and PrEC cells (SPON2, GAGEs) had low transcript expression in their corresponding cell lines. HIC1, which is methylated in both LNCaP and PrEC, was minimally expressed in both. (B) Meta-analysis (cancer vs. normal) on 13 different prostate cancer datasets for genes methylated and repressed in LNCaP cells were also mostly repressed in tumors. Genes are ranked according to percent repression across the datasets, with GSTP1 indicated as the top-repressed gene in this meta-analysis. Forty four out of 81 methylated genes appear within the top 25% of repressed genes in at least one study. Previously characterized methylated genes are indicated (red). Analysis was derived from www.oncomine.org (see Materials and Methods).

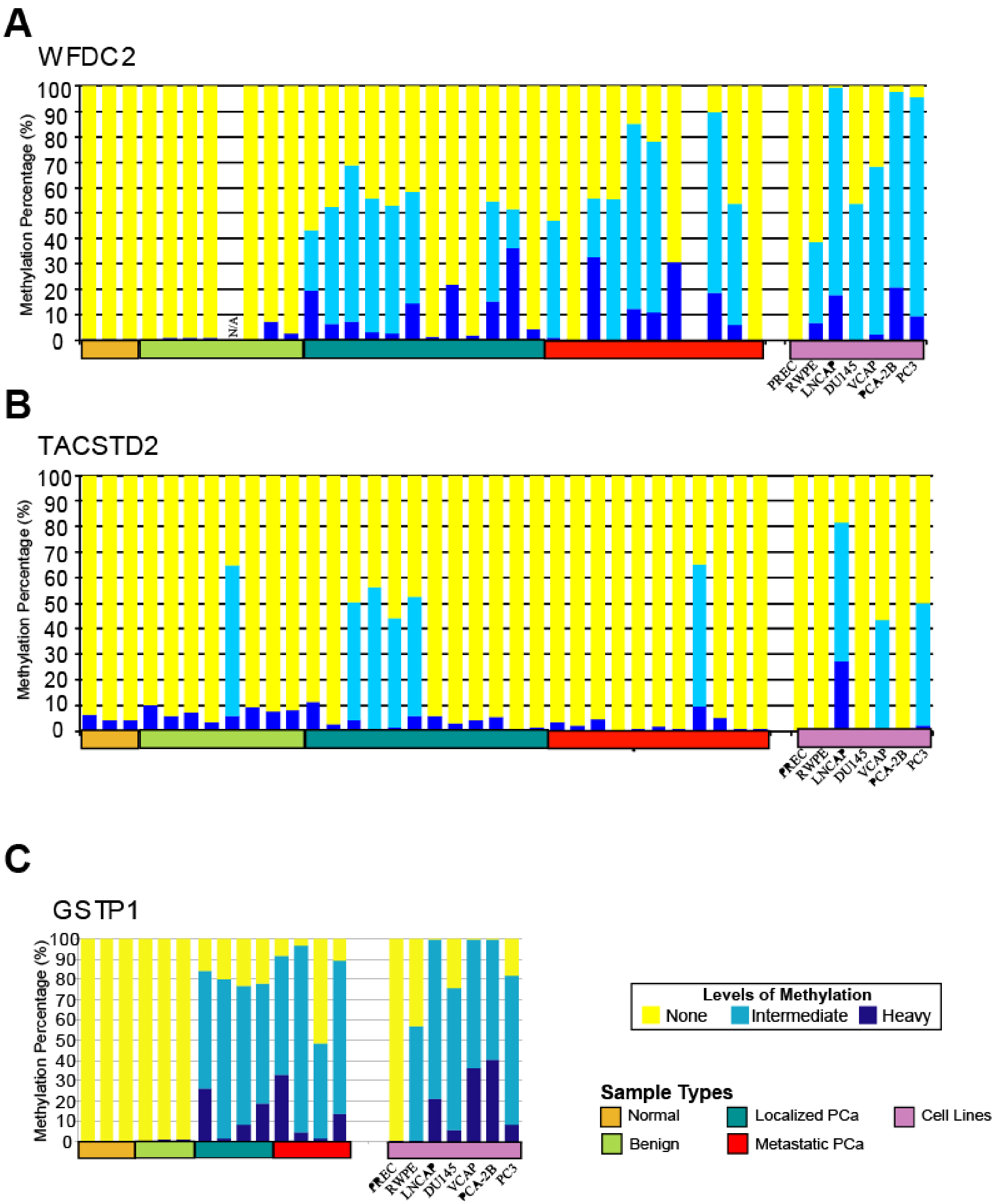


Figure 4.10 Methylation target gene validation on prostate tissues. DNA methylation is associated with gene repression and WFDC2 and TACSTD2 are validated as methylated target genes in both LNCaP and prostate cancer tissues. (A) WFDC2, (B) TACSTD2 and (C) GSTP1 genes were assessed by qPCR on 14 prostate tissue samples

(3 normal, 8 benign adjacent, 12 PCa, and 11 Mets) and 7 different prostate cell lines. GSTP1 and WFDC2 shows high level of methylation in majority of cancer samples (n = 17 out of 22 samples), compared to normal and benign prostate specimens (n = 10), while TACSTD2 was methylated predominantly in localized PCa samples.

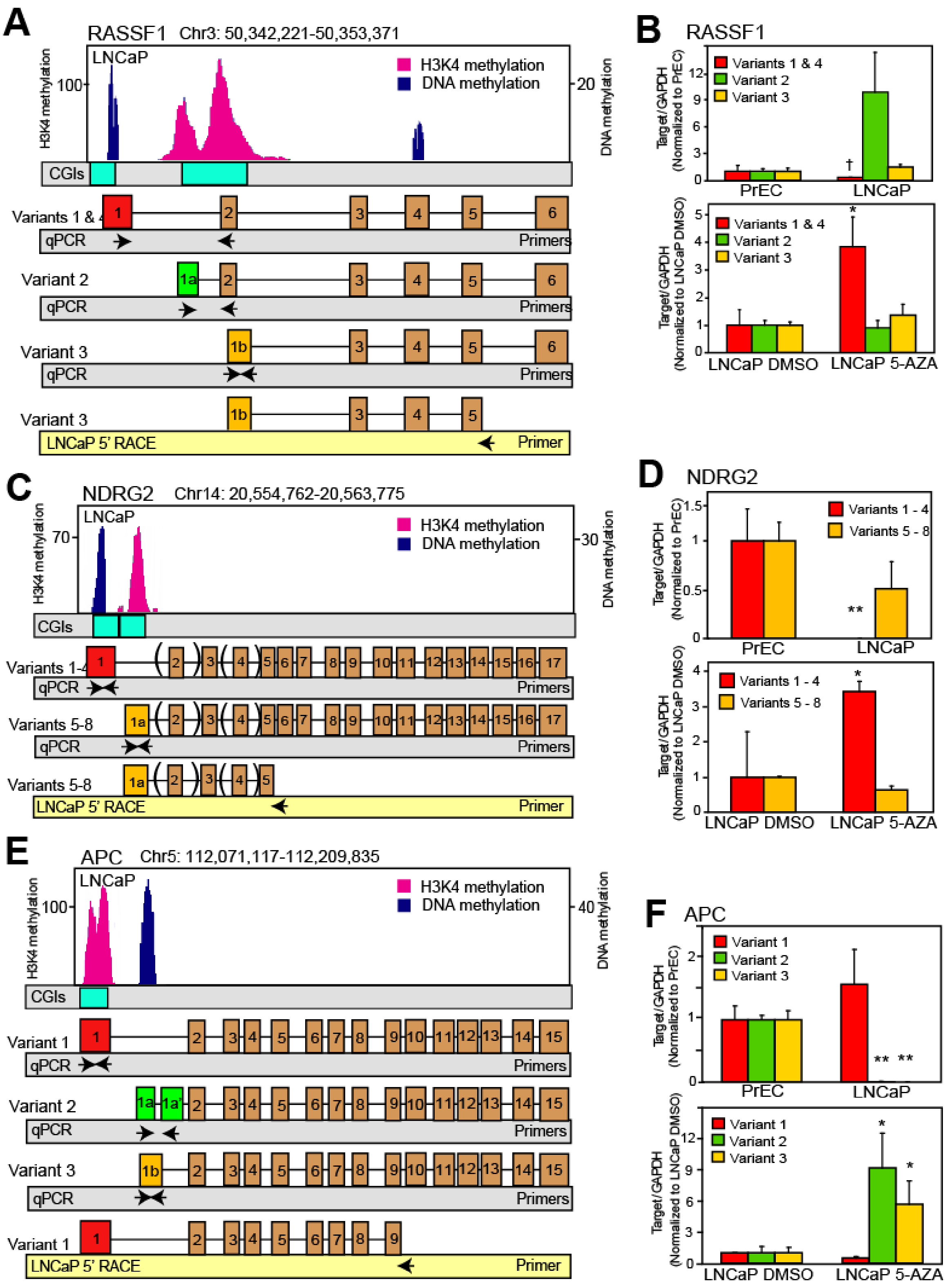


Figure 4.11 Cancer-specific DNA methylation enables switching of alternate transcriptional start sites (TSS) leading to transcript isoform regulation. Next generation sequencing for DNA methylation and histone 3 lysine 4 trimethylation

(H3K4me3) in LNCaP cells reveals genome-wide patterning that couples CpG methylation with H3K4 marks to repress or activate, specific transcript variants (A,C,E). Independent epigenetic modifications mark specific alternative TSS. In RASSF1 (A) and NDRG2 (C), CpG methylation occurs at the TSS of the longer variants, with H3K4me3 marks positioned on the TSS of the shorter variants. By contrast, APC (E) exhibits the reverse, with H3K4me3 on the longer variant 1 and CpG methylation at the shorter variants 2 and 3. (B,D,F) Preferential silencing and 5-Aza-induced re-expression of CpG-methylated variants in LNCaP cells. Variants exhibiting CpG methylation on their TSSs show preferential silencing compared to variants with H3K4me3 marks in LNCaP cells. These variants show preferential re-expression upon treatment of 48 hr androgen-starved LNCaP cells with 6uM 5-Aza. qRT-PCR data is normalized to variant expression levels in PrEC prostate primary epithelial cells or DMSO-treated LNCaP cells in the respective panels. (A,C,E) 5’RACE results validated RASSF1 variant-3, NRDG2 variants 5-8 and APC variant-1 expression in LNCaP cells.

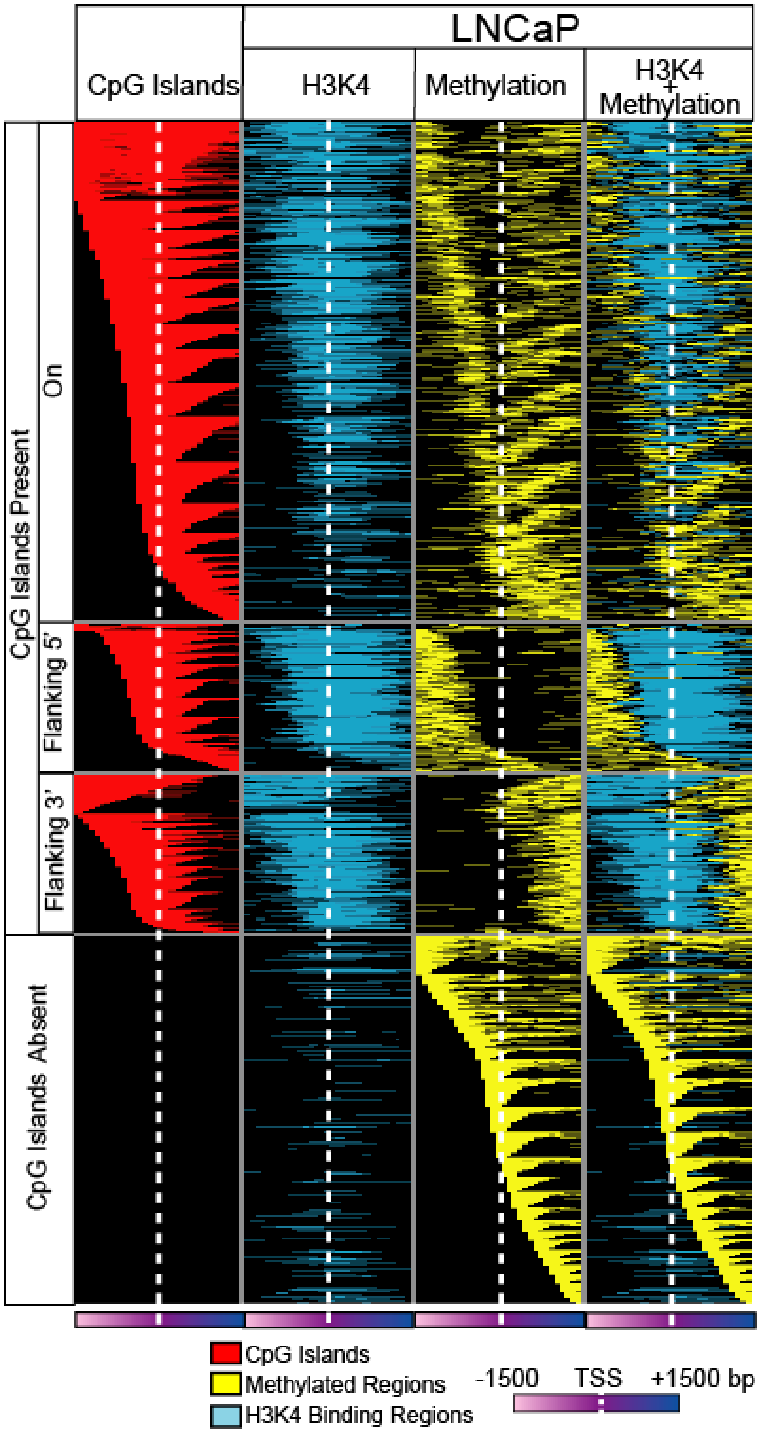


Figure 4.12 Promoter DNA methylation and histone H3K4me3 marks in LNCaP prostate cancer cell line. Promoter DNA methylation and histone H3K4me3 marks are mutually exclusive in LNCaP. Each row represents a unique promoter region, ±1,500 bps of the transcription start site (white dotted line) at 100 bp window size. CpG island location is indicated in red in first column. The second column represents histone H3K4me3 marks (blue), and the third column (yellow) depicts DNA methylation

observed in the corresponding location in LNCaP. Superimposed data is displayed in the fourth column.

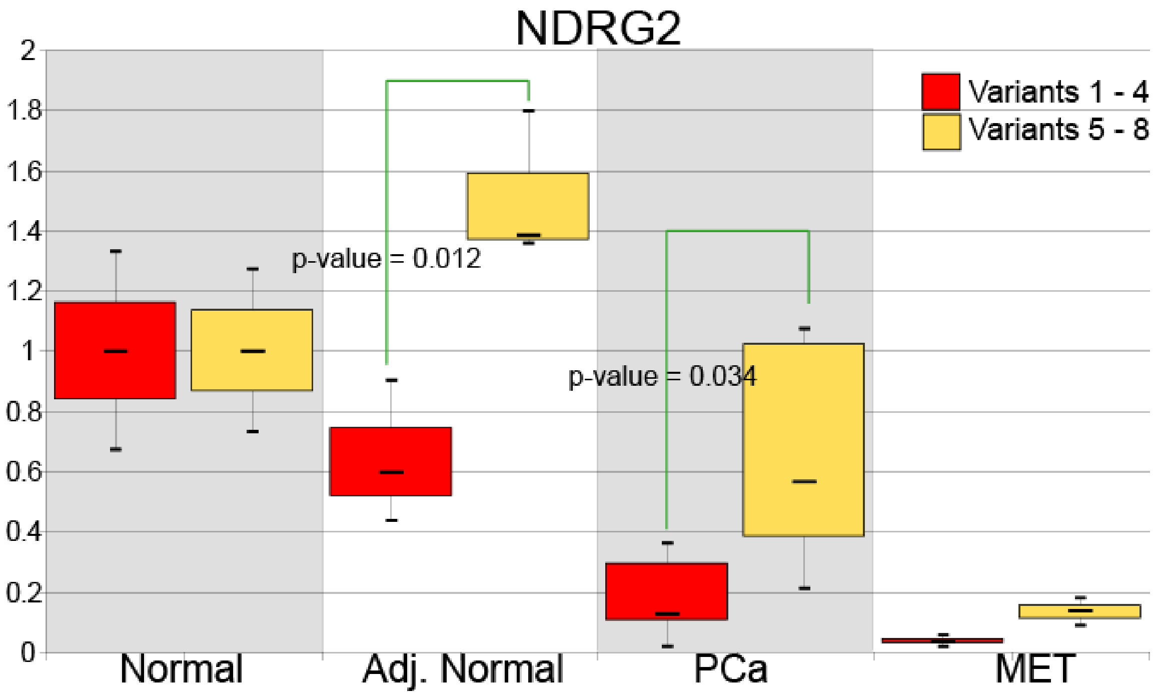


Figure 4.13 Isoform-specific expression patterns of NDRG2 in the prostate tissue cohort (n=12). qRT-PCR results on variants 1-4 (red) and 5-8 (yellow) of NDRG2 in 12 prostate tissue samples are shown as box plots. Variants 5-8, which share a common transcriptional start site, were significantly over-expressed compared to variants 1-4, which share a different common transcriptional start site, in adjacent normal (p-value=0.012) and localized prostate cancer (p-value=0.034) samples.

References

1. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6-21 .
2. Illingworth RS & Bird AP (2009) CpG islands--'a rough guide'. *FEBS Lett* 583(11):1713-1720 .
3. Nelson WG, De Marzo AM, & Yegnasubramanian S (2009) Epigenetic alterations in human prostate cancers. *Endocrinology* 150(9):3991-4002 .
4. Yegnasubramanian S, *et al.* (2004) Hypermethylation of CpG islands in primary and metastatic human prostate cancer. *Cancer Res* 64(6):1975-1986 .
5. Li LC, Carroll PR, & Dahiya R (2005) Epigenetic changes in prostate cancer: implication for diagnosis and treatment. *Journal of the National Cancer Institute* 97(2):103-115 .
6. Kron K, *et al.* (2009) Discovery of novel hypermethylated genes in prostate cancer using genomic CpG island microarrays. *PLoS ONE* 4(3):e4830 .
7. Yegnasubramanian S, *et al.* (2008) DNA hypomethylation arises later in prostate cancer progression than CpG island hypermethylation and contributes to metastatic tumor heterogeneity. *Cancer Res* 68(21):8954-8967 .
8. Ongenaert M, *et al.* (2008) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res* 36(Database issue):D842-846 .
9. Cokus SJ, *et al.* (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452(7184):215-219 .
10. Meissner A, *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454(7205):766-770 .
11. Down TA, *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology* 26(7):779-785 .
12. Weber M, *et al.* (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 37(8):853-862 .
13. Rauch TA & Pfeifer GP (2009) The MIRA method for DNA methylation analysis. *Methods in molecular biology (Clifton, N.J* 507:65-75 .
14. Brunner AL, *et al.* (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res* 19(6):1044-1056 .
15. Hodges E, *et al.* (2009) High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* .
16. Issa JP (2004) CpG island methylator phenotype in cancer. *Nat Rev Cancer* 4(12):988-993 .
17. Jones PA & Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature reviews* 3(6):415-428 .
18. Takai D & Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99(6):3740-3745 .

19. Irizarry RA*, et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 41(2):178-186 .

20. Rhodes DR*, et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9(2):166-180 .

21. Tomlins SA*, et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39(1):41-51.

22. Eckhardt F*, et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38(12):1378-1385 .

23. Lodygin D, Epanchintsev A, Menssen A, Diebold J, & Hermeking H (2005) Functional epigenomics identifies genes frequently silenced in prostate cancer. *Cancer Res* 65(10):4218-4227 .

24. Rehman I*, et al.* (2005) Promoter hyper-methylation of calcium binding proteins S100A6 and S100A2 in human prostate cancer. *Prostate* 65(4):322-330 .

25. Dammann R*, et al.* (2005) The tumor suppressor RASSF1A in human carcinogenesis: an update. *Histology and histopathology* 20(2):645-663 .

26. Dammann R*, et al.* (2000) Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat Genet* 25(3):315-319 .

27. Kuzmin I*, et al.* (2002) The RASSF1A tumor suppressor gene is inactivated in prostate tumors and suppresses growth of prostate carcinoma cells. *Cancer Res* 62(12):3498-3502 .

28. Thomson JP*, et al.* (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* 464(7291):1082-1086 .

29. Hu M, Yu J, Taylor JM, Chinnaiyan AM, & Qin ZS (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* 38(7):2154-2167 .

30. Rhodes DR*, et al.* (2007) Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia* 9(5):443-454 .

31. Mootha VK*, et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34(3):267-273 .

32. Subramanian A, Kuehn H, Gould J, Tamayo P, & Mesirov JP (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 23(23):3251-3253 .

33. Subramanian A*, et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550 .

34. Maher CA*, et al.* (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* 458(7234):97-101 .

35. Tusher VG, Tibshirani R, & Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9):5116-5121 .

36. Yan PS*, et al.* (2006) Mapping geographic zones of cancer risk with epigenetic biomarkers in normal breast tissue. *Clin Cancer Res* 12(22):6626-6636 .

37. Li LC & Dahiya R (2002) MethPrimer: designing primers for methylation PCRs. *Bioinformatics* 18(11):1427-1431 .

38. Bock C, *et al.* (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* 21(21):4067-4068 .

39. Han B, *et al.* (2008) A fluorescence in situ hybridization screen for E26 transformation-specific aberrations: identification of DDX5-ETV4 fusion protein in prostate cancer. *Cancer Res* 68(18):7629-7637 .

40. Tomlins SA, *et al.* (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* 448(7153):595-599 .

41. Vandesompele J, *et al.* (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3(7):RESEARCH0034 .

CHAPTER 5

INTEGRATIVE ANALYSIS OF PROSTATE CANCER TISSUE METHYLOME, COPY NUMBER AND GENE EXPRESSION CHANGES

DNA methylation is one of the mechanisms of gene repression, when it targets CpG rich regions usually found in small clusters termed CpG islands in gene promoters(1-2). In cancer, hypermethylation of gene promoters commonly marks disease progression, silencing putative tumor suppressor genes. More recently, using the methylation profiling, there were several attempts to classify cancer types based on methylation patterns, and to identify methylated gene predictors, which could be used as a diagnostic tool for determining samples with good vs. poor outcome. In addition to DNA methylation, the genomic aberrations such as deletion and amplification also play a significant role in transcriptional regulation. Defining tumor subtypes based on the presence of genomic aberrations and evaluating its prognostic value is a key research area in post-microarray era. In this chapter, we present results from our global methylation analysis obtained using M-NGS strategy (described in **Chapter 4**) of normal prostate and various clinical specimens that represent different stages of prostate cancer. We next layout the results from our integrative analysis of DNA methylation, copy number and gene expression dataset obtained from the same clinical samples. As described in the introduction part, DNA methylation events are known to occur throughout prostate cancer progression, and while genes such as GSTP1 and APC which show methylation changes early on, might play a role in tumor initiation, other methylation targets such as p14 and ER are thought to play a role in more advanced stages of prostate cancer (3). The sequence of copy number alteration also display a similar pattern, where certain genomic alterations such as chromosome 8p deletion are known to occur even in precursor lesion PIN stage, while clinically advanced stages of prostate cancer show several late onset genomic aberrations affecting vast regions of multiple chromosomes. Both DNA methylation and genomic alterations could drive transcriptional dysregulation. These two events are believed to act independently in gene regulation, and they could either act in the same direction (such as hypermethylation-deletion leading to gene repression, or hypomethylation-amplification leading to gene over-expression) or act as an opposing force (such as deletion and amplification). Knudsen’s two-hit hypothesis(4-5), that was formulated more than 20 years ago, was visited using the methylation event as one of the hits, and heterozygous deletion as the other which together could lead to gene silencing. Most of the tumor suppressors require two hits for complete loss of their activity, as seen in RB1, TP53, and APC in various tumor types (6-8). Although a nonsense or a frameshift mutation is a usual 1st hit(5), DNA methylation has been reported as the first hit event in several tumor suppressor genes such as VHL in renal and MLH1 in colon and gastric tumors(9-10). Unlike mutation or the methylation event, allelic loss involving deletion, mitotic recombination, and chromosomal disjunction has been rarely seen as the 1st hit and is mostly considered a 2nd hit. However, recent discovery in gene fusion events commonly occurred in prostate cancer indicates that the allelic loss may not be a rare 1st hit. In this chapter, we also identify potential target regions for a two-hit model by assessing for co-occurrence of DNA methylation and loss of heterozygosity due to 1 copy deletion in the prostate cancer genome. As a next level of analysis we integrated aCGH, gene expression profiling and M-NGS datasets obtained from the same specimens to investigate the relation between these genome wide changes in prostate cancer, and the results are discussed here.

Results and Discussion:

Characterization of DNA Methylation in Prostate Cancer Tissues

We generated eleven next generation sequencing libraries with M-NGS workflow from prostate tissue samples which include transplant normal, benign adjacent, localized, and metastatic tissues (**Table 5.1**). As described in **Chapter 4**, fifty nanograms of genomic DNA were used as input to create Methylplex libraries. The Methyplex libraries described above were constructed through the commercial service option provided by Rubicon Genomics Inc, Ann Arbor, MI as a part of early access to the technology. The very low input requirement of tissue DNA in this procedure is a definite advantage when one deals with low sample availability and holds a promise for characterization of FFPE tissue sections that might provide most valuable information. The sequencing libraries size selected for 350-450 bp size range were sequenced using single-read option on the Illumina Genome Analyzer II (see Methods for protocol details in Chapter 4). We obtained an average of 6 million mappable reads per M-NGS sample (see **Table 5.2**). CG dinucleotides in mapped M-NGS reads were enriched up to five-fold as compared to previously obtained pan-histone Chip-Seq data and this implied a good enrichment for methylation target regions by the Methylplex procedure (**Table 5.2**). A Hidden Markov Model (HMM)-based algorithm described in the Methods (from **Chapter 4**) section (11) was used to detect regions highly represented by the mapped reads obtained in each sequencing run. The coverage from tissue M-NGS runs ranged from 28 to 56 Mb which was comparable to cell line study presented in chapter 4 (**Table 5.2 and Table 4.1**). The benefit of using this application over raw reads was discussed in Chapter 4. Comparison of sequencing data obtained from two independent LNCAP cell line DNA M-NGS library preparations, processed in separate batches, showed a high experimental reproducibility with over 80% correlation between the runs (data not shown).

Global Differences in Prostate Tissue CpG Methylation

The genomic distribution of methylated regions in all the prostate tissue samples analyzed is shown in **Figure 5.1A**. Regardless of tissue types, most of the methylation occurred within intergenic and intronic regions. The normal specimens had the lowest, and metastatic samples the highest methylation level in all genomic locations among the various prostate tissue types analyzed (**Fig 5.1**). The promoter hypermethylation plays a role in gene repression (12) and tumorigenesis(13). Analysis of the CpG island / promoter methylation in our tissue samples was done essentially following the cell line dataset analysis described in **Chapter 4**. Of the 71,120 (56Mb) CpG islands identified by using Takai Jones criteria in the human genome (14) on average the cancer samples showed higher number of CGIs methylated 9,507 (ranged from 7,371 to 11,566) compared to normal 7,676 (ranged from 6,845 to 8,475) (**Table. 5.2**). A total distance of 3000bps (1,500bps on either side of transcription start site) was considered as promoter region for each gene in this analysis. On average, 2,816 genes from normal tissues harbored methylation within its promoter, and this number increases to 3,859 in tumor samples and this difference was statistically significantly (p-value < 0.0125) (**Table 5.2**). The methylation frequency of promoters with CpG islands was twice higher compared to promoters that lack CpG islands in all tissue types (**Fig. 5.1B**). Unlike the methylation pattern observed in genes with promoter CpG islands, the comparable number of methylation on gene promoter without CpG islands was observed from benign adjacent, localized, and metastatic samples. Overall, twice many numbers of genes were methylated in metastatic compared to transplant normal samples regardless of the presence and absence of CpG islands (**Fig. 5.1B**).

In order to get an unbiased visualization of promoters targeted by DNA methylation in our prostate cohort we queried our data set for presence of promoter methylation in any one sample. In total, 6,619 unique transcriptional start sites (corresponding to 6,077 unique Refseq genes) contained methylation in at least one sample and this information is displayed in a heatmap format for better visualization. (**Fig. 5.2**) Heatmap representation revealed a clear distinction in promoter methylation patterns present among the prostate tissue types analyzed (**Fig. 5.2**). Broadly, the promoters fell into 2 groups based on the presence or absence of a CpG island. Interestingly, 32.5% of promoters (n=2,154) lacked CpG islands but nevertheless exhibited methylation around the transcription start site (TSS) (**Fig. 5.2**). The remaining 67.5% (n=4,465) (I to IV, CpG islands indicated with red bars from 1st column) had CpG islands spanning the TSS and 3 distinct methylation patterns as shown in cell line model from Chapter 4 were again observed in the tissue samples: (1) Methylation was mostly confined to the island with higher frequency (2) Methylation was positioned 5’ to the CpG island and (3) Methylation was positioned 3’ to the CpG island. In addition to the correlation between methylation and location/ presence of CpG islands on gene promoter, we also observed a tissue type specific methylation pattern (regions I, II, III and IV) (**Fig. 5.2**). Importantly there exists a cancer-specific methylation pattern that encompasses region I (N=1,045; metastatic samples) and II (N=1,436; localized and metastatic samples) which may play a pivotal role in tumor progression. The methylation present in all sample types observed in region IV is thought to be prostate tissue specific (N=2,737) and only comparison with methylation dataset from other organs will provide a confirmation. Lastly, promoters that were methylated in prostates derived from prostate cancer patients (Normal adjacent, localized and metastatic prostate cancers) and not in normal prostates obtained individuals with no known prostate disease are clustered in region III (III, N=1,401). The well studied gene GSTP1 is present in this cluster and in general the methylation observed on these promoters could be due to factors such as age, cancer field effect etc. The normal cases in our tissue cohort came from a relatively younger population group (21 and 46 years of age), and hence age can be a factor in differential methylation observed mainly in region III. The cell line LNCaP and PrEC methylation profiles are displayed along with tissues in **Figure 5.2** for comparison and to show the overall resemblance between metastatic tissues and the metastatic prostate cell line LNCaP. Another interesting observation here was the higher similarity between PrEC and benign adjacent tissue methylomes compared to normal tissue, indicating possible cell culture induced quasi normal-states. In summary this analysis identified 2,481 unique gene promoters that constituted a cancer-specific methylation signature

Promoter Methylation and Transcriptional Repression in Prostate Cancer

After indentifying class specific promoter methylation patterns as described in the previous section, we next investigated the relationship between promoter DNA methylation and gene expression. The gene expression data (described in materials and methods section) was obtained by hybridizing the labeled probes prepared from the corresponding samples to Agilent human gene expression microarray platform. A two fold cutoff was used in the gene expression dataset to classify genes to be either under or overexpressed. First we looked at expression pattern of genes in each of the four regions identified in **Figure 5.2**. Interestingly genes in region II (methylated only in PCa and metastatic tissues) and region I (methylated only in metastatic tissues) showed a marked enrichment of gene repression (upto 43%) in cancer tissues compared to both type of normal tissues employed (p-value < 1.57e-11) **(Fig. 5.3A** and **Table 5.3)**. Genes with and without promoter CpG islands were indistinguishable in this analysis showing an equal effect of DNA methylation on gene repression even in promoters that lacked CGI. Genes from region IV (methylated in all samples) however, did not show a cancer specific enrichment in gene repression as expected **(Fig. 5.3A)**. The percent correlation between gene repression and DNA methylation was the highest (60-70%) in cancer samples when the analysis was restricted to matched samples with statistical significance **(Fig. 5.3B** and **Table 5.3).** The correlation rate between gene repression and promoter methylation with and without the presence of CpG islands were no differential (**Fig. 5.3B,** CpG vs. Non-CpG), suggesting that methylation-mediated gene repression does not require a CpG island-containing promoter, but rather all gene promoters may be functionally regulated by DNA methylation. Not all genes with promoter methylation are subjected to gene repression, however using the gene list from cancer-specific region I+II from **Figure 5.2**, we were able to distinguish normal samples from tumor samples (**Fig. 5.4**). In hope of finding the presence of certain transcription factors or genes involved in certain pathways that gives more vulnerability to the methylation-driven transcription, the Genomatix software is utilized. The presence of the differences in transcription binding factors and pathways involved between repressed and non-repressed genes harboring promoter methylation is accessed, and found no significant difference between two groups.

Analysis of aCGH data, mRNA expression and methylation profiles in matched prostate specimens

The eleven prostate tissue samples listed in **Table 5.1** were simultaneously characterized by M-NGS for global genome-wide methylation analysis, agilent gene expression microarray to monitor transcript expression and by Agilent aCGH microarrays to document copy number aberrations. The results from the integrative analysis between gene expression and DNA methylation were discussed in the previous section. We next investigated the relationship between gene expression and aCGH results to identify regions that exhibit coordinate copy number and transcript changes. Here we first listed the chromosomal regions that displayed significant aberrations in each sample using the Nexus copy number analysis software (BioDiscovery Inc., CA) using rank segmentation algorithm. Threshold used for high gain and loss was 0.7 and -0.75, and single copy gain and loss was 0.4 and -0.25, respectively. We then looked at the transcript expression patterns of genes located within these chromosomal regions in the corresponding sample's gene expression data. Gene expression pattern of the chromosomal regions with significant aberrations in the analyzed sample cohort is presented as a heat map in **Figure 5.5A**. One of our normal prostate sample derived from a chromosome 21 trisomy individual (sample obtained from Down's syndrome patient) displayed 1 copy gain in chromosome 21 as expected and served as an internal control (**Fig. 5.5A**). While the metastatic samples had multiple expansive regions of aberration (**Fig. 5.5A and 5.5B**). The heterozygous loss was the most common event in localized PCa tissues. In metastatic group, 2 out of 3 samples harbored comparable numbers of gain and loss, while homozygous loss was not frequently observed except in single copy Y chromosome (**Fig. 5.5B**). We also observed an association between mRNA under-expression and chromosomal loss (**Fig 5.5C**), where as high as 55% concordance (average value was 46%) in moderate level mRNA expression ($\log_2$ ratio<-1), was observed for tumor samples including MET and PCA tissues. The concordance rate between gain and over-expression was as high as 48% (average value was 35%). A previous prostate cancer aCGH and coupled gene expression study estimated a 38% (in average) association between highly amplified genes and transcript over-expression in MET samples (15). In breast cancer, Pollack et al. have reported a 62% (representing 54 unique genes) association between 117 highly amplified genes and transcript over-expression (16), and Hyman *et al*. have reported 44% highly amplified genes associated with over-expression, and 10.5% of highly over-expressed genes to be amplified in breast cancer (17). Integrative analysis of our aCGH and gene expression data allowed a direct comparison between the change in copy number and transcript expression levels, and genes within regions of significant genomic alterations showed concordance at the mRNA expression level. In respect with the methylation events, the enrichment of repression is shown among genes that are both methylated and deleted with highest concordance at 60% among tumor samples. In presence of both methylation and amplification, these regions are enriched with neither down-regulation nor up-regulation due to the presence of two opposite driving forces. Despite of having small number of prostate samples analyzed in this chapter, we were still able to demonstrate the major role of gene copy number variation in their transcriptional regulation in tumor samples.

Knudson's Two-hit Hypothesis

In cancer, loss of gene function by simultaneous promoter methylation and loss of heterozygosity in tumor suppressor genes such as RB1, VHL, CDKN2A, and CDKN2B, has been reported supporting Knudson's two-hit hypothesis (18). The first hit usually is a mutation (5), followed by inactivation of the second allele through genomic loss or DNA methylation. In case of VHL and MLH1 genes, DNA methylation was previously shown to serve as a first hit in various cancer (9-10, 18). Promoter methylation as a second hit is believed to be a rare event; however some incidents have been reported in von Hippel-Lindau syndrome and gastric cancer (5). Integrative analysis between methylation and copy number analysis opens up the possibility for insightful understanding of cooperating methylation-deletion events in Knudson's two-hit model (**Fig. 5.6A**). We examined the frequency of obtaining methylation-deletion events in our met samples (N=3). We divided all Ref Seq gene promoters (UCSC build 36.1, 2006) in each sample into two groups, 1) Promoters that contain DNA methylation and 2)

Unmethylated promoters. The groups were then queried in their respective sample aCGH data to know if they lie in a LOH region. From our analysis in metastatic samples, while a small percent (3%) of genes with promoter methylation was found to lie in regions of genomic amplification, as high as 28% (average of 23%) of genes with promoter methylation are harbored in regions of heterozygous loss (**Fig. 5.6B**). Hence the frequency of methylation-deletion event was significantly higher (p-value < 0.016) compared with methylation-amplification event (the highest with 6.9%, the average of 3.2%) (**Fig. 5.6B**). We also performed this analysis on localized PCa samples where the incident of methylation-deletion events in this tissue group were significantly lower at less than 5% (data not shown). This difference indicates possible occurrence of methylation-deletion two-hit event at later stage of cancer progression. Among genes lacking promoter methylation methylation, even though genomic deletion is observed at higher frequency than genomic amplification; this difference was not statistically significant (**Fig. 5.6C**). The frequency of deletion among the genes harboring methylation or lacking methylation showed no statistical difference. Methylation-deletion event is not the only scenario for two-hit activation. As mentioned previously, somatic mutations can make a significant contribution towards a first hit. Hence only an integrative analysis inclusive of the somatic mutation data in these samples will provide a complete picture. The frequency of the methylation-deletion event can be compared with no mutation/no methylation-deletion event for more complete analysis in future (**Fig. 5.6C**). Based on our analysis with metastatic samples, cooperation between promoter methylation and genomic deletions could play a major role in a two-hit inactivation model for prostate tumorigenesis. In order to represent candidate loci that exhibit frequent two-hit methylation-deletion, a “Two-hit frequency score” was calculated for all refseq genes and plotted in **Figure 5.6D**. The highest peak indicates most frequent site for two-hit event, and interestingly, NKX3.1, a known tumor suppressor gene in prostate cancer, resides within that locus. Few other loci with high peaks harbored well-known tumor suppressor genes such as WT1, RB1, and SMAD4 and the sites of LOH in prostate cancer such as chromosome 16 and 19 (**Fig. 5.6D**). The list of genes that contain both methylation and deletion in more than two metastatic samples is listed in **Table 5.4**. When these genes were mapped to the heatmap represented in **Figure 5.2**, all were located within cancer-specific methylated region I and II, implicating the role of two-hit inactivation in tumorigenesis. Characterization of a larger sample set will definitely strengthen this analysis in identifying causal genetic loci. Again, characterizing multiple samples from the same patient might provide the information on whether the methylation comes in as a fist hit or a second hit player in this model, which remains to be demonstrated by global analysis.

In summary, we used a high throughput M-NGS strategy to characterize the DNA methylome map of prostate tissue specimen using a minimal amount of input DNA. In addition to the identification of distinct patterns of DNA methylation around TSSs that frequently occur on promoters either containing or lacking a CpG island, the cancer-specific methylation observed only among localized and metastatic samples and prostate-specific observed among all sample types were discovered. Upon integrative analyses with copy number, gene expression and methylation data from identical samples we found the evidence for methylation-deletion two-hit events, identified the frequent two-hit loci, and the candidate genes with two-hit inactivation in metastatic samples.

Materials and Methods:

Reagents and prostate tissue samples

Grossly dissected human prostate tissue samples from University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core (SPORE) were collected with informed consent of the patients and prior institutional review board approval. A total of 2 normal, 2 benign adjacent, 3 localized cancer and 3 metastatic prostate samples (n=10) were characterized by M-NGS, aCGH and GE microarray (**Table 5.1**). Genomic DNA was isolated from tissue using DNeasy Blood and tissue kit (Qiagen Inc, Valencia, CA) according to manufacturer’s instructions. For M-NGS library generation and hybridization, please see materials and methods in **Chapter 4**.

Agilent GE and aCGH Hybridization

Total RNA isolated from tissue samples was further purified using RNAeasy Micro Kit (Qiagen) according to the manufacturer's instructions. Expression profiling was performed using the Agilent 44K expression array. Briefly, one microgram of total RNA was converted to cRNA and then labeled according to the manufacturer's protocol (Agilent). The aCGH profiling was performed on the Agilent Human Genome CGH 244K array. Genomic DNA was isolated from tissue samples and further purified using DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions. Two microgram of gDNA was labeled according to the manufacturer’s protocol (Agilent).

Hybridizations were performed for 16 h at 65 °C. Scanned images from Agilent microarray scanner were analyzed and extracted using Agilent Feature Extraction Software 9.1.3.1, with linear and lowess normalization performed for each array. The prostate tissue samples (Cy5) were hybridized against control pooled prostate RNA normal samples (BD Clontech, Heidelberg, Germany) (Cy3).

Statistical Analysis

HMM analysis of M-NGS data. Hidden Markov Model (HMM) based next generation sequencing analysis is conducted in a two-step process that takes in raw reads and outputs refined boundaries of enriched chromosomal regions. In detail, please refer to materials and methods in **Chapter 4**.

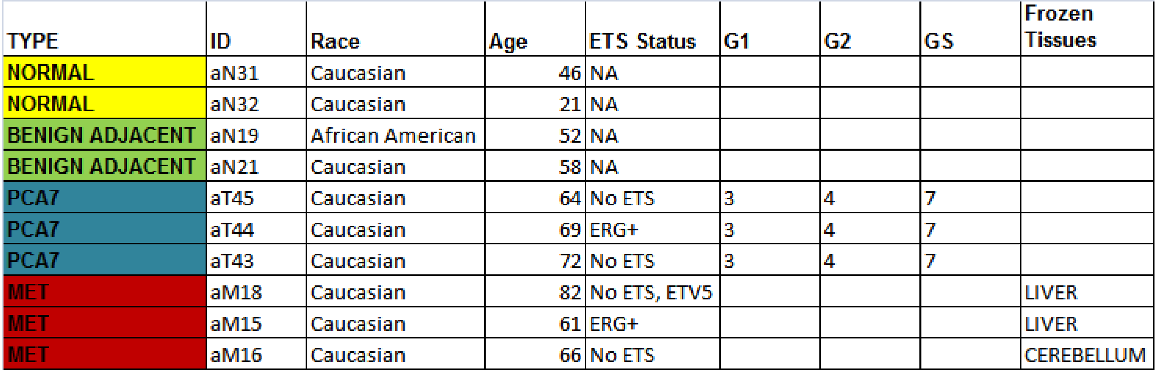
Nexus Copy Number Analysis. The detailed protocol on Nexus Copy Number Analysis is available **(19)**. The output from Agilent Feature Extraction Software were imported to Nexus copy number analysis tool. The transplant normal sample with known chromosome 21 trisomy served as an additional internal control for one copy gain and it was included as one of our insetting threshold for the copy number analysis.

Genomatix analysis. The GenomatixSuite is a web-based software available at www.genomatix.de. RegionMiner analysis was used to identify the global over-representation of transcription factors binding sites. Using the list of genes as input, BiblioSphere Pathway allowed to rank and explore the multiple pathways.

Two-hit Frequency Analysis

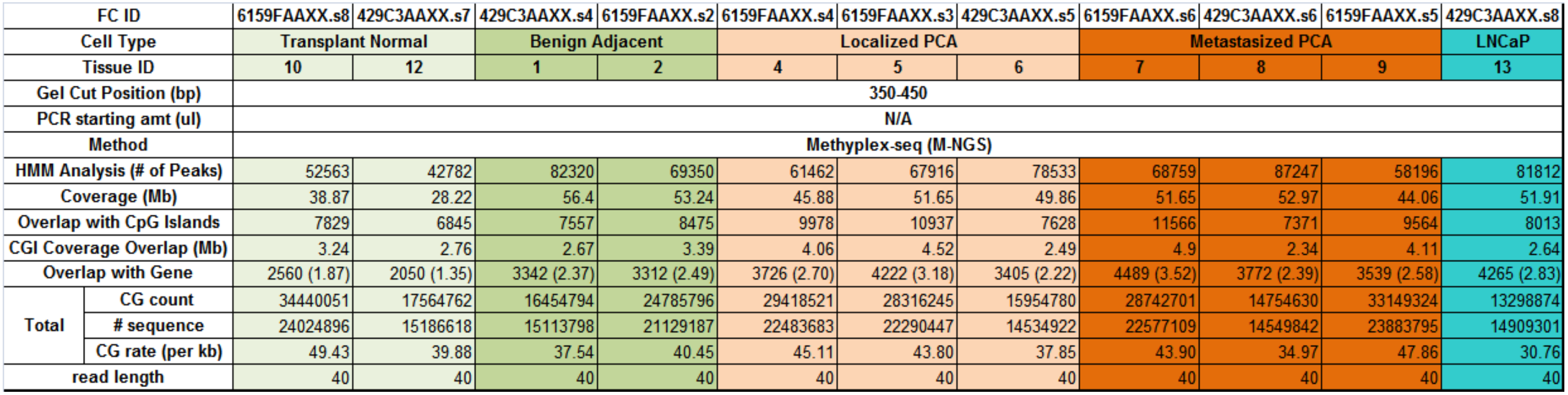
The RefSeq genes (UCSC genome build 36.1, March 2006) are arranged according to their chromosomal location. In each sample, a score was given to each gene based on the presence of heterozygous loss and promoter methylation. For each gene, the scores from metastatic samples were combined, then the moving average value of consecutive 20 genes are plotted to identify the frequent two-hit locus.

Table 5.1 Clinical information of prostate tissue samples profiled by M-NGS analysis



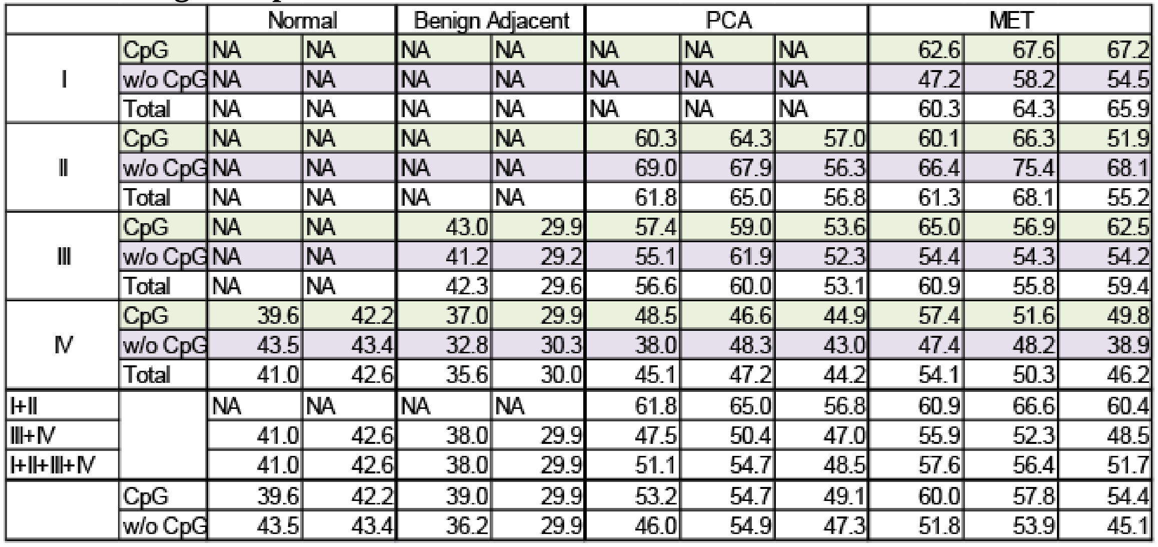
TYPE	ID	Race	Age	ETS Status	G1	G2	GS	Frozen Tissues
NORMAL	aN31	Caucasian	46	NA				
NORMAL	aN32	Caucasian	21	NA				
BENIGN ADJACENT	aN19	African American	52	NA				
BENIGN ADJACENT	aN21	Caucasian	58	NA				
PCA7	aT45	Caucasian	64	No ETS	3	4	7	
PCA7	aT44	Caucasian	69	ERG+	3	4	7	
PCA7	aT43	Caucasian	72	No ETS	3	4	7	
MET	aM18	Caucasian	82	No ETS, ETV5				LIVER
MET	aM15	Caucasian	61	ERG+				LIVER
MET	aM16	Caucasian	66	No ETS				CEREBELLUM

Table 5.2 Summary of Prostate Tissues Methylplex Sequencing Results



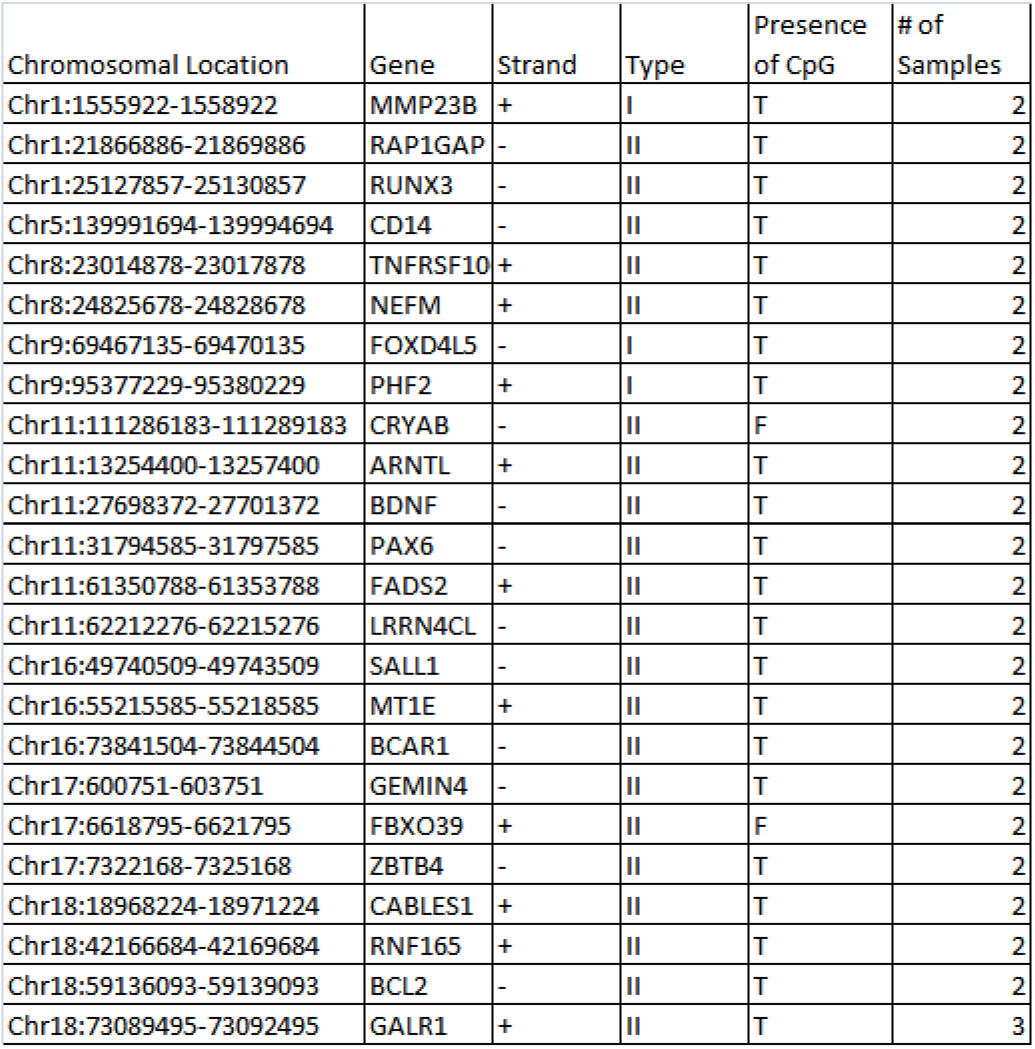
FC ID		6159FAAXX.s8	429C3AAXX.s7	429C3AAXX.s4	6159FAAXX.s2	6159FAAXX.s4	6159FAAXX.s3	429C3AAXX.s5	6159FAAXX.s6	429C3AAXX.s6	6159FAAXX.s5	429C3AAXX.s8
Cell Type		Transplant Normal		Benign Adjacent		Localized PCA			Metastasized PCA			LNCaP
Tissue ID		10	12	1	2	4	5	6	7	8	9	13
Gel Cut Position (bp)		350-450										
PCR starting amt (ul)		N/A										
Method		Methyplex-seq (M-NGS)										
HMM Analysis (# of Peaks)		52563	42782	82320	69350	61462	67916	78533	68759	87247	58196	81812
Coverage (Mb)		38.87	28.22	56.4	53.24	45.88	51.65	49.86	51.65	52.97	44.06	51.91
Overlap with CpG Islands		7829	6845	7557	8475	9978	10937	7628	11566	7371	9564	8013
CGI Coverage Overlap (Mb)		3.24	2.76	2.67	3.39	4.06	4.52	2.49	4.9	2.34	4.11	2.64
Overlap with Gene		2560 (1.87)	2050 (1.35)	3342 (2.37)	3312 (2.49)	3726 (2.70)	4222 (3.18)	3405 (2.22)	4489 (3.52)	3772 (2.39)	3539 (2.58)	4265 (2.83)
Total	CG count	34440051	17564762	16454794	24785796	29418521	28316245	15954780	28742701	14754630	33149324	13298874
	# sequence	24024896	15186618	15113798	21129187	22483683	22290447	14534922	22577109	14549842	23883795	14909301
	CG rate (per kb)	49.43	39.88	37.54	40.45	45.11	43.80	37.85	43.90	34.97	47.86	30.76
read length		40	40	40	40	40	40	40	40	40	40	40

Table 5.3 The concordance between differentially methylated regions identified by M-NGS and gene repression



		Normal		Benign Adjacent		PCA			MET		
I	CpG	NA	NA	NA	NA	NA	NA	NA	62.6	67.6	67.2
	w/o CpG	NA	NA	NA	NA	NA	NA	NA	47.2	58.2	54.5
	Total	NA	NA	NA	NA	NA	NA	NA	60.3	64.3	65.9
II	CpG	NA	NA	NA	NA	60.3	64.3	57.0	60.1	66.3	51.9
	w/o CpG	NA	NA	NA	NA	69.0	67.9	56.3	66.4	75.4	68.1
	Total	NA	NA	NA	NA	61.8	65.0	56.8	61.3	68.1	55.2
III	CpG	NA	NA	43.0	29.9	57.4	59.0	53.6	65.0	56.9	62.5
	w/o CpG	NA	NA	41.2	29.2	55.1	61.9	52.3	54.4	54.3	54.2
	Total	NA	NA	42.3	29.6	56.6	60.0	53.1	60.9	55.8	59.4
IV	CpG	39.6	42.2	37.0	29.9	48.5	46.6	44.9	57.4	51.6	49.8
	w/o CpG	43.5	43.4	32.8	30.3	38.0	48.3	43.0	47.4	48.2	38.9
	Total	41.0	42.6	35.6	30.0	45.1	47.2	44.2	54.1	50.3	46.2
I+II		NA	NA	NA	NA	61.8	65.0	56.8	60.9	66.6	60.4
III+IV		41.0	42.6	38.0	29.9	47.5	50.4	47.0	55.9	52.3	48.5
I+II+III+IV		41.0	42.6	38.0	29.9	51.1	54.7	48.5	57.6	56.4	51.7
	CpG	39.6	42.2	39.0	29.9	53.2	54.7	49.1	60.0	57.8	54.4
	w/o CpG	43.5	43.4	36.2	29.9	46.0	54.9	47.3	51.8	53.9	45.1

Table 5.4 Top candidate gene list for two-hit inactivation



Chromosomal Location	Gene	Strand	Type	Presence of CpG	# of Samples
Chr1:1555922-1558922	MMP23B	+	I	T	2
Chr1:21866886-21869886	RAP1GAP	-	II	T	2
Chr1:25127857-25130857	RUNX3	-	II	T	2
Chr5:139991694-139994694	CD14	-	II	T	2
Chr8:23014878-23017878	TNFRSF10	+	II	T	2
Chr8:24825678-24828678	NEFM	+	II	T	2
Chr9:69467135-69470135	FOXD4L5	-	I	T	2
Chr9:95377229-95380229	PHF2	+	I	T	2
Chr11:111286183-111289183	CRYAB	-	II	F	2
Chr11:13254400-13257400	ARNTL	+	II	T	2
Chr11:27698372-27701372	BDNF	-	II	T	2
Chr11:31794585-31797585	PAX6	-	II	T	2
Chr11:61350788-61353788	FADS2	+	II	T	2
Chr11:62212276-62215276	LRRN4CL	-	II	T	2
Chr16:49740509-49743509	SALL1	-	II	T	2
Chr16:55215585-55218585	MT1E	+	II	T	2
Chr16:73841504-73844504	BCAR1	-	II	T	2
Chr17:600751-603751	GEMIN4	-	II	T	2
Chr17:6618795-6621795	FBXO39	+	II	F	2
Chr17:7322168-7325168	ZBTB4	-	II	T	2
Chr18:18968224-18971224	CABLES1	+	II	T	2
Chr18:42166684-42169684	RNF165	+	II	T	2
Chr18:59136093-59139093	BCL2	-	II	T	2
Chr18:73089495-73092495	GALR1	+	II	T	3

Type: indicates the regions from heatmap represented in **Figure 5.2**
T- CpG Islands present
F-CpG Islands Absent

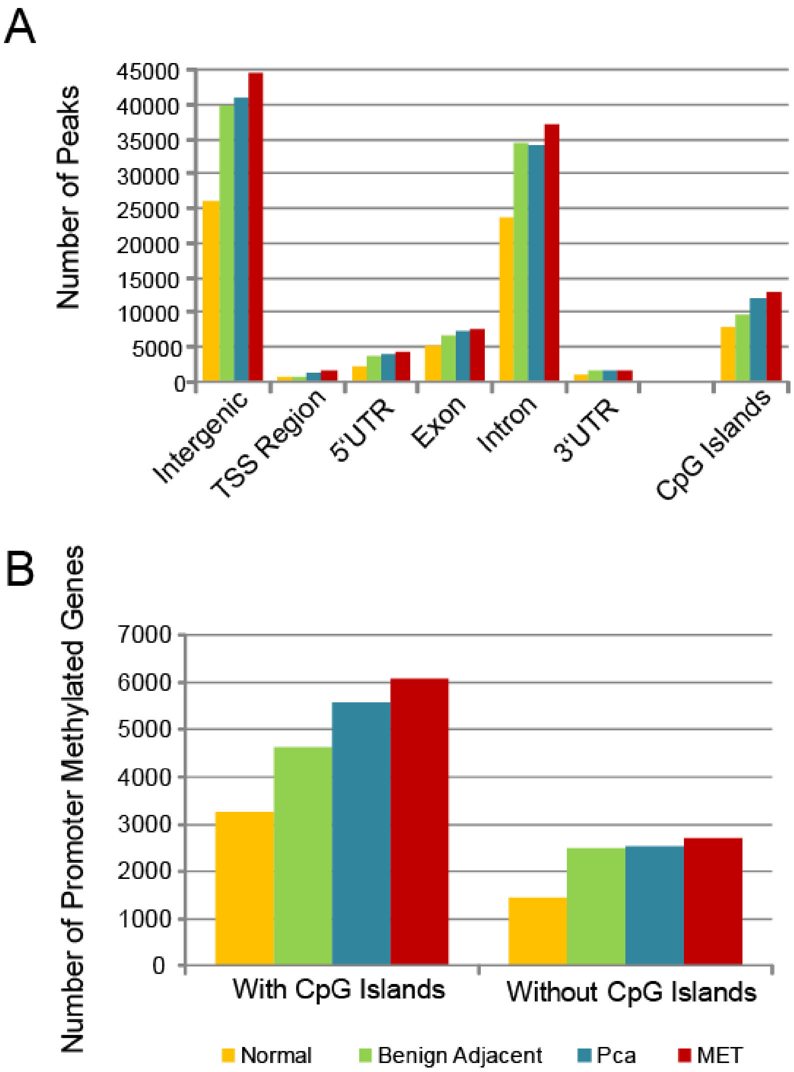


Figure 5.1 Genomic regions targeted by DNA methylation in prostate tissue samples as revealed by Methylplex-Next generation sequencing (M-NGS). (A) DNA methylation in intergenic and intronic regions represents a large proportion of all observed DNA methylation in tissue samples. While the normal samples exhibited the lowest number of methylation peaks, total number of methylated peaks in benign adjacent, localized and metastatic samples was comparable. (B) Methylated promoters in all tissue types are two fold more likely to have a CpG island. While the number of promoter methylation with CpG islands is increased gradually with the cancer progression from normal to metastatic samples, this increase is not evident among promoters without CpG islands.

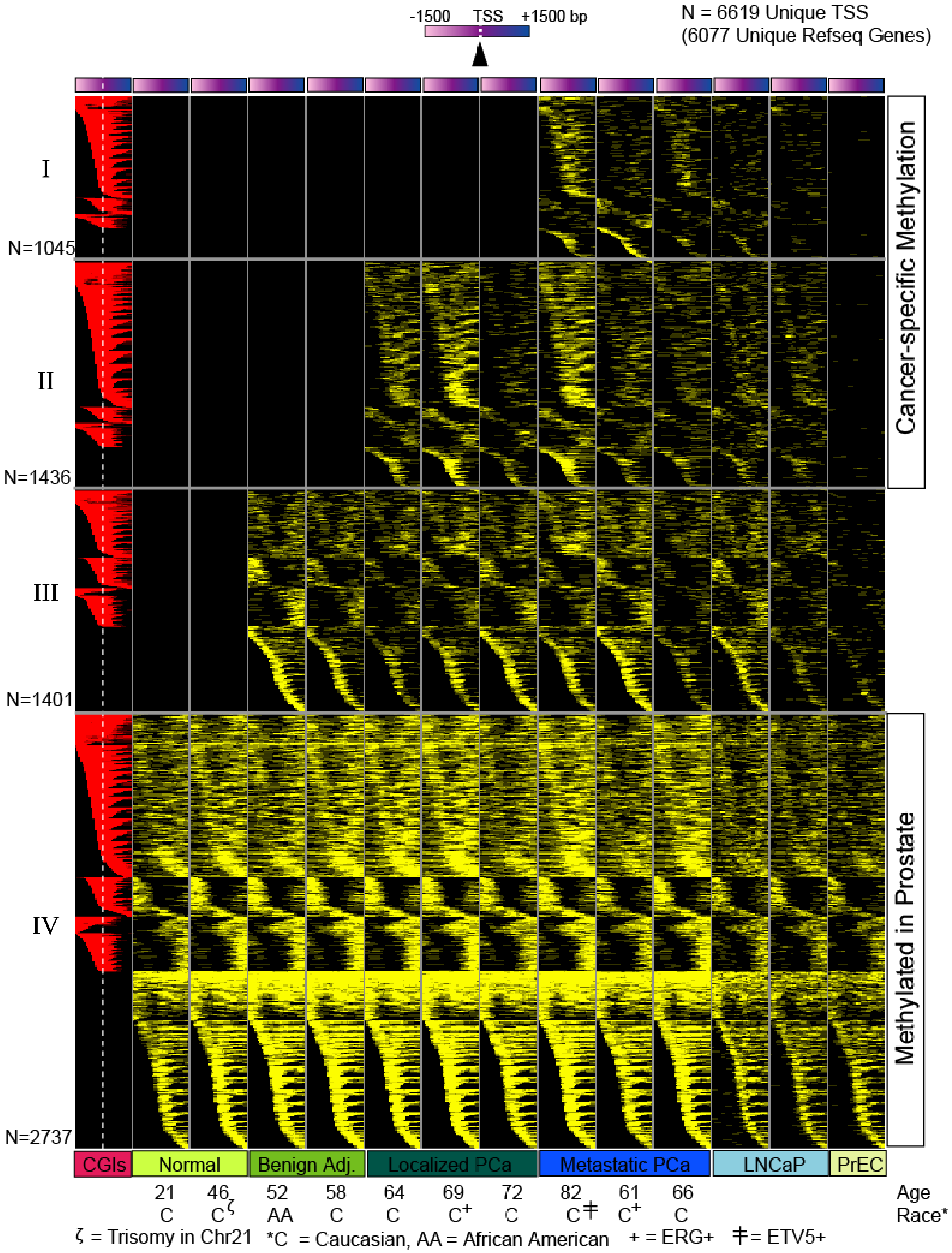


Figure 5.2 Prostate tissue samples with distinct patterns of promoter methylation. Analysis of gene promoter regions (±1,500 bp from TSS) identified methylation in 6,619 RefSeq transcriptional start sties (corresponding to 6,077 unique genes) present in any one tissue sample. Each row represents a unique promoter region, ±1,500 bp from the

transcription start site (white dotted line on 1st column) at 100 bp window size. CpG island location is indicated in red in the first column. The methylation (yellow) observed in the corresponding location in each tissue and cell line is indicated. Among promoters represented, 3 distinct patterns of methylation were found. Methylation occurred either (1) on CpG islands, (2) in regions flanking the island (5' or 3') or (3) in promoters without any CpG island. In addition, the tissue-type specific methylation was also present (cluster I, II, III, and IV). Cluster I and II is defined as cancer-specific, and cluster IV is defined as the methylated genes in prostate tissues. The information on age and race as well as their ETS fusion status of each sample are provided.

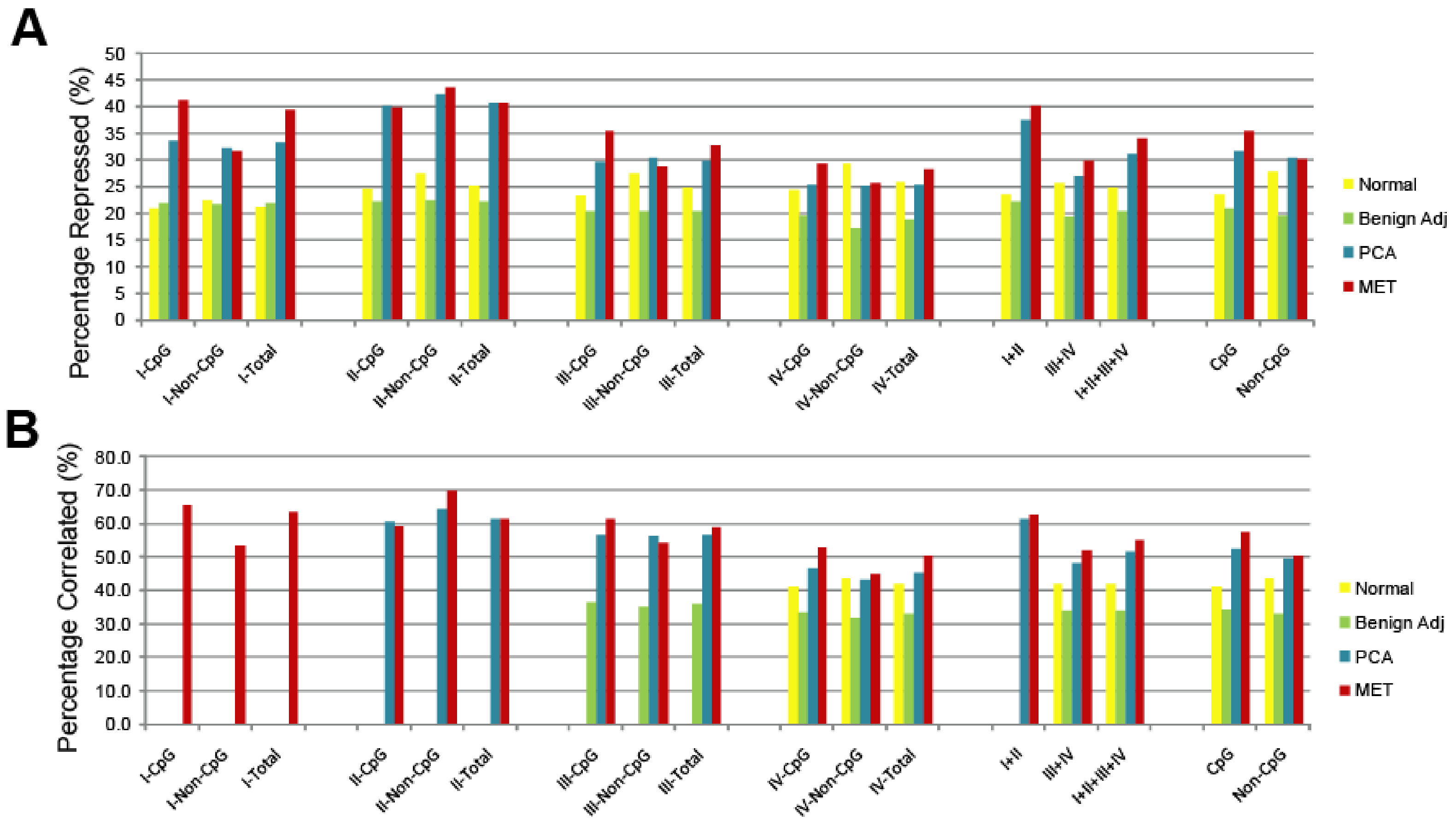


Figure 5.3 DNA methylation is associated with gene repression. (A) Genes located within cancer-specific clusters I and II are enriched with repressed genes among cancer samples compared to the ones located within clusters III and IV (p-value < 1.57e-11). There was no difference in repression rate between genes with and without CpG islands. The benign adjacent samples showed the lowest number of repressed genes among all tissue types in all cluster types. (B) Integrative analysis revealed the overlap between differentially methylated regions identified by M-NGS and gene repression. The percent correlation between gene repression and DNA methylation was the highest (60-70%) in cancer samples when the analysis was restricted to matched sample. The values used in **Figure 5.3** can be found in **Table 5.3**.

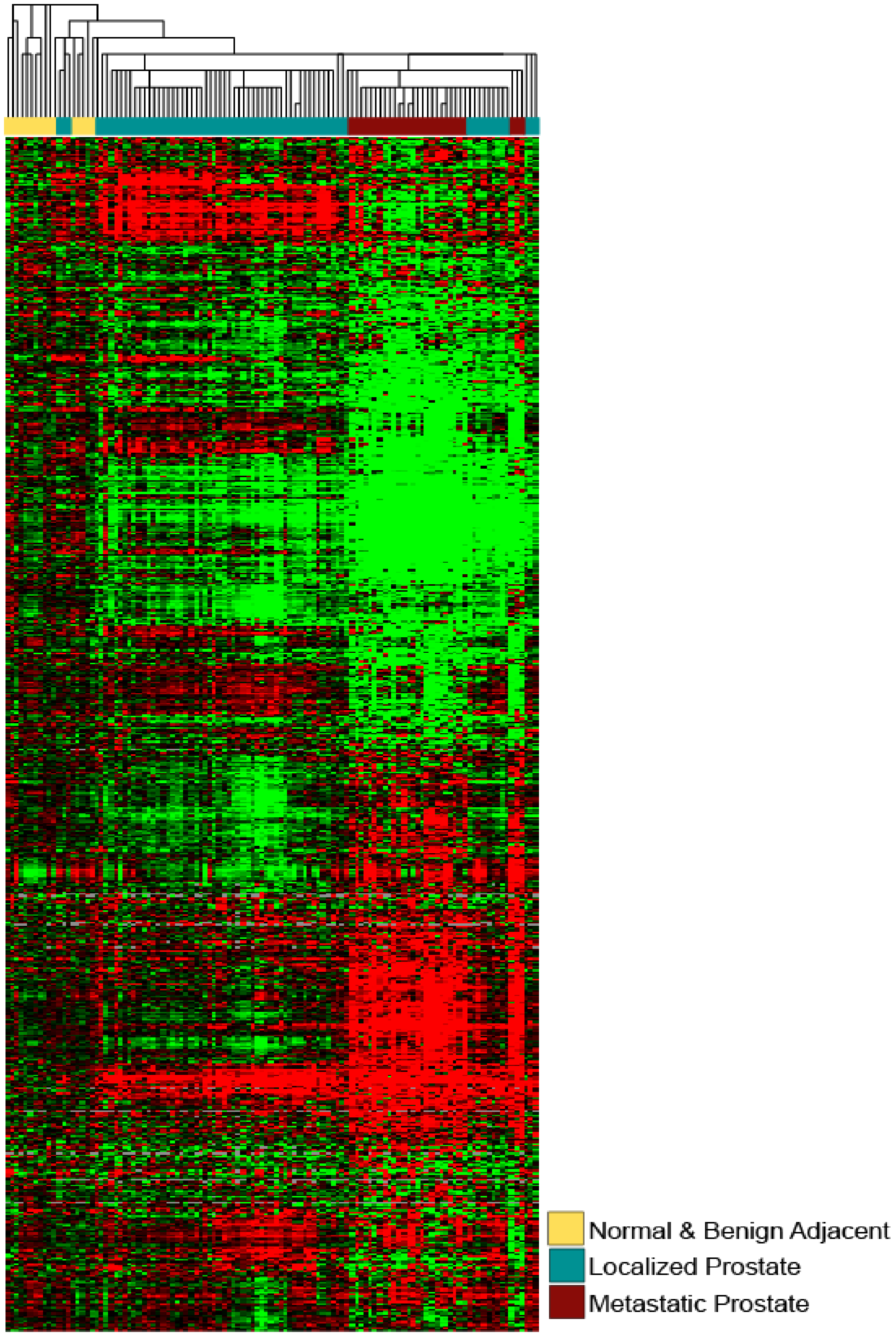


Figure 5.4 Hierarchical clustering of prostate cancer transcriptome (microarray data) by methylation target genes. Hierarchical clustering of gene expression values (prostate cancer microarray dataset n=155) of methylated targets in prostate cancer (clusters I and II from **Fig. 5.2**) was able to differentiate between normal and cancer samples.

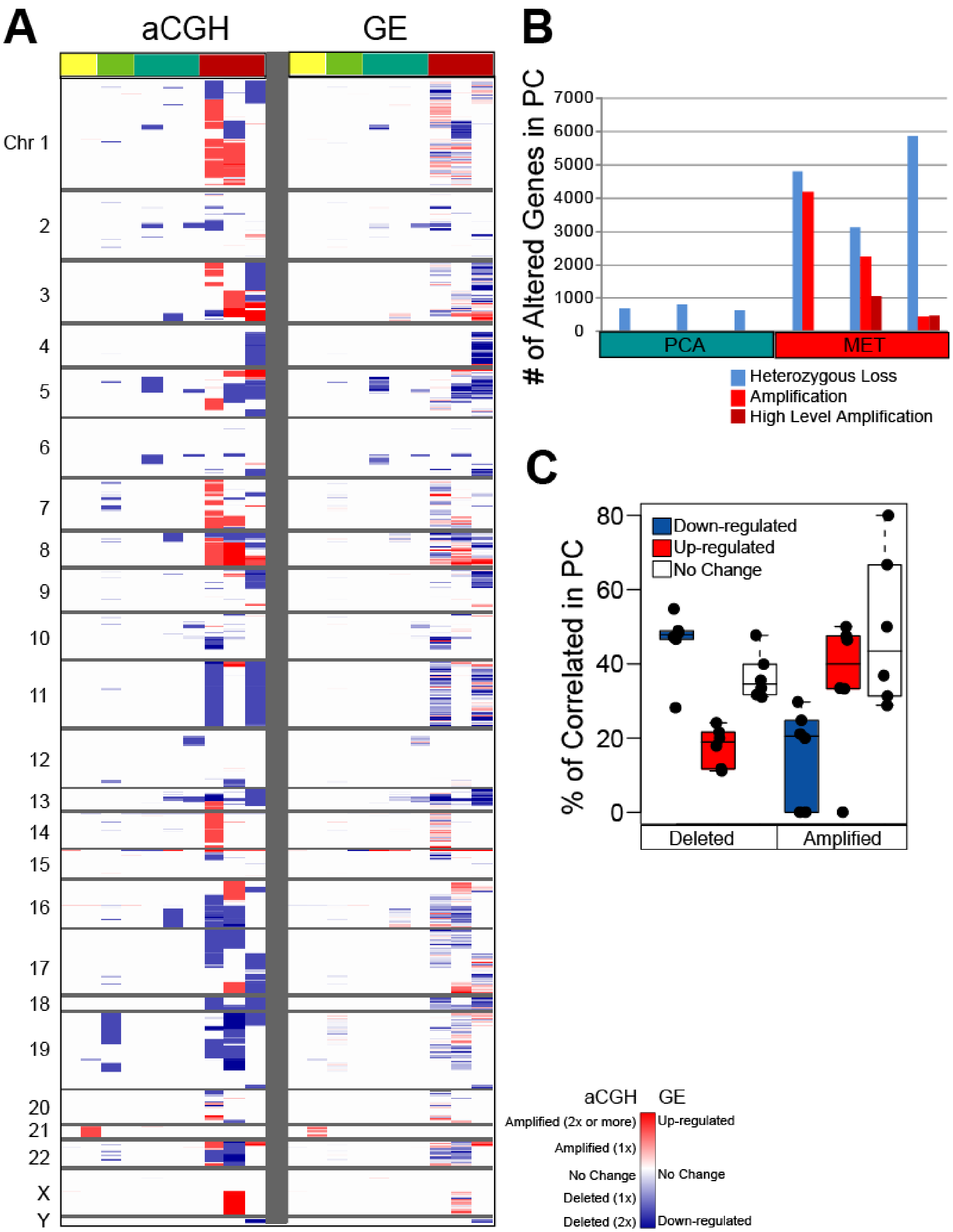


Figure 5.5 Genome-wide chromosomal alterations and their matched gene expression changes in prostate cancer. (A) Chromosomal aberrations are depicted for tissue samples, normal(yellow), followed by benign (green) and tumor prostate samples (cyan and dark red). Each row represents one of 20,990 unique genes, ordered by genome map position from chromosome 1 to Y (red reflects fold-amplification, blue reflects fold-deletion, and white indicates no change on left panel). Right panel displays the mRNA

expression of matched samples within regions of significant genomic alteration. (red reflects over-expression, blue reflects under-expression, and white indicates no change on right panel). Trisomy in chromosome 21 is observed in one of transplant normal samples as single copy gain, and corresponding changes in gene expression is also shown. (B) The number of genomic aberrations including heterozygous loss, single copy gain, and high copy gain in each tumor sample is shown. The heterozygous loss was the most frequently observed event in localized samples, while metastatic samples harbor comparable numbers of deletion and amplification in two out of three samples. (C) The association between deletion and gene repression and amplification and gene up-regulation in tumor samples is displayed. Around 50% of deleted genes were repressed and 40% of amplified genes were over-expressed in prostate tissues.

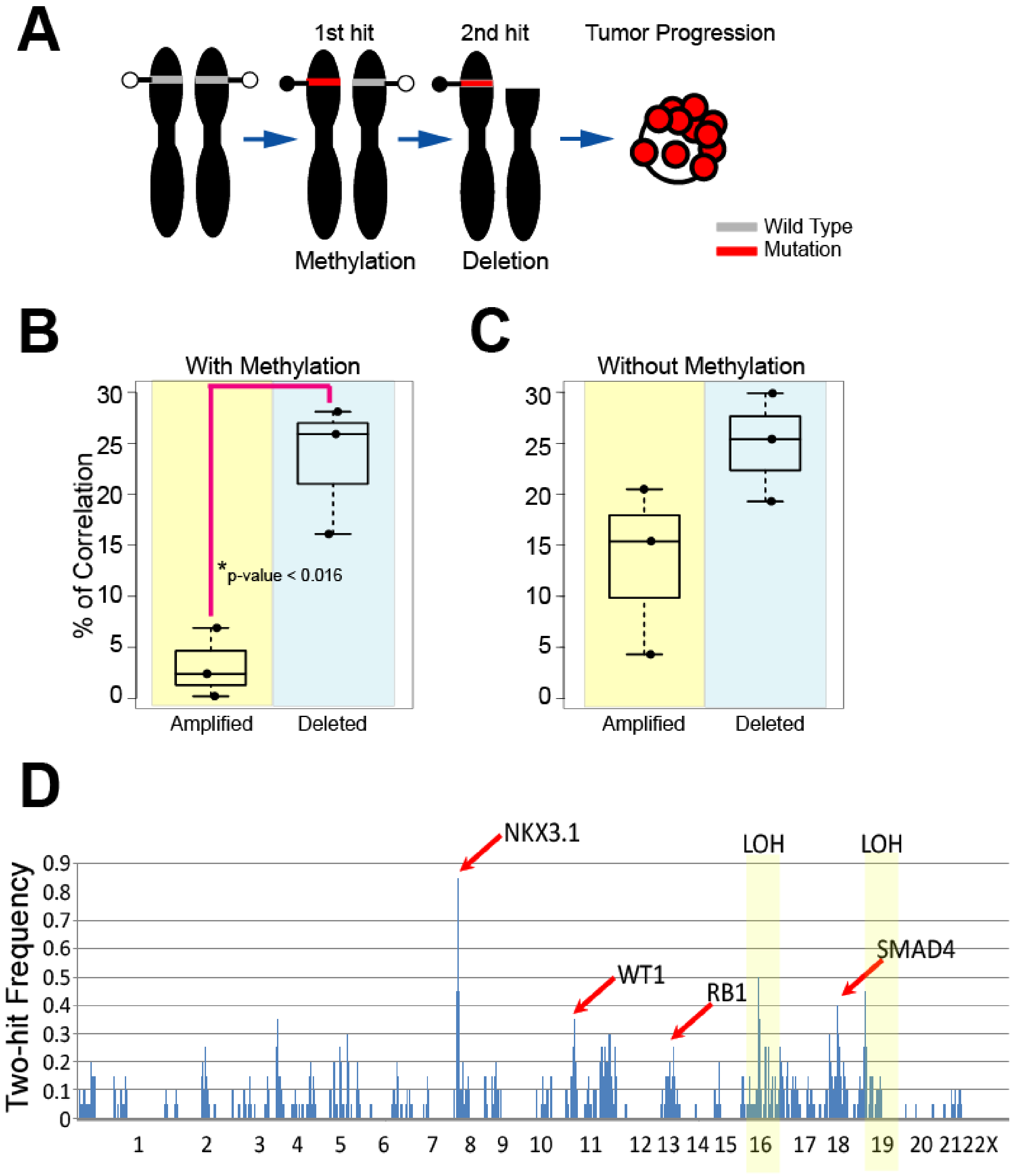


Figure 5.6 Association between copy number alteration and DNA methylation. (A) A schematic of two-hit hypothesis involving methylation and deletion is provided. (B) Among genes with promoter methylation, significantly more number of regions was deleted than amplified (p-value < 0.016). (C) Among the genes without promoter methylation, deletion was more frequently observed than amplification, this difference was not statistically significant. (D) Two-hit frequency scores for each RefSeq gene was calculated and plotted. Tumor suppressor genes such as NKX3.1, WT1, RB1, and SMAD4 and frequently reported regions of LOH such as chromosome 16 and 19 was among the regions with highest scores.

References

1. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6-21 .
2. Illingworth RS & Bird AP (2009) CpG islands--'a rough guide'. *FEBS Lett* 583(11):1713-1720 .
3. Yegnasubramanian S*, et al.* (2004) Hypermethylation of CpG islands in primary and metastatic human prostate cancer. *Cancer Res* 64(6):1975-1986 .
4. Knudson AG (1996) Hereditary cancer: two hits revisited. *J Cancer Res Clin Oncol* 122(3):135-140 .
5. Tomlinson IP, Roylance R, & Houlston RS (2001) Two hits revisited again. *J Med Genet* 38(2):81-85 .
6. Watanabe T, Nakamura M, Yonekawa Y, Kleihues P, & Ohgaki H (2001) Promoter hypermethylation and homozygous deletion of the p14ARF and p16INK4a genes in oligodendrogliomas. *Acta Neuropathol* 101(3):185-189 .
7. Esteller M*, et al.* (2000) Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. *Cancer Res* 60(16):4366-4371 .
8. Pogribny IP & James SJ (2002) Reduction of p53 gene expression in human primary hepatocellular carcinoma is associated with promoter region methylation without coding region mutation. *Cancer Lett* 176(2):169-174 .
9. Prowse AH*, et al.* (1997) Somatic inactivation of the VHL gene in Von Hippel-Lindau disease tumors. *Am J Hum Genet* 60(4):765-771 .
10. Myohanen SK, Baylin SB, & Herman JG (1998) Hypermethylation can selectively silence individual p16ink4A alleles in neoplasia. *Cancer Res* 58(4):591-593 .
11. Hu M, Yu J, Taylor JM, Chinnaiyan AM, & Qin ZS (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res* 38(7):2154-2167 .
12. Jones PA & Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature reviews* 3(6):415-428 .
13. Issa JP (2004) CpG island methylator phenotype in cancer. *Nat Rev Cancer* 4(12):988-993 .
14. Takai D & Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99(6):3740-3745 .
15. Kim JH*, et al.* (2007) Integrative analysis of genomic aberrations associated with prostate cancer progression. *Cancer Res* 67(17):8229-8239 .
16. Pollack JR*, et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A* 99(20):12963-12968.
17. Hyman E*, et al.* (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res* 62(21):6240-6245.
18. Jones PA & Laird PW (1999) Cancer epigenetics comes of age. *Nat Genet* 21(2):163-167 .
19. Darvishi K (2010) Application of Nexus copy number software for CNV detection and analysis. *Curr Protoc Hum Genet* Chapter 4:Unit 4 14 11-28 .

CHAPTER 6

CONCLUSION

Prostate cancer is the most common epithelial cancer and second leading cause of cancer death in men in the US. The list of molecular events that take place during cancer initiation and progression includes, a) transcriptomic changes, b) epigenetic changes such as DNA methylation and histone modification and c) copy number alterations d) somatic mutations. The goal of my thesis work was to gain understanding of prostate cancer biology by investigating the relationship between some of these major genome-wide events through integrative analysis.

In **Chapter 2**, we monitored genome-wide copy number changes using array comparative genomic hybridization (aCGH) of laser-capture microdissected prostate cancer samples (N=62) on a cDNA microarray platform. The samples represented multiple stages of prostate cancer progression, including precursor lesions (PIN and PIA), clinically localized disease, and metastatic cancer in addition to normal epithelium and stromal compartments. Minimal common regions (MCR) of copy number alterations were defined for each sample type, and metastatic samples displayed the most number of alterations. We identified several novel MCRs in addition to detecting known regions of copy number aberration. The genomic aberrations identified in each stage of tumor comprised of both unique and recurrent events. There were genomic aberrations unique to localized and metastatic cancers, but not in normal samples. Also there were shared aberration sites throughout progression starting with PIN stage. Identifying these shared regions is of great interest as the candidate genes that lie within might play a role in tumor progression, especially when the alteration is detected in early precursor stage and is preserved or becomes more frequent in other progressive stages of the cancer. The role of PIN as neoplastic precursor lesion for prostate cancer is supported by our study. First, the whole arm amplification, often observed among advanced form of the disease, was also present in at least one PIN sample (8q amplification). Secondly, several genetic alterations indentified in our PIN samples resembled the events observed in PCA including the amplified regions such as 8q22.2-q24.12 or deleted regions such as 18q21.1-q23.

From **Chapter 2**, we identified the regions with genomic aberrations in the entire spectrum of prostate cancer from normal to metastatic prostate cancer. We hypothesized that this genomic aberration observed within each sample group is one of the key underlying mechanisms for transcriptional regulation and responsible for transcriptional differences among sample groups. This hypothesis was tested in **Chapter 3**, which describes the integrative analysis performed on genomic aberration and gene expression dataset obtained simultaneously from identical samples. In our analysis, we identified 42% of amplified genes to be over-expressed in metastatic samples. This percentage was lower in localized prostate samples, only 22% of amplified genes were over-expressed. Top-altered genes with corresponding gene expression change were identified from our aCGH, and this includes well-studied MYC, TPD52, and PTEN among several novel candidates. Recently, using the same prostate cohort on gene expression array, our lab has identified several outlier genes that are overexpressed in a subset of cancer tissues. Further characterization of one such outlier namely ERG (an ETS family transcription factor), lead to the discovery of the first gene fusion phenomenon in prostate cancer involving the genes TMPRSS2 and ERG. ETS fusion are found in upto 60% of the prostate cancers and may play crucial role in prostate cancer development. We speculated that prostate cancer samples based on their ETS fusion status (fusion positive or negative) might use alternative pathways / mechanisms to promote tumorigenesis and differential genomic aberrations, and the genomic difference may be present between the two categories. Analysis of copy number differences between ETS fusion positive and negative samples identified alterations in various chromosomal regions including 6q21 present only among fusion negative tumor samples. The genomic region 6q21 whose deletion is present in over 45% of our non-ETS cases contains genes such as FOXO3A and CCNC that are reported to play a role in prostate carcinogenesis. This significant finding will help in further characterizing the ETS fusion negative prostate cancers samples.

In addition to genomic aberrations, there are epigenetic events that occur during cancer progression. Histone and DNA modifications are key mechanisms for transcriptional regulation. Unlike histone modification marks, which are associated with both gene activation and repression, DNA methylation is generally considered to be involved with gene repression. However, according to the latest literature, an epigenetic mark can either repress or support transcription based on the location of the mark in the gene. In **Chapter4**, we uncovered the genome-wide DNA methylation events that mark a normal (PrEC) and cancer (LNCaP) prostate cell line genomes using novel technology termed Methylplex-Next Generation Sequencing (M-NGS). First we performed extensive independent validation by bisulfite sequencing and methylation-specific QPCRs and confirmed the robustness of this methodology in identifying methylated target regions. Detailed promoter analysis revealed the presence of diverse methylation patterns in two cell lines around transcription start sites, including direct methylation of CpG islands, methylation of regions flanking CpG islands, and methylation of promoters devoid of CpG islands.

Our integrative analysis showed high correlation between methylated promoters and gene repression in the corresponding prostate cells. Further analysis of LNCaP methylated and repressed genes (oncomine meta-analysis) with publically available prostate cancer gene expression dataset identified several potential biomarkers similar to well-characterized gene GSTP1. From this list, we characterized genes including WFDC2 and TACSTD2, which exhibited cancer-specific DNA methylation pattern.

In our next analysis we integrated DNA methylation (considered as a silencing mark) and histone 3 lysine 4 trimethylation (H3K4me3, considered as an active histone mark) ChIP-Seq data in LNCaP cell line to study the relationship between the two epigenetic modifications. This analysis revealed several interesting scenarios that are listed below 1) Methylated promoters that contain CpG islands (CGI) had more H3K4me3 compared to methylated promoters without CGI 2) Among CGI promoters that contained both DNA methylation mark and H3K4me3, the modifications had distinct boundaries and were mutually exclusive. 3) Alternate transcription start sites of several genes had mutually exclusive genomic modification with either H3K4me3 or DNA methylation marks, showing differential regulation of transcript isoforms. The last observation later enabled us to identify differential regulation of specific transcript isoform expression by DNA methylation. Based on H3K4me3 occupancy and DNA methylation on genes with multiple TSSs, we nominated candidates including RASSF1 and NDRG2 for isoform-specific methylation and they were validated. This isoform-regulation by DNA methylation is not limited to the cell line samples and also present in patient tissues. We demonstrated differential expression of NDRG2 transcript isoform in a tissue panel by qRT-PCR as inferred from the cell line model. As seen in our LNCaP cells, variants 1-4 were significantly under-expressed as compared to variants 5-8 in localized PCa (p-value=0.034) and adjacent benign prostate (p-value=0.012), but not in normal (non-prostate cancer) samples. These results indicate the presence of a global cancer-specific DNA methylation that regulate transcript isoform expression by regulating the use of alternate transcription state sites.

Some of the limitations of our M-NGS methodology are the inability to detect all methylated regions in the genome mainly due to the lack of restriction enzyme recognition sites in a given target region, does not provide information on hemi-methylation status, and quantitative measure of percent methylation of each CG residue cannot be calculated. The last limitation can be addressed by including a bi-sulfite conversion in our M-NGS workflow and this is one of our future goals.

Finally, we expanded our M-NGS methylation study using M-NGS from cell lines to a panel of prostate tissue specimens that include normal, benign adjacent, localized and metastatic prostate cancer in **Chapter 5**. As seen in **Chapter 2** with genomic copy number changes, we hypothesized each tumor type may exhibit methylation patterns that distinguish each other. As expected, distinct methylation patterns were revealed in each sample group, and the cancer-specific regions, had the highest concordance with gene repression among the samples. Additionally, we identified the methylation patterns observed on gene promoters, coding regions, intergenic regions, CpG islands, and regions that harbored microRNAs, among others in each sample type.

In tissues, DNA methylation had major effect on gene repression that nearly 60% of methylated genes in prostate cancer samples were repressed. In an independent analysis, hierarchical clustering with the list of genes methylated (M-NGS data, cluster I and II) in prostate cancer of a large prostate cohort (n=155) gene expression dataset robustly differentiated normal and cancer samples, and we found an enrichment in gene repression. The mechanisms behind remaining 40% of genes that are methylated, but not repressed remain to be characterized. However, the one mechanism that we have an explanation for is gene amplification. The genomic aberration and DNA methylation are two independent mechanisms for transcriptional regulation; however they can co-occur within same genomic regions. In case of the co-occurence of genomic deletion and DNA methylation, both events were driving the gene to get repressed. However, in the event of genomic amplification and DNA methylation, due to two opposite forces in transcriptional regulation, neither the enrichment of gene over-expression nor under-expression is observed.

Related to these studies in methylation is Knudson’s two-hit hypothesis, established 20 years ago, which postulates that two genomic or epigenomic events is required to inactivate a tumor suppressor gene. While the frame-shift or nonsense mutations are the most frequently reported first hit, the DNA methylation event is also proposed to be one of two hits for tumor suppressor genes such as VHL and MLH1. We used a 3-way integrative analysis on copy number, DNA methylation, and gene expression, to identify regions displaying co-occurrence of copy number loss and DNA methylation, supporting a two-hit model for gene silencing. The methylation-harboring regions were mostly associated with methylation-deletion over methylation-amplification events with statistical significance, and moreover, the genomic regions with highest two-hit frequency include tumor-suppressor genes such as NKX3.1, RB1, and SMAD4 and several novel candidates.

Though our current study **Chapter 5** has a small sample size, and addition of more samples in future will add strength. The observations made here are very significant and in addition was able to classify gene expression dataset from a large sample cohort. This implies that our dataset has captured the key predominant epigenetic events (DNA methylation) that underlie prostate tumorigenesis.

In summary, we characterized two major types of genome-wide events that occur in prostate cancer in order to understand their global implications on transcriptional regulation and contributions to this disease phenotype. Some aspects that will be further investigated in the lab are 1) The genetic aberrations that characterize ETS fusion negative samples. 2) We validated one of our methylation candidate genes, WFDC2, which is both repressed and frequently methylated in prostate cancer tissues at high levels similar to the previously known gene GSTP1, and is also repressed in prostate tumor samples. Evaluating the performance of this and other methylation targets identified in this study as methylation biomarkers is of immediate interest. 3) Studies will be carried out on select epigenetic target regions with functional relevance to further our understanding of prostate cancer biology. 4) Characterizing more samples by methylation sequencing and increase our sample number. Finally, the datasets generated in this thesis work will be a valuable public resource for several genome-wide analyses in future.

APPENDIX

CHAPTER 2

Minimal Common Regions (MCRs) table for Chapter 2 is available online at the following address:

http://cancerres.aacrjournals.org/cgi/data/67/17/8229/DC1/4

Minimal Common Regions (MCRs) Table. The genes within the MCRs that meet the cutoff threshold are listed under amplified/deleted candidate gene category. The over- and under-expressed genes from matching mRNA data are also listed, where the cutoff values for over- and under-expressed genes are defined as $\log_2$ ratio of >0.4 and <-0.4 (±4 standard deviations of the middle 50% quantile of corresponding data).

Multiple individuals contributed to the work presented in these chapters. Contributions of individuals for each chapter are as follows:

CHAPTER 2 and 3

Jung Kim, Saravana Dhanasekaran, and Arul Chinnaiyan conceived the experiments and wrote the manuscript represented in this chapter. Jung Kim performed all in silico analyses. Jung Kim and Saravana Dhanasekaran performed aCGH profiling of prostate cancer tissues and cell lines. Scott Tomlins performed gene expression profiling of prostate cancer tissues. We like to thank A. Menon for help with microarray production, D. Rhodes, B. Laxman, S. Subramaiam, for helpful discussions, S. Bhagavathula, J. Siddiqui, R.Varambally for tissue database help. A.M.C. is supported by a Burroughs Wellcome Foundation Award in Clinical Translational Research. K.J.P. is supported by an American Cancer Society Award. S.A.T. is supported by the Medical Scientist Training Program and a Rackham Pre-doctoral Award. This research was supported in part by National Institutes of Health Grant RO1 CA97063 (A.M.C and D.G.); RO1 CA102872 (K.J.P.); U01 CA111275 (A.M.C and D.G.); P50 CA69568 (K.J.P., A.M.C and D.G.); Department of Defense PC040517 (R.M.); and PC051081 (A.M.C.); and the Ralph Wilson Medical Research Foundation Grant (K.J.P.).

CHAPTER 4 and 5

Jung Kim, Saravana Dhanasekaran, and Arul Chinnaiyan conceived the experiments and wrote the manuscript represented in this chapter. Jung Kim performed all in silico analyses. Jung Kim, Saravana Dhanasekaran, and Dan Robinson performed next generation sequencing preparation of cancer tissues and cell lines. Jung Kim and Saravana Dhanasekaran performed bisulfate sequencing and methylprofiler qPCR. Xuhong Cao performed Agilent aCGH profiling of prostate cancer tissues. Xiaojun Jing performed Agilent GE profiling of prostate cancer tissues. John Presner performed 5’RACE and qRT-PCR on methylation-driven isoforms. We thank Vladimir Marakarov, Emmanuel Kamberov Takao Kurihara, and John Langmore from Rubicon Genomics for

early access to the methylplex technology. We thank Robert Lonigro, Shanker Kaylana-Sundaram, Terrence Barrette, Mike Quist, and Jianjun Yu for help with data analysis, Roger Moorey for help with illumina sequencing, Xiaojun Jing, Rajal Shah, Bo Han, and Javed Siddiqui for providing prostate tissue samples, and Jill Granger for editing this manuscript. C.A.M. was supported by a National Institutes of Health (NIH) Ruth L. Kirschstein postdoctoral training grant, and currently derives support from the American Association of Cancer Research Amgen Fellowship in Clinical/Translational Research and the Canary Foundation and American Cancer Society Early Detection Postdoctoral Fellowship. J.Y. was supported by NIH Grant 1K99CA129565-01A1 and Department of Defense (DOD) Grant PC080665, and her current address is Division of Hematology/Oncology at Northwestern University, IL. J.R.P was supported by a NIH Cancer Biology Training Grant CA009676-18. J.R.P is a Fellow of the University of Michigan Medical Scientist Training Program. A.M.C. was supported in part by the NIH (Prostate SPORE P50CA69568, R01 R01CA132874), the DOD (BC075023, W81XWH-08-0110), the Early Detection Research Network (U01 CA111275), a Burroughs Welcome Foundation Award in Clinical Translational Research and a Doris Duke Charitable Foundation Distinguished Clinical Investigator Award. This work was also supported by National Center for Integrative Biomedical Informatics Grant U54 DA021519.