MUTATIONAL ANALYSIS OF NON-CODING GENOME IN PROSTATE CANCER USING WHOLE GENOME SEQUENCING.

Sanamjeet Singh Virdi
University of Tampere
Institute of Biomedical Technology
Bioinformatics

Master's Thesis
June 2014

Acknowledgement

My sincere gratitude to the head of Computational Biology group, also the supervisor for the thesis Prof. Matti Nykter, for accepting and providing me the opportunity to work on such a fascinating topic. Under whose humble and inspiring supervision I was able to learn and successfully execute the project. His spontaneous replies and comments to the questions were appreciable. Knowledge and experience gained during the time spent in the research group no doubt will be beneficial in future. The whole learning experience with working in the group was great.

I am also grateful to the former head of the department Prof. Mauno Vihinen for accepting me in the program, my sincere thanks to then coordinator of the program Dr. Martti Tolvanen for his patience regarding all the courses he supervised, his advices and intelligent discussions on interesting topics over the years had a great impact on my studies.

I also like to mention my colleagues in the group, the PhD students Matti Annala, Tommi Rantapero and Francesco Tabaro to thank them for their consistent help in the project by clearing my doubts and improving my technical skills. It was a memorable experience working with them.

My thanks to the classmates of the bioinformatics program for their support, also big thanks to all my beloved friends whom I met in Finland during the masters making my stay memorable for lifetime.

In the end, thanks to my family and friends back home for their love, support and encouragement without whom it could not have been possible.

June 2014
Sanamjeet Virdi

MASTER'S THESIS

Place:	University of Tampere Master's Degree Program in Bioinformatics Institute of Biomedical Technology Tampere, Finland
Author:	Sanamjeet Singh Virdi
Title:	Mutational analysis of non-coding genome in prostate cancer using whole genome sequencing.
Pages:	68
Supervisor:	Prof. Matti Nykter
Reviewer:	Prof. Matti Nykter and Juha Kesseli
Time:	June 2014

Abstract

Prostate Cancer is a lethal disease characterized as progressive and possessing distinct molecular heterogeneity during its timespan. Vast number of abnormalities has been reported till now with number of somatic mutations and germline risk factors in addition to rearrangements in the chromatin. All of them together makes the architecture of prostate genome very complex and hard to understand. Transcription machinery is central to gene regulation, with a considerable vacuum in information till date related to abnormalities in the regulatory regions in prostate genome we have tried to explore the non-coding genome and detect mutations in regulatory regions to find if they hold any significance in prostate cancer.

Using prostate cancer sample data from the whole genome sequencing study by Berger *et al.*, 2011 which were seven matched normal-tumor genomes, encouraging results were produced by applying a pipeline of independently selected tools for variant analysis. DNase I hypersensitivity sites (DNaseI HSs) for LNCaP cell lines obtained from The Encyclopedia of DNA Elements (ENCODE) Consortium were used as markers for regulatory region in the genome and from initial 54679 SNVs detected across 7 samples, 21 promoter and 621 enhancer mutations were detected overlapping the DNaseI HS peaks. Out of which 4 and 21 mutations in promoters and enhancers respectively were the finally filtered out whose genes where found relevant directly or indirectly in prostate and other cancers by performing extensive search in the published literature.

For concrete evidence validation would be the next step as results were mere implications, but our study is a diligent effort in delineating the intricate genomics involved in prostate cancer outside the gene coding regions.

Contents

Abbreviations

BWT	Burrows wheeler transform
ChIP	Chromatin immuno-precipitation
CNV	Copy number variation
dbSNP	SNP database
DNA	Deoxyribonucleic Acid
DNase	Deoxyribonuclease
DNaseI HS	DNase I hypersensitivity site
FET	Fisher’s exact test
GRC	Genome Reference Consortium
INDEL	Insertions/Deletions
LCR	Locus control regions
LOH	Loss of heterozygosity
MAR	Matrix-associated element
PCa	Prostate Cancer
RNA	Ribonucleic Acid
SAM	Sequence Alignment/Map
SAR	Scaffold-attachment region
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
TF	Transcription factor
TFBS	Transcription factor binding sites
TSS	Transcription start site
TTS	Transcription termination site
UCSC	University of California, Santa Cruz
UTR	Untranslated region
WGS	Whole genome sequencing

List of Figures

List of Tables

1. Introduction

Prostate Cancer (PCa) is one of major cancers affecting men which is lethal whose rate of incidence is higher among older men specifically in developed countries where it ranks second in cancer related deaths (Huang *et al.*, 2014). The disease has characteristics of being progressive, metastatic and ultimately leading to death of the patient.

With advent of new technologies like Next generation sequencing (NGS) and accumulation of huge amounts of data, advancements have been made in unraveling the molecular genetics of PCa. Tomlins *et al.*, 2009 in their study have described PCa as aberrations in tumor suppressors and promoter oncogenes. In addition to it they have revealed a high complexity of molecular events such as most common *TMPRSS2:ERG* and other ETS fusions which was seen consistent with the followed studies by Jindan Yu *et al.*, 2010 and Berger *et al.*, 2011. Distinct patterns can be seen in the disease fore example, mutations only specific to a subtype of PCa like *SPOP* gene occurring only in ETS negative tumors (Barbieri *et al.*, 2012) divides PCa into two fusion subtypes. Although most of the PCa subtypes are indolent but there are cases where disease is very aggressive and evades clinical therapy through different molecular mechanism which all together makes understanding the PCa very challenging.

Heterogeneity of the disease has been revealed by numerous studies such as by Berger *et al.*, 2011, Barbieri *et al.*, 2012 and others, which describe the mutational landscape of the PCa genome. Both studies were emphasized on protein coding genes and argue that the genesis of disease lies in chromatin and transcription aberrancies. Transcription in itself is a complex phenomenon involving different regions of genomes which regulate it. Though it can be seen in the studies highlighting various mutations in the genes and the rearrangements they have left out a part of puzzle which involves the regulatory regions. As gene regulation is central to its transcription it becomes important to investigate this aspect of PCa genome. An evidence to this hypothesis can be seen in study by Huang *et al.*, 2014 where PCa risk allele was located within a functional transcription factor binding sites (TFBS).

The aim of the research was to focus on the before mentioned aspect and is stressed on to explore aberrations in regulatory regions of transcription machinery. This would enable us to broaden the understanding of PCa on a whole genome level. The study has utilized the data from the work done by Berger *et al.*, 2011 and used the already sequenced genome data to identify regions of interest in the PCa genome primarily to look for the somatic mutations in the regulatory regions.

The present study with the help of advanced NGS tools and techniques delineate the mutations in non-coding regions and examines weather it is of any relevance to PCa through establishing a link to literature. A direct comparison of seven high quality normal-tumor samples (sequenced) was conducted which provided us with results where mutations in promoters of 4 genes and in enhancers of 21 genes were found related directly or indirectly to PCa or other cancers. Furthermore, validation of mutations and their presence in the TFBS as evidence of their participation in PCa is needed for development of more effective clinical therapies.

2. Review of Literature
2.1 Human Genomic Scenario

Nuclear DNA and the mitochondrial DNA together combined are termed as human genome. More than 99 % of total genetic information is confined to nuclear genome (Strachan T, 1999a). The double Stranded nuclear DNA combined with histones and non-histone proteins constitutes the chromosome, which is distributed among 22 pairs of autosomes and two sex chromosomes X and Y. Human chromosomes are formed of accessible (less condense euchromatic) and inaccessible (dense heterochromatic) domains. DNA sequencing and identification of exons has revealed that majority region of the human genome is considered to be non-coding (Lodish H *et al.*, 2000). Apparently, only 3 percent of human genome represents coding sequences (Strachan T, 1999a). The remaining non-coding genome does not account for the fact that it is totally non-functional or junk. Thorough analysis of transcripts and regulatory information are essential for the identification of genes and regions which regulate them which can help us in working of whole genome (Djebali *et al.*, 2012; Gerstein *et al.*, 2012; Neph *et al.*, 2012; Thurman *et al.*, 2012).

Functional elements of genome are not only genes coding for proteins, it also includes genes coding for non-coding RNA and genomic portions taking part in any kind of biochemical pathway be it protein binding or chromatin structure. It has been found that majority (80.4%) of human genome assists or is a part of at least one biochemical event whether it is transcript or chromatin associated (Bernstein *et al.*, 2012). The study done by the Encyclopedia of DNA Elements (ENCODE) Consortium provides evidence to the fact that this majority non-coding genome cannot be neglected, most of which lies near regulatory regions of DNA or regions of open chromatin, transcription start sites, intergenic and intronic.

Human genome is not merely linear sequence of coding (of which 75 % is transcribed at some point in some cells (Djebali *et al.*, 2012)) and non-coding regions. It is a very intricate architecture of regulatory networks (proteins binding to DNA), non-coding transcripts, regions of open chromatin and chromosome modifications in addition to promoters and enhancers working as an organization resulting in proper functioning and specialization of the cell. Short and tandem repeats, regulatory factor binding regions and small RNAs, broad histone marks, transcripts (protein coding and non-protein coding), transposable elements and pseudogenes, segmental duplications and structural variants, these elements are all parts of non-coding region. Abnormalities whether at single base level (SNPs or SNVs) or at chromosomal level can result into genetic dysfunction.

2.1.1 Regulation of Gene Expression

Human genome is very highly organized and complexity it's of regulation is extraordinary. Transcriptional control can be defined as regulation of protein synthesis by controlling formation of respective mRNA. The regulation of genes through complex machinery defines the cell type, its function and phenotype. Gene expression regulation has its significant roots at transcription initiation. These are mainly proteins binding onto DNA (also non-coding RNAs) or DNA itself structurally organized to form this machinery. Transcription initiator RNA polymerase has access to every promoter to which it can bind in absence of enhancer and insulators or repressors. Though access to promoters is restricted, for transcription to initiate chromatin undergoes many structural changes. The regulation can be both negative and positive, positive regulation being the dominant one. Any genome dysregulation can cause aberrant gene expression which lies at the heart of multiple diseases. Promoters, enhancers, silencers, insulators or repressor and locus-control regions (LCRs) are the regions or sequences of DNA in genome to which transcription factor bind and regulate genes (Birney *et al.*, 2007; Maston *et al.*, 2006)(Figure 2.1).

2.1.1.1 Transcription Factors

Human genome contains about 1,700–1,900 transcription factors (TFs) (Vaquerizas *et al.,* 2009). Out of 20,000-25,000 genes in human genome each of them having distinct expression pattern and very few of them encodes for TFs, which confers different permutations of same regulatory elements controlling gene expression given the fact that promoters there are multiple regulatory elements. Binding sites of transcription factors can be characterized at genome scale by chromatin immune-precipitation followed by microarray: ChIP–chip or sequencing: ChIP-seq (Birney *et al.,* 2007; Johnson *et al.,* 2007; Robertson *et al.,* 2007).

Transcription factors are the proteins which regulate the transcription of genes by binding to the gene specific sequences (TFBS). Transcription factors in human genome are responsible for regulation of hundreds to thousands of downstream genes (Johnson *et al.*, 2007). Not only genes but TFs also regulates other TFs to form a cross regulatory network. These networks are the foundations of the biological networks. Any variation in TFBS can affect the binding of the transcription factor resulting in dysregulation of genes.

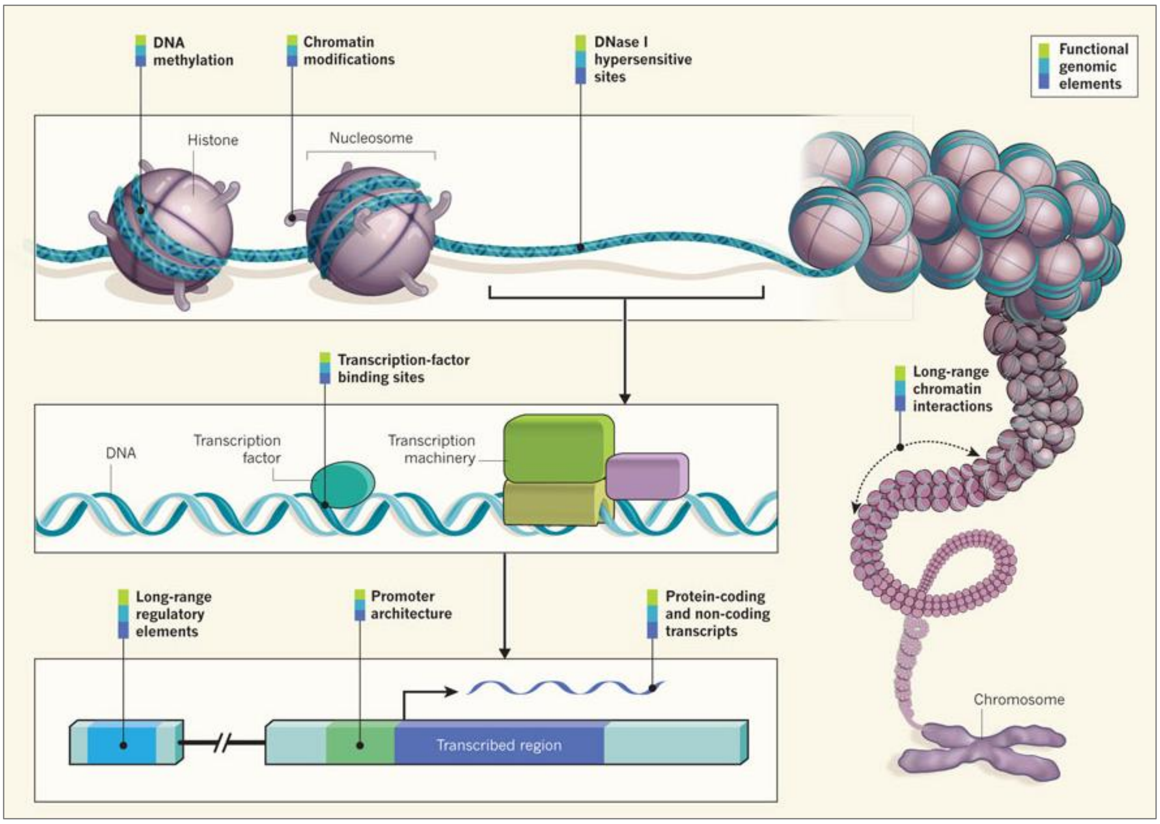

DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

Figure 2.1: Genome architecture, which describes the functional genomic elements that orchestrate the development and function of a human.
(Adapted: ENCODE Consortium, Bernstein *et al.*, 2012)

2.1.1.2 Promoters

Promoters are the crucial regulatory sites that alter the expression of the nearest gene. "These are the loci (overlapping transcription start sites) of accurate initiation of transcription mediated by transcription factors and various chromatin-remodelling and -modifying enzymes" (Cairns, 2009). Promoters (core promoter and proximal regulatory elements) are *cis*-elements for *trans*-acting TFs on which they bind to form transcription pre-initiation complex, whose catalytic activity is to assist in accurate positioning of DNA-dependent RNA polymerase (Lenhard *et al.*, 2012). The interaction between RNA polymerase and promoter sequence defines regulation of transcription initiation. Promoter sequences can vary considerably which affects the binding affinity of RNA polymerase thus affecting transcription initiation. Promoters can be of many classes characterized by their respective functions. It is the core promoter which acts as a binding site for basic transcriptional machinery and where pre-initiation complex (PIC) is assembled before the transcription start site (TSS). Other regulatory proteins termed as activators binds to the proximal promoters or TFBS (<1kb from core promoters) which synergistically act as a regulatory elements. A mutation at any of these sites can cause aberrant transcription (Maston *et al.*, 2006).

2.1.1.3 Enhancers

Enhancers are also the *cis*-regulatory elements which increase the transcription of genes. They can be located up to <1Mb from the promoters. Human genome contains hundreds and thousands of enhancers scattered across 98% of the genome which is non-protein-coding. Enhancer function is independent of their orientation and distance from their target promoter (or promoters) (Pennacchio *et al.*, 2013). They are the clusters of close TFBSs which work cooperatively to enhance transcription (Maston *et al.*, 2006). Enhancers can be found upstream, downstream or in introns suggesting that the location relative to their genes in *cis* is highly variable. Individual enhancers have also been found to regulate more than one gene (Mohrs *et al.*, 2001). Also it is not necessary that they act on the promoters located in their vicinity, enhancers can also regulate genes located distantly along the chromosome thus bypassing the nearby genes.

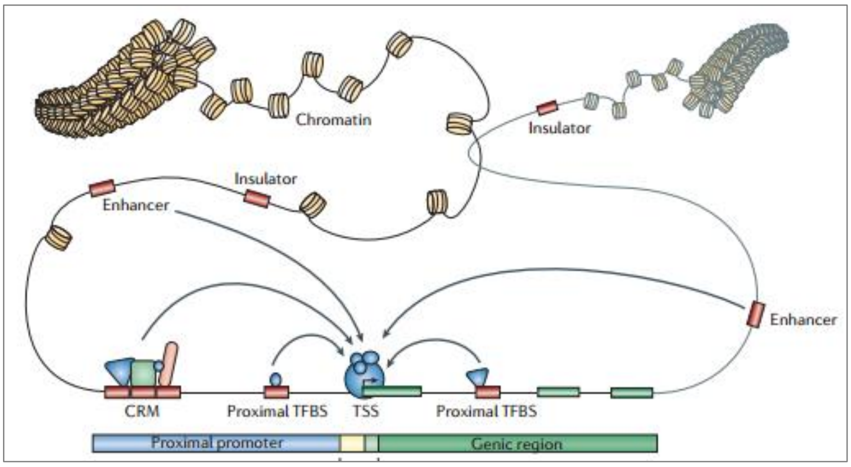


Figure 2.2: A summary of promoter elements and regulatory signals (Adapted: Maston et al., 2006) Yellow genomic portion is the smaller core promoter just in the vicinity of TSSs.

2.1.1.4 Silencers

Just like enhancers but opposite in their function, Silencers are the negative regulatory elements.

They have a silencing or repressing effect on gene expression and act as binding sites for repressors.

2.1.1.5 Insulators

"Insulators are DNA elements that insulate genes located in one chromatin domain from promiscuous regulation by enhancers or silencers in neighboring domains" (Raab & Kamakaka, 2010). They block the effect of transcription on nearby gene by binding onto the regulatory sequence or operator of that gene.

2.1.1.6 Locus Control Regions (LCRs)

LCRs are a composed group of enhancers, silencers, insulators and nuclear matrix or chromosome scaffold-attachment regions (MARs or SARs), which regulate entire gene cluster or genetic locus (Maston *et al.*, 2006). The collective activity of all the elements in the region defines the functionality of LCRs. They are located in the region upstream of their respective gene, but can also be found within introns.

2.1.1.7 Histone Modification and Open Chromatin Sites

DNA is densely packaged into chromatin and this cell type specific organization affects the access and activity of many regulatory elements. Structurally chromatin is formed by DNA folding into nucleosome which is in turn an octamer of histone proteins onto which 147 bp of DNA is wrapped. "The nucleosome positions along with histone variants and modifications make up the primary structure of chromatin"(Zhou *et al.*, 2011). DNA in chromatin may remain accessible to DNA-binding proteins such as TFs and RNA polymerase II (RNAPII) or may be further compacted. Chromatin can also organize into higher-order structures such as nuclear lamina-associated domains and transcription factories. Aspects of gene and genome regulation are reflected at every level of organization (Zhou *et al.*, 2011). The main feature of chromatin organization of DNA is whether it is transcription active or inactive. Transcriptionally active parts of chromatin are structurally different from the inactive parts. The compact structure in addition to methylation of cytosine acts a barrier to protein binding, thus access to open chromatin affects the regulation machinery.

"Specific chromatin configurations may be dictated by DNA sequence, DNA methylation patterns, transcription factors and other regulatory proteins, and transcriptional activity, and may be maintained through epigenetic controls that are rooted in the chromatin machinery" (Margueron & Reinberg, 2010; Zhou *et al.*, 2011). Better understanding of these modifications may provide insight onto how the genome is regulated in normal and diseased cell type.

2.1.2 Transcriptional Dysregulation and Diseases

Mutations in the regulatory elements of transcription can have an abrupt effect on gene expression. In addition, mutations in chromatin remodeling factors have been shown to be associated with cancers (Maston *et al.*, 2006). For example, in "*Brahma (Brm)* and *brahma-related gene-1 (Brg1)* (which are mammalian homologues of SWI/SNF chromatin-remodeling factor subunits, that can regulate both transcriptional activation and repression) are found to be deleted in numerous cancer cell lines, leading to the altered expression of genes that influence cell proliferation and metastasis"

(Banine *et al.*, 2005). Another example of mutations in regulatory elements is Bernard-Soulier syndrome where the proximal promoter of Glycoprotein-Ibbeta harbored mutation which changes the GATA consensus binding site, disrupting GATA1 binding to the mutated site and decreases promoter activity by 84% (Ludlow *et al.*, 1996). In PCa, somatic mutations have been detected in the transcription factor *ATBF1*, which impair its function one of which is to inhibit cell proliferation (Sun *et al.*, 2005). Chromosomal rearrangements (translocations) have also been detected in various types of cancers, which can result in involvement of regulatory regions causing aberrant gene expression where enhancer or promoters of one gene may get linked to a proto-oncogene, as seen in activation of *C-MYC* gene (the cellular homologue of the avian myelocitic leukemia virus) in Burkitt's lymphoma and acute T-cell leukemia respectively where T-cell receptor genes fuse to *cMYC* oncogene (Popescu & Zimonjic, 2006).

Most oncogenic and gene fusion products are transcription factors, and the disruption of transcriptional activity as a result might be critical in cancer types harboring rearrangements and fusions. Other case of fusion result can be generation of a completely new protein with altered activity where a transcription factor fuse with another protein, as seen in *BCR-ABL* fusion associated with chronic myelogenous leukemia (Saglio & Cilloni, 2004). Interestingly, fusions now have been associated in PCa, where *TMPRSS2* genes fuse with ETS family TFs, which causes androgen-regulated expression of *ERG* (a type of ETS family of TFs). But major role or detection of disrupted transcription regulatory machinery in other parts of PCa genome is still unknown.

2.1.3 Human Genomic Variations

Though major portion of genome is identical in humans, the variations in it make individuals different from each other. Variations can be seen as a mechanism evolution has made us survive and adapt. These variations which occur in many forms may be similar in different populations. They can occur at various levels in genome, structural in form of chromosomal re-arrangements, in DNA sequences within or outside genes and also at protein level. Nature of variations can vary from beneficial, innocuous and to deleterious.

2.1.3.1 Mutations

Mutations are the changes in the DNA sequence in which one or more than one base pair has gone under change. These changes can be random caused because of errors in DNA replication or due to external factors related to environment. These can be classified as germline or somatic. New mutations arise in single individuals, in somatic cells or in the germline. If a germline mutation can be transmitted to the offspring it can spread to other sexual members of the population (Strachan T, 1999b). Mutations can lead to changes in the structure of an encoded protein, in its decrease or complete loss in its expression. Any change in the DNA sequence affects all copies of the encoded transcript which in turn affects protein unless repaired by cell repair machinery. The result can be deleterious to a cell or organism. But, if alterations occur in the sequences of RNA or protein molecules during their synthesis, they are less damaging because many copies of each RNA and protein are synthesized. DNA alterations can be both small and large.

Different types of mutations

Point mutations: These involve alteration in a single base pair, and small deletions generally directly affecting the function of the gene.

Missense: These lead to a change of a single amino acid in the encoded protein sequence.

Nonsense: These lead to formation of stop codon resulting in unfinished encoded protein and premature termination of translation.

Frameshift: These cause change in the reading frame by insertion or deletion of any number nucleotides but three.

Silent: Mutation which do not cause change in the phenotype.

Synonymous: Mutation(substitution) which causes no change in resulting amino acid or generates an amino acid with similar properties as before but can affect transcription, mRNA transport, splicing and translation.

2.1.3.2 Chromosomal re-arrangements

These are the alterations involving large chunks of DNA. Chromosomal rearrangements encompass several different classes of events: deletions, duplications, inversions and translocations. Each of these events can be caused by breakage of DNA double helices in the genome at two different locations, followed by a rejoining of the broken ends to produce a new chromosomal arrangement of genes, different from the gene order of the chromosomes before breakage. There are two general types of rearrangements, balanced and imbalanced. Balanced rearrangements only changes the gene order rather than resulting in any loss or gain of genomic material. “The two simple classes of balanced rearrangements are inversions and reciprocal translocations in which no chromosomal material is gained or lost” (Griffiths AJF *et al.,* 1999). Phenotypic abnormalities can be caused if the rearrangement involves deletion or duplication of very large segment of DNA.

2.1.3.3 Polymorphism

Polymorphism is the phenomenon of the existence of two or more variants (alleles, phenotypes, sequence variants, chromosomal structure variants) at significant frequencies in the population. These are differences in the DNA which is not a mutation. These are the variations which can be common in different types of populations. Single nucleotide polymorphism (SNPs) is the most common form of variation which occur once every 1000s base pairs of nucleotides. If 1% of the population has same variation, it is called as a SNP otherwise it is a mutation. These can be found spread across all the genome, within coding and non-coding regions.

2.1.3.4 Copy number variations (CNVs)

Copy number variations (CNVs) are as important as component of genomic diversity as single nucleotide polymorphisms (SNPs). "Redon *et al.* (2006) defined a CNV as a DNA segment of one kilobase (Kb) or larger" that is present at a variable copy number in comparison with a reference genome (Clancy S, 2008). They can or cannot have an effect on the phenotype.

2.2 Chromatin Accessibility in Human Genome

DNA in human genome is densely packed with histones and non-histone proteins to form the chromatin. Most essential role in chromatin structure is of the histone proteins. The DNA together combined with histones is so compact and complex that it makes it inaccessible for regulatory proteins to interact with DNA sequences. Direct interaction of DNA with protein with regulatory rules needs chromatin to accommodate these proteins, so chromatin modifies its structure accordingly. The most common entity of chromatin is nucleosome which is a 147 bp octamer of DNA wound on 4 pairs of histones. The repeated units of nucleosomes form the basic structure of chromatin. In this packed structure of DNA with histones a portion of histone protrudes from the nucleosome core octamer which is vulnerable to chemical interaction known as histone acetylation which is important for genetic function. Other types of chemical interaction of tails of core histones is with methyl and phosphate groups which all together combined termed as histone modifications. These modifications can influence chromatin structure and can have a significant impact on cellular activity (Brown TA, 2002).

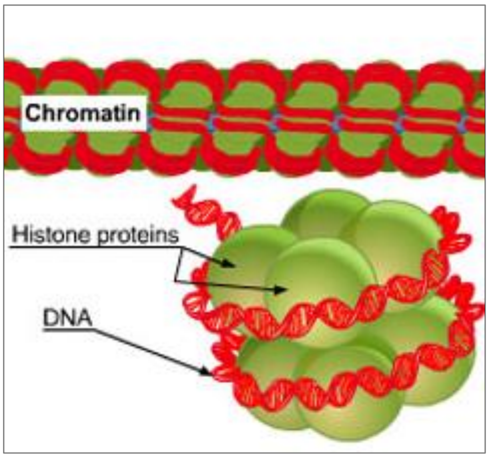


Figure 2.3: Structure of chromatin in human genome. DNA is inaccessible for regulatory proteins for interaction, thus chromatin uncoiling must take place before assembling of proteins (TFs and RNA polymerase) for transcription. (Adapted: Scott Freeman 2011)

Before transcription the compact structure of chromatin in the neighborhood of genes is disrupted (histone acetylation) at nearby promoters and enhancers which give rise to a short stretch of DNA (hypersensitive sites) which have a property of unusually sensitive to nucleases and chemical probes (Felsenfeld *et al.*, 1996). These DNase I hypersensitive sites (DNase I HSs) in chromatin can be regarded as markers of regions of gene regulation are used to map regulatory DNA regions.

2.2.1 DNAseI HSs as Markers for Regulatory Regions

DNase hypersensitive sites encompass all the *cis*-regulatory elements of which it overlies directly and it is maximal at core region of regulatory occupancy (Thurman *et al.*, 2012). The precise delineation of regulatory sites by DNaseI has made genome wide mapping of DHSs followed by NGS methods an exemplary way of discovering and cataloguing regulatory region of DNA. DHSs are flanked by nucleosomes, thus along the genome the density of DNaseI cleavage can be seen as a quantitative continuous measure of chromatin accessibility as function of genome position. Those flanked regions can also be termed as peaks as the density will be considerably higher than the occupied flanking regions.

ENCODE project consortium have done extensive work to collect the information regarding most aspects of human genome. Majority (97.4 %) of DNaseI HSs encompass the elements was validated by comparing to the validated regulated elements already located (Thurman *et al.*, 2012). Thus confirming DNaseI HSs as markers for regulatory DNA. Not only marking of the regulatory regions, a change in DHSs dynamics can also be used as a predictor for cell type TF binding. DHSs encompass variety of regions that are related to many TFs but also as in their study He *et al.*, 2012 have shown quantitative DNaseI dynamics can be used to predict TF binding, where AR and ESR1 are shown to have highly predictive binding in PCa and breast cancer respectively. He *et al.,* outlined three major factors describing DHS. First in accordance with other studies, stating majority of DHS occurs in nucleosome free regions. Second, a new finding of DHS frequently arise due to TF binding and those regions not necessarily being nucleosome free and finally DHS can change with removal of cofactors thus depending on binding of transcription factors . As two different TFs which were central to their study, both *AR* and *ESR1* having distinct interaction modes with chromatin suggests DHS dynamics can be used a general approach for predicting cell type specific *cis*-regulatory elements (He *et al.*, 2012). Though high throughput genome wide detection of DHSs can be done to detect the regulatory elements but DHSs only gives an implication of a functional regulatory element, to determine if the site is an actual binding site for the protein to regulate a different powerful technique of chromatin immunoprecipitation emerged as an advantage. Also, this technique and detect an actual binding site but they do not demonstrate the functionality of the element in regulation of a target gene (Maston *et al.*, 2006).

2.2.2 Collection and Annotation of Functional Elements of Human Genome

With advent of state of art techniques like ChIP-seq (Landt *et al.*, 2012), DNase-seq (Song & Crawford, 2010) various research groups (more importantly ENCODE project consortium) have accumulated vast amount of data for transcription factor binding sequences, TFs and other *cis*-regulatory elements in different cell types. For example Thurman *et al.*, 2012 under the umbrella (ENCODE, Birney *et al.*, 2007) of annotating all the functional elements in human genome, have identified over 2.9 million DHSs through genome wide profiling in 125 diverse cell type and tissues. The findings by Thurman *et al.* were quite intriguing as they quantified and compared DNase I data from ENCODE common cell cline K562 with ChIP-seq signals from 42 TFs from ENCODE ChIP-seq to reveal the parallels between ChIP signals and DNase sensitivity across the genome. With many number of TFs shown to bound hypersensitive site as in case of beta-globin locus, the correlation between ChIP-seq signal and nuclease sensitivity at hypersensitive site II was explained by interaction between DNA bound proteins to other interactive protein which increased the likelihood of accessibility of chromatin (Thurman *et al.,* 2012). These finding were consistent with one of the findings in study done by He *et al.,* 2012.

With extensive resources available due to ENCODE, like active regulatory elements identified as open chromatin in multiple cell types (which has revealed novel relationships between chromatin accessibility and transcription) has widen the scope of analyzing cell selective gene regulation and recognize patterns of regulation also from distal elements (Thurman *et al.*, 2012). Additionally, “genetic variations in regulatory elements are a key driver of evolution and disease” (Amit *et al.*, 2009; Nicolae *et al.*, 2010; Wray, 2007). The information regarding the location of the functional gene regulatory elements in the genome can have a major impact in characterizing non-coding variants.

2.3 Prostate Cancer – An Overview of Genomic Complexity

In the developed world PCa is the most common non-skin cancer among men. Major risk factors include age, ethnicity, diet and heredity (Understanding Prostate Cancer). "The prostate sits in the pelvis, surrounded by the rectum posteriorly and the bladder superiorly" (Oh WK *et al.*, 2003). It is a fluid secreting gland which secretes 30 % of the sperm and the prostatic secretion facilitates the sperms movement during ejaculation (Guyton & Hall, 2006). Many factors have been associated with the genesis of PCa. Progression of PCa into metastasis generally results patient's death. PCa can have long time span of disease progression, varying from localized PCa to developing into a lethal aggressive disease. Recent whole genome studies have revealed that germline SNPs are associated with increase in the risk of PCa, also somatic mutations in addition to chromosomal rearrangements and recurrent gene fusions, which all together makes it a multi-factoral cancer. PCa genomic alterations scenario is however incomplete (Berger *et al.*, 2011), with limited therapeutic treatments available there is a critical need to understand the genetic underpinnings lying at the heart of PCa and develop new strategies to reveal novel therapeutic targets. As Barbieri *et al.* describe it in their review it has a variable clinical course, and moreover molecular heterogeneity makes it a disease with unpredictable behavior (Barbieri & Tomlins, 2014).

To understand the molecular heterogeneity of prostate cancer whole genome studies as well exome sequencing studies have been performed utilizing NGS platforms. Recent technological advancements have enabled us in characterization of somatic mutations in large number of tumor samples through massively parallel high throughput sequencing of DNA (Watson *et al.,* 2013). As evident from studies on other types of cancers "Large-scale cancer genome characterization projects studying glioblastoma, lung, colon, pancreas and breast cancers have provided critical new insights into the molecular classification of cancers and have the potential to identify new therapeutic targets" (Cancer Genome Atlas Research Network, 2008; Ding *et al.*, 2008; Jones *et al.*, 2008; Parsons *et al.*, 2008; Sjoblom *et al.*, 2006; Weir *et al.*, 2007; Wood *et al.*, 2007). Similarly, with sequencing of whole genome of tumor samples complete map of carcinogenesis of PCa can be generated which can provide new insights into genome variation. Additionally clinical and biological heterogeneity of PCa makes it difficult to accurately predict tumor aggressiveness, thus affordable NGS techniques enables us to characterize unrevealed genomic features (Wu *et al.*, 2012).

2.3.1 Chromosomal rearrangements involving gene fusions

A number of variations, gene fusions and aberrant expression of genes which are common to PCa have been detected in PCa genome (Witte, 2009). Along with many genomic alterations between androgen regulated genes and ETS transcription factors, the most common gene fusion occurs between transmembrane protease serine 2 gene (*TMPRSS2*), which is androgen regulated and E twenty-six (ETS) transcription factors. The fusion *TMPRSS2-ERG* (One of the ETS transcription factors) has been observed in approximately 50% of the tumors studied (Kumar-Sinha *et al.,* 2008; Tomlins *et al.*, 2005, 2007).

Various studies have revealed that there is complex pattern of balanced rearrangements involved in the arousal of oncogenic gene fusions. Berger *et al.* (2011) have found range of rearrangements per genome in their samples of the study. They have stated three genes disrupted from rearrangements were *ZNF407*, *CHD1* and *PTEN* other than chromosomal rearrangements involving *TMPRSS2* and *ERG*. Balanced rearrangements as identified in the study have closed chained patterns which suggest variation resulting in interaction between multiple genomic regions is driven by localization. E26 oncogene homolog gene v-ets erythroblastosis virus (*ERG*) and the ETS variant 1 gene (*ETV1*) are the two ETS transcription factors mostly involved in recurrent gene fusions. It has been revealed that the fusion occurs between 5' UTR of the prostate specific androgen induced *TMPRSS2* and the respective ETS gene (Tomlins *et al.*, 2005). Benign prostate tissues were devoid of these fusions or they were not detected, and the fusions to *ERG* or *ETV1* were only seen when they were overexpressed. Novel 5' and 3' fusion partners have been identified by various studies. They include fusion with ETS variant 4 gene (*ETV4*) and the ETS variant 5 gene (*ETV5*) (Helgeson *et al.*, 2008; Tomlins *et al.*, 2006). Not all PCa cell lines acquire ETS gene fusions, thus they can be characterized as ETS fusion positive and ETS fusion negative subclasses. Both cell lines have found to have distinct transcription profiles (Arber *et al.,* 2000; Bohlander, 2005; Iwamoto *et al.,* 2000; Trojanowska, 2000) suggesting they are different diseases. Though it has been indicated in studies there are numerous 3' partners for *TMPRSS2*, 5' fusion partners for *ETV1* have also been identified other than *TMPRSS2* which are 5' UTRs from *SLC45A3*, *HERV-K_22q11.23*, *C15ORF21*, and *HNRPA2B1*. These partners identified can fuse with other members of ETS family creating possibilities of rare gene fusions in PCa (Helgeson *et al.*, 2008).

2.3.1.1 TMPRSS2-ERG gene fusion

"In prostate cells, androgen signaling has been shown to induce co-localization of *TMPRSS2* and *ERG*, thereby allowing double strand breaks to facilitate gene fusion formation" (Haffner *et al.*, 2010; Lin *et al.*, 2009; Mani *et al.*, 2009). The fusion resulted in androgen-regulated expression of ERG (which is a transcript regulator). Thus androgen responsive elements that normally restrict expression of *TPMRSS2* drove aberrant overexpression of 5' truncated *ETS* oncogenes (Tomlins *et al.*, 2007). Though this fusion has been seen in 90 % of the samples in the study by Tomlins *et al.*, 2005 and overexpression of *ERG* too, it confers fusion driven mechanism of *ERG* overexpression as revealed by Demichelis *et al.*, 2007 in their study of gene fusion associated with lethal prostate cancer. The role of the fusion has been inferred by Yu *et al.*, 2010 in attenuating androgen signaling via inhibition of *AR* expression including attenuation of *AR* at gene specific loci.

TMPRSS2 and *ERG* are located <3 Mb apart on chr21, so fusion can occur either with inter-chromosomal insertion or deletion of intervening region, also due to chromosomal translocations. Recent studies have shown varied frequency of this particular fusion across different tumors, highest in clinically localized samples (tumor confined to prostate) and lowest in benign prostate hyperplasia samples (Kumar-Sinha *et al.*, 2008). Many different spliced isoforms of mRNA of this particular gene fusion have been observed. For example isoform involving exon 2 of *TMPRSS2* in frame with exon 4 of *ERG* has been associated with aggressive form of PCa (J. Wang *et al.*, 2006). It can be added to stated results that various isoforms produce varied affects to PCa progression depending on their expression levels.

2.3.1.2 Other major gene rearrangements: CADM2, PTEN, MAGI2, TP53, SKP2, BRCA2.

Numerous research works has been published describing the somatic aberration landscape in prostate genome. Few of them in agreement with results of other studies validating them and expanding the work to a new level whereas, some contradictory. Most notably we have advanced to the stage where PCa is seen to be classified with distinct patterns of aberrations in the genome. In their study of whole genome sequencing of seven prostate tumors and patient-matched normal samples, Berger *et al.* reported rearrangements in 3 out of 7 samples other than *TMPRSS2* and *ERG*. *CSMD3* and *CADM2* (cell adhesion molecule) were also stated to be disrupted in PCa. "*CADM2* encodes a nectin-like member of the immunoglobin-like cell adhesion molecule" (Berger *et al.*, 2011) and several of this type of nectin like proteins have function in tumour suppression.

PTEN has been found to be a tumor suppressor gene in studies carried out that suppresses tumor cell growth and may regulate tumor cell invasion and metastasis (J. Li *et al.*, 1997). Notably, Berger *et al.* found *PTEN* gene disrupted and in addition *MAGI2* gene which encodes *PTEN* interacting protein. The disruption was a dinucleotide insertion in the *PTEN* coding sequence in one of their samples. The rearrangements disrupting both genes affects their functionality specially *PTEN* either directly or indirectly through *MAGI2*, dysregulating PI3 kinase pathway in PCa. In agreement with Berger *et al.*, in a review *PTEN* locus was reported to be deleted in 40 % of primary PCas (Barbieri & Tomlins, 2014). Notable observation in the same study was that *PTEN* rearrangements resulted in chromosome copy loss whereas *MAGI2* rearrangements were balanced. In accordance with this study Robbins and colleagues in their work of lethal metastatic prostate tumors report a homozygous deletion in one of their tumor samples which encompasses many genes including *PTEN*. Furthermore in addition to *PTEN* disruption, double strand breakage was of the *BRCA2* tumor-suppressor gene is also one of the somatic alteration stated by them (Robbins *et al.*, 2011).

Taylor *et al.*, in their study of 200 or more tumor samples have carried out analysis for copy number changes. Interestingly, they have implicated in 5 of 17 metastatic lesions studied, significant changes in *TP53, SKP2* and *PTEN*. In accordance with this study Robbins *et al.*, 2011 supported the concept of before-mentioned somatic alterations driving lethal PCa but also reveal novel somatic point mutations in genes including *MTOR*, *BRCA2*, *ARHGEF12*, and *CHD5*. Other notable deletion of gene is *CHD1*, "which encodes a chromo-domain helicase DNA-binding protein that acts to remodel chromatin states and is involved in transcriptional control across the genome". As Barbieri *et al*,. have reported *CHD1* was recurrently seen to be deleted in 10-25 percent of both primary and metastatic tumors. Although rearrangements and point mutation have been detected but lesions in *CHD1*, specifically focal homozygous deletions are restricted to ETS-negative tumors. There is considerable increase in other genomic rearrangements harbored by PCas which have *CHD1* deletion as stated in review by Barbieri *et al.* (Barbieri & Tomlins, 2014; Barbieri *et al.*, 2012; Burkhardt *et al.*, 2013; Grasso *et al.*, 2012; Taylor *et al.*, 2010).

2.3.2 Germline Variations

Risk of inheriting cancer is elevated due to germline mutations in genes (Park BH, 2003). This fact has been shown in various genome wide association (GWA) studies as germline SNPs were detected to be associated with PCa (Witte, 2009). Lichtenstein *et al.* have shown that there is a very high probability of PCa inheritance and given this fact, the risk involving germline mutations in genes cannot be denied. Several chromosome loci have been shown by studies to harbor these mutations such as mutations in chromosome 8q24 region which are associated with PCa and this region has also been associated with colorectal, breast, ovarian and bladder cancers (Amundadottir *et al.*, 2006; Witte, 2009). In addition chromosome 24 loci and other PCa variant loci on chromosome 10 and 17 have also been identified. The one of the two loci is found to be associated with SNP (located near promoter of *MSMB* gene) in PCa (Witte, 2009). Beke *et al.* found in their study that deregulation of *MSMB* gene increases PCa advancement (Beke *et al.,* 2007). All these GWA studies have put a new insight in PCa understanding by deciphering role of germline variations associated with the risk of disease and its development but overall view is still unclear about how potential use of these SNPs can explain varied aggressiveness of PCa. But studies have shown SNPs to be linked with PCa independently but when combined together they can increase the risk of the disease (Witte, 2009).

2.3.3 Somatic Mutations

The important somatic variations are the rearrangements which have been discussed already but apart from them many genes have been linked to PCa which are found to be mutated. Berger *et al.*, 2011 have given an elaborate description of somatic mutations in their study. They have stated that more than 80% of the PCa genome harbored somatic mutations (7 PCa samples in total), especially CpG dinucleotides where mutation rate was ten times more than the other regions in genome. Various genes were reported to harbor mutations though not the all the samples encompassed every mutation again enlightening complexity of PCa. Genes mentioned to be mutated were *SPTA1*, *SPOP,* chromatin modifiers *CHD1*, *CHD5*, *HDAC9,* and members of *HSP-1* stress response complex (*HSP2*, *HSPA5* and *HSP90AB1*). Other cancer genes such as *PRKCI* and *DICER* were also stated to be mutated by the study. Targeted mutational analysis done by Robbins *et al.* on matched normal-tumor samples novel mutations were revealed in *MTOR*, *BRCA2*, *ARHGEF12* and *CHD5* (Robbins *et al.*, 2011).

Of all the mutated genes listed in various studies the significant mutated gene is *SPOP*. As emphasized by Barbieri *et al.*, 2012 in their study of primary tumors and matched normal-tumor samples, in addition to 5000 mutations they found *SPOP* was “most” mutated one. In localized prostate tumors *SPOP* mutation is a common point mutation. Again, Barbieri *et al.* 2014 in their review state *SPOP* mutations to be around 6-15 % of multiple cohorts. As an evidence of distinct patterns in PCa tumors *SPOP* mutations were found to be mutually exclusive of ETS fusions (Barbieri *et al.,* 2012). In addition to this fact *SPOP* has been found to mutually exclusive with deletions and mutations in *TP53* tumor suppressor (Barbieri *et al.*, 2012; Lindberg *et al.*, 2013). A distinct pattern of *SPOP* mutation was also shown by them in the study was *CHD1* deletions were associated with *SPOP* mutations. These *CHD1* deletions are restricted in ETS negative tumors like *SPOP* mutations which make ETS negative prostate tumor class distinct from other PCas.

Furthermore, recurrent mutations have also been found in *FOXA1* (Grasso *et al.*, 2012) and *MED12* (Barbieri *et al.*, 2012). *FOXA1* is known to modulate androgen receptor transcriptional activity. In another finding in their review, Barbieri *el al.*, 2013 notify 25 to 30 percent of localized PCas harbor lesions in *TP53* which happen to an "earlier" event in PCa progression (Barbieri *et al.*, 2012). Suggesting a possibility how variation occurs in PCa in a timeline fashion.

Somatic mutations described till now were results of comprehensive work done on exomes of PCa with help of NGS. With many factors affecting abnormal working of PCa genome, the variation in other parts of chromatin than exome cannot be neglected given the fact in these parts of the chromatin resides the regulatory elements which control the gene expression precisely and carefully in spatial and temporal fashion. Genome deregulation can have major impact on cellular functions, and if proteins (TFs, activators, repressors, co-activators and co-repressors) are involved it can affect gene expression control, also in DNA repair and maintenance mechanisms (Barbieri & Tomlins, 2014). For example Enhancer of Zeste Homolog 2 (*EZH2*) reported dysregulated in variety of cancers. It can be through mutations, overexpression and other mechanisms. "*EZH2* acts as a histone methyltransferase to silence gene expression and plays a critical role in chromatin regulation" (Barbieri & Tomlins, 2014). Nature of PCa can be driven by aberrant expression as Varambally *et al.*, 2002 have shown overexpression in PCa is associated with PCa metastasis and aggressiveness. A number of genes involved in chromatin regulation (histone modification) have also been found to be mutated in PCa, including *KDM6A/UTX*, *MLL2*, and *MLL3* (Barbieri *et al.*, 2012; Lindberg *et al.*, 2013; Taylor *et al.*, 2010).

Given the genomic alterations in PCa genome, an altogether different level of understanding of the complexity genomic alterations has been made possible through whole genome sequencing rather than targeted sequencing. Gene fusions are seen as most common genetic abnormality in PCa. If we can complete the picture involving germline SNPs and somatic mutations, the understanding of mechanisms of PCa genesis and progression can be achieved. The disruption caused by them is not confined to overexpression and dysfunction, they also affect the major pathways affecting the cell function. Multiple AR-pathway signaling components are seen to have been mutated or dysregulated (Grasso *et al.*, 2012; Taylor *et al.*, 2010). AR gene disruption has been shown to have increased activity in PCa. It is amplified in 40 % of the metastatic tumors and mutated in 10 % treated metastatic tumors but absent in localized PCa (Barbieri & Tomlins, 2014; Grasso *et al.*, 2012; Taylor *et al.*, 2010). In their review Barbieri *et al.* have reported many alterations, such as in genes like *FOXA1* and *NCOA2* which increases androgen signaling and other transcriptional activity mediated by androgen that could initiated genomic rearrangements in PCa making androgen signaling pathway of critical importance in primary and advanced PCa (Barbieri & Tomlins, 2014). As Berger *et al.*, 2011 state "whole genome sequencing of large numbers of relapsing primary and metastatic PCas promises to define a genetic cartography that assists in tumor classification, elaborates mechanisms of carcinogenesis and identifies new targets for therapeutic intervention". But of more importance is discovering the timeline of occurring events which can be critical in cancer progression and aggressiveness (Barbieri & Tomlins, 2014). Some progress has been seen, as *SPOP* mutation and ETS rearrangements which are mutually exclusive events occurring earlier whereas *PTEN*, *TP53* and *AR* disruptions seen in advanced stages of tumors (Beltran *et al.*, 2013). All these findings help to make a framework of timespan of PCa from initiation, progression and in advanced cases death. With many more studies undergoing at whole genome level, it may become possible in future that PCa can evolve from poorly understood molecular heterogeneous to homogenous clinical types and more identifiable by molecular criteria amenable to specific management strategies (Barbieri & Tomlins, 2014).

2.4 Tools for Variant Analysis by Whole Genome Sequencing

To extract information from the DNA, the sequence of bases must be revealed which stores all the coding information for proteins. As classical sequencing methods are not very efficient and cost friendly at large scale, various technologies have been developed for sequencing DNA which is in use at present since few years. A generic name of "next generation sequencing" (NGS) is termed which includes many platforms for sequencing DNA depending on the sequencing approach and output. These NGS technologies have proved to very cost efficient, with high throughput but producing manageable data output and have straightforward interpretation of analysis results (Mardis, 2013). As cost of NGS methods decreased, the amount of data produced from sequencing increased exponentially thus resulting in further development of various analytical tools and algorithms. Depending on the type of analysis varies the tools selection. One of such type application of NGS is in variant analysis which involves series of five distinct steps; quality assessment, sequence alignment, variant identification, variant annotation and visualization (Mardis, 2013).

2.4.1 Next Generation Sequencing methods

There are number of platforms for NGS in the market which follow common principles of template preparation, sequencing and imaging but differ in the chemistry applied, how sequencing is performed and data output. NGS instruments instead of cloning the DNA, construct a library of fragments from the DNA to be sequenced. Each fragment to be used as a template which is covalently attached by DNA ligase to a universal adapter (specific to platforms) used to polymerase amplify the library fragments (Mardis, 2013). These templates are usually immobilized on a solid supports depending upon the platform. Each fragment yields a single sequence focus which is required to produce sufficient signal from the DNA sequencing steps that determines the sequenced data, which gives it a digital nature. This type of sequencing is also referred as massively parallel sequencing because of the steps simultaneously performed on each fragment in a stepwise manner. Nucleotide addition being the first and detection of that incorporated nucleotide on the fragment being sequenced and finally washing step which can be either performed for chemically removing unattached nucleotides after first step or to remove fluorescent labels or blocking groups (Mardis, 2013). The enormous data sets produced at the end are results of millions to billions of reaction foci to be sequenced at a time, in addition these methods perform sequencing and detection step simultaneously in a sequence, latter followed by former.

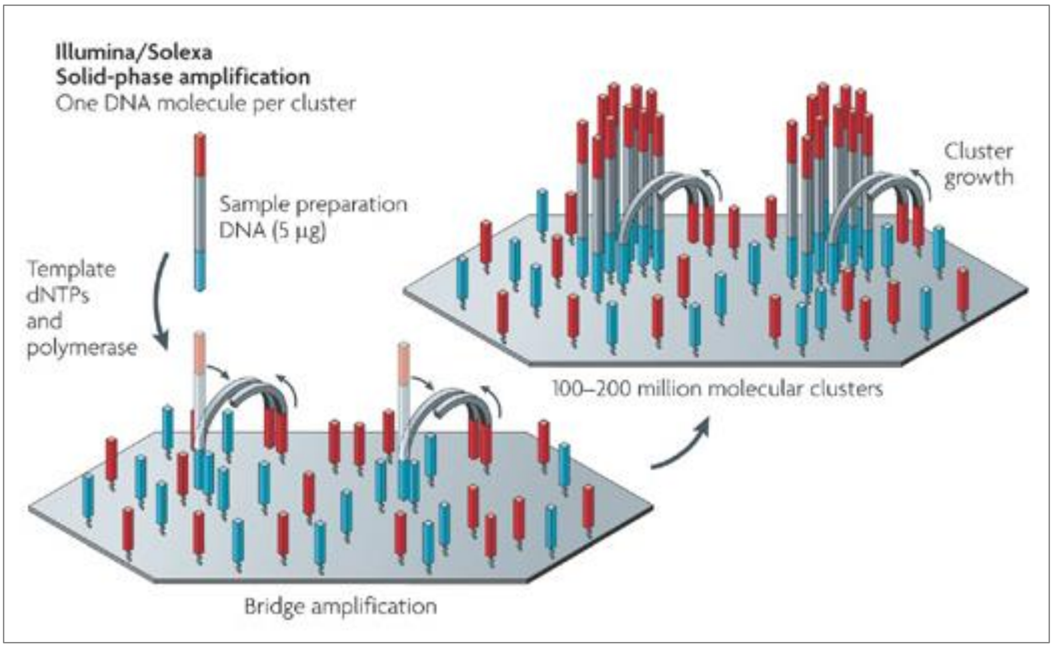


Figure 2.4: First two steps of sequencing performed in Illumina/Solexa GA, library preparation and cluster generation. (Adapted: Metzker, 2010).

Among various technologies available Illumina/Solexa Genome Analyzer (GA) is dominant one in the market currently (Metzker, 2010). This technology works in three principle steps. First, library preparation where the single DNA molecule (repaired and adenylated after the fragmentation 200~500 bp) are ligated adapters at both ends which are then isolated and purified. In second step which is cluster generation, single molecules are immobilized by hybridizing to the fixed oligos which were already further attached to the solid support (flow cell). Bound fragments are then clonally amplified to make copies through extension and bridge amplification with adjacent primer to form hundreds and millions of clusters (Metzker, 2010). Reversed strands are cleaved and washed away. With ends blocked sequencing primers are hybridized to the DNA templates, now the clusters are ready for the third and last step which is sequencing. In sequencing all the clusters are sequenced simultaneously (hence the term massively parallel sequencing is often used) base by base in parallel by adding all four fluorescently labelled reversibly terminated nucleotides where all compete for one site. The termination of DNA synthesis after addition of a nucleotide is an important step. After each step of synthesis remaining nucleotides are washed away and imaging is performed where clusters are excited by a laser to detect the incorporated nucleotide. This step is followed by a cleavage step where inhibiting/terminating group or fluorescent dye is removed allowing addition of the next base (Metzker, 2010).

Though sequencing performed is highly specific and of high quality but it is not error free, where substitutions are the most common, most number of this type of errors occurring where previous base detected is 'G' (Dohm *et al.,* 2008). The technology is vulnerable to amplification bias during template preparation which results in underrepresentation of AT-rich and GT-rich regions (Bentley *et al.*, 2008; Dohm *et al.*, 2008). The error percentage of resulting reads is said to be at most 0.5 % i.e. 1 in 200 bases. Some fragments which lag behind in extension than other fragments due to incomplete de-blocking in the previous cycle or oppositely complete lack of blocking which allows more than one nucleotide to add contributes as noise termed as phasing. Another issue which adds to noise is the fluorescence labels which fail to cleave after one cycle adds residual interference (Mardis, 2013).

Read length after sequencing has been improved in this technology from 25 bp single end reads to 150 bp paired end reads. Short reads pose a challenge of assembling as the extent of shared sequence is limited. In addition, size and complexity of the genome add to the limitation due to repetitive content of half of the genome which includes gene families (Mardis, 2013). To overcome this limitation, reads length have to be increased to improve the certainty of the origin which is provided in the recent sequencers as in paired-end sequencing (Mardis, 2013). In this read pairs on basis of length covered can be obtained as paired end reads or mate pair reads.

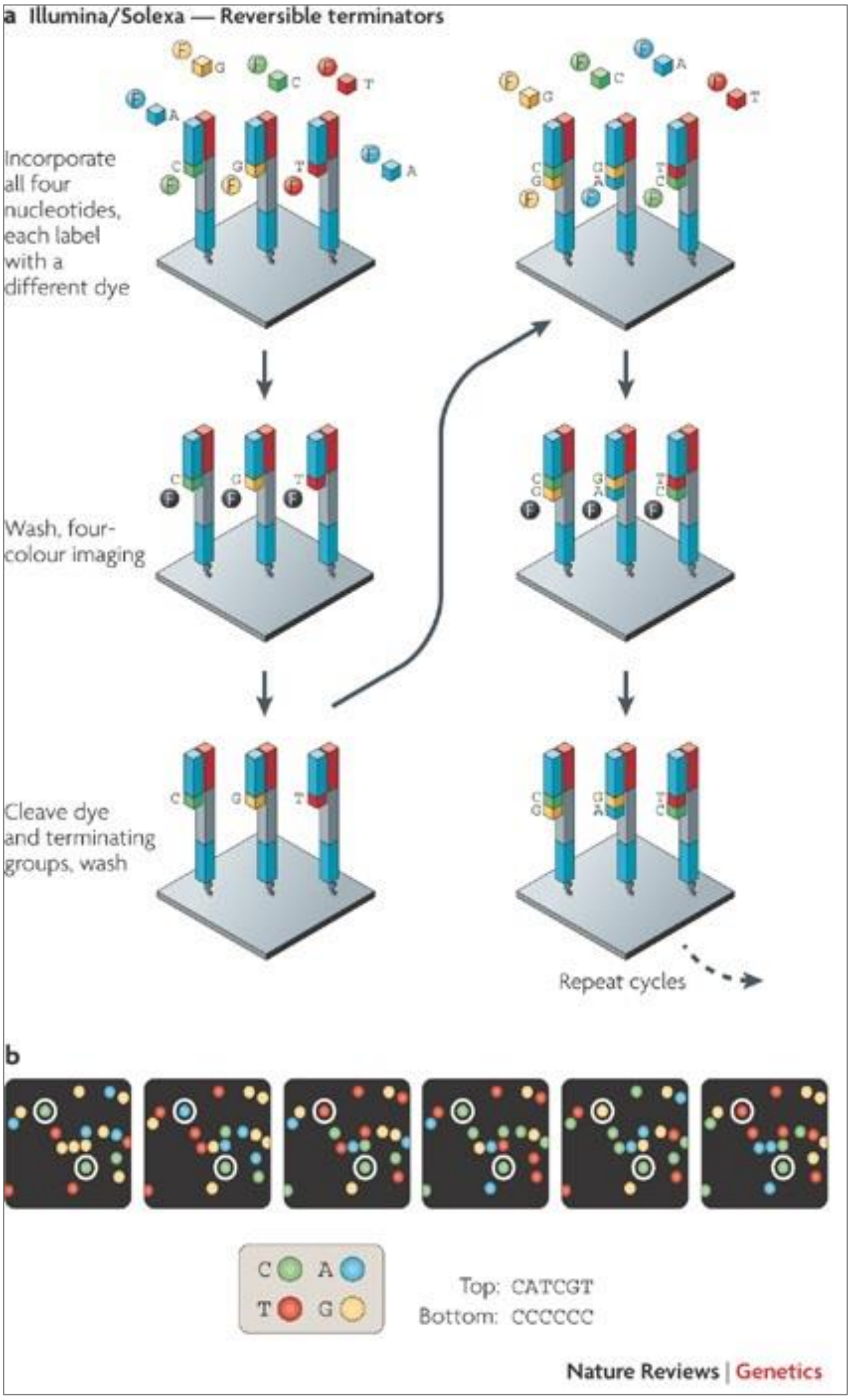

a Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
b
C A
T G
Top: CATCGT
Bottom: CCCCCC
Nature Reviews | Genetics

Figure 2.5: a. Cyclic reversible termination, the above described third step of sequencing in Illumina/Solexa GA.

b. Imaging output from sequencing data. Each colored dot represent a single cluster (Adapted: Metzker, 2010).

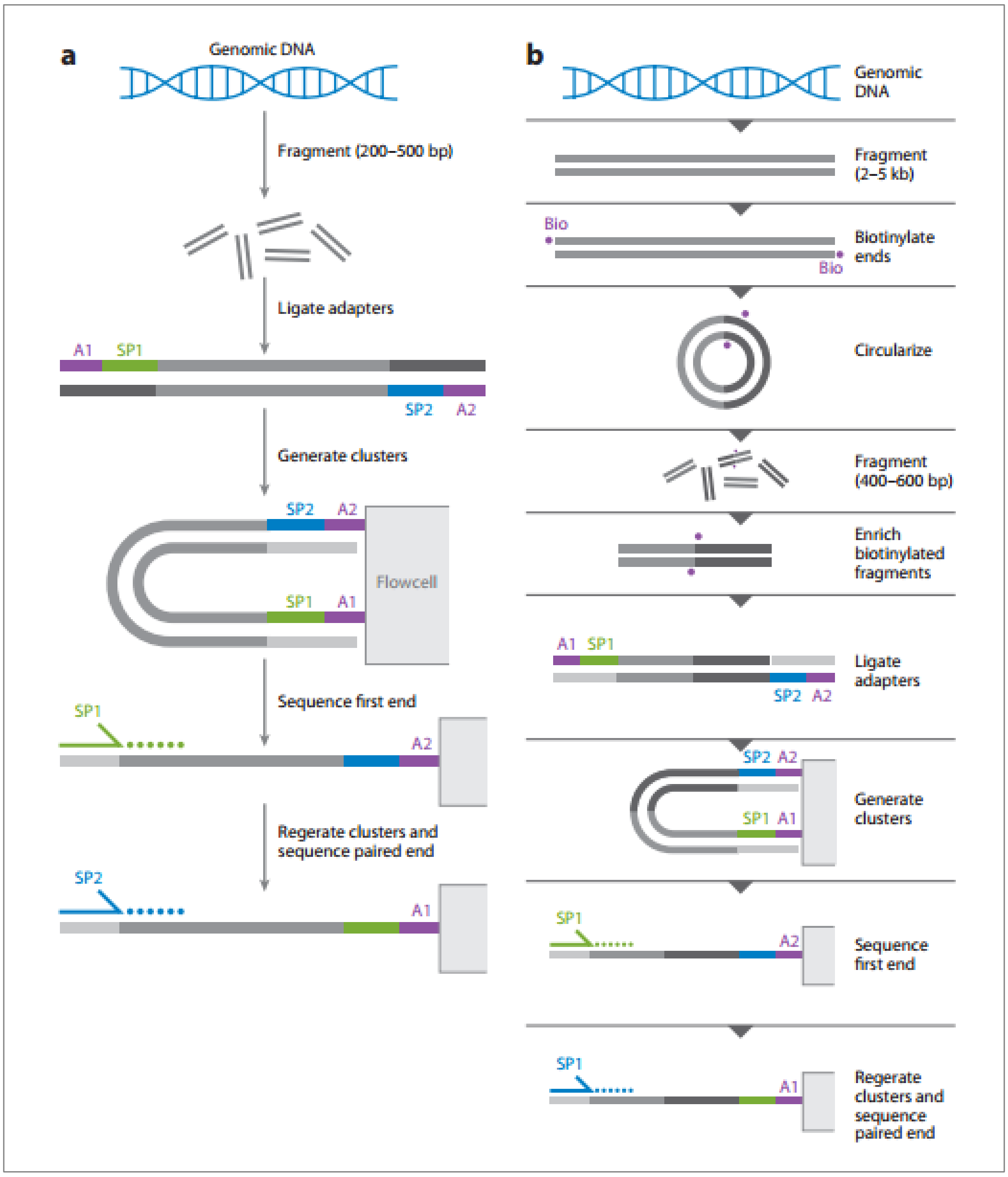


Figure 2.6:
a. Paired end sequencing
b. Mate pare sequencing.
(Adapted: Mardis, 2013)

To obtain reads from both ends of the fragments for more coverage a linear fragment <1kb is used to which two different adapters are covalently ligated having different primer sites. The sequencing is designed in such a way where fragment ends primed only at one end are extended to clusters with numerous cycles but are removed so that the extension of fragments with different set of adapters primed can take place which is also the opposite end. In mate pair sequencing the sequencing procedure is almost the same but initially fragments to be sequenced undergoes a series of chemical and biological steps to get library of enriched fragments which contain both ends of the initial fragment joined by a central adapter (Mardis, 2013). The fragment length is more than 1kb in this case. The combination of the above two approaches can provide coverage up to 20 kb specially in obtaining long range assembly in difficult regions of the genome (Gnerre *et al.*, 2011).

After sequenced reads are generated and filtered, they are ready for the next step in variant analysis of aligning to the reference genome. There are two repositories for human reference genome assemblies namely UCSC (hg) and GRC (GRch) which are identical (Genome bioinformatics group) but have different in nomenclature. UCSC provide hg18 and hg19 versions of human genome assembly while GRCh36 and CRCh37 are provided by GRC (Pabinger *et al.*, 2013).

2.4.2 Genome alignment: BOWTIE 2

Sequencing machines results in huge amounts of data in terms of number of reads (short DNA sequences) generated. As reads are short and to reveal from which part of the genome they are generated alignment to the reference genome is required. As human genome is huge the computational costs for alignment millions of reads is also high. Studies like Ley *et al.*, 2008 where number of reads generated reach in billions and short read alignment tools like Maq (Li H., Ruan *et al.*, 2008) and SOAP (Li, R. *et al.*, 2008) will take 5 CPU-months and 3 CPU-years align them (Langmead *et al.*, 2009). To decrease the computational time number of CPUs has to be increase which results in high computing cost.

To overcome above outlined shortcomings BOWTIE (Langmead *et al.*, 2009) was developed which is ultrafast, memory efficient as compared with Maq and SOAP. It uses a novel indexing strategy on reference genome uses burrows-wheeler transform (BWT) (M. Burrows, n.d.) based on full-text minute-space (FM) index (Ferragina & Manzini, 2000). Initially developed for data compression BWT-based indexing when applied allows large reference genome to be searched efficiently.

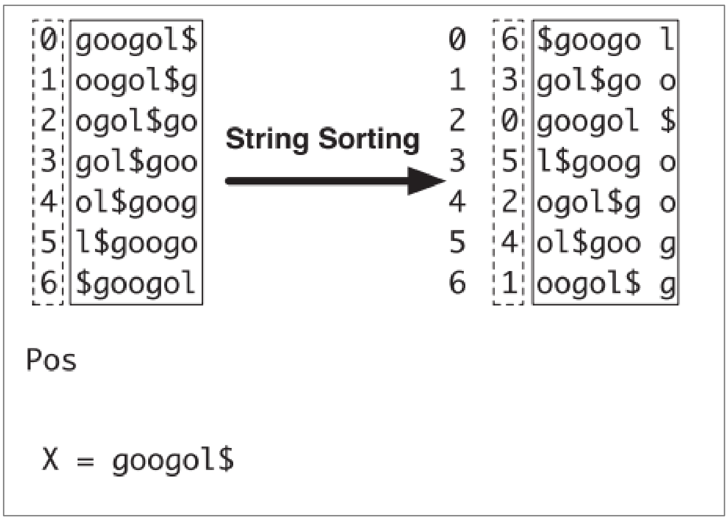
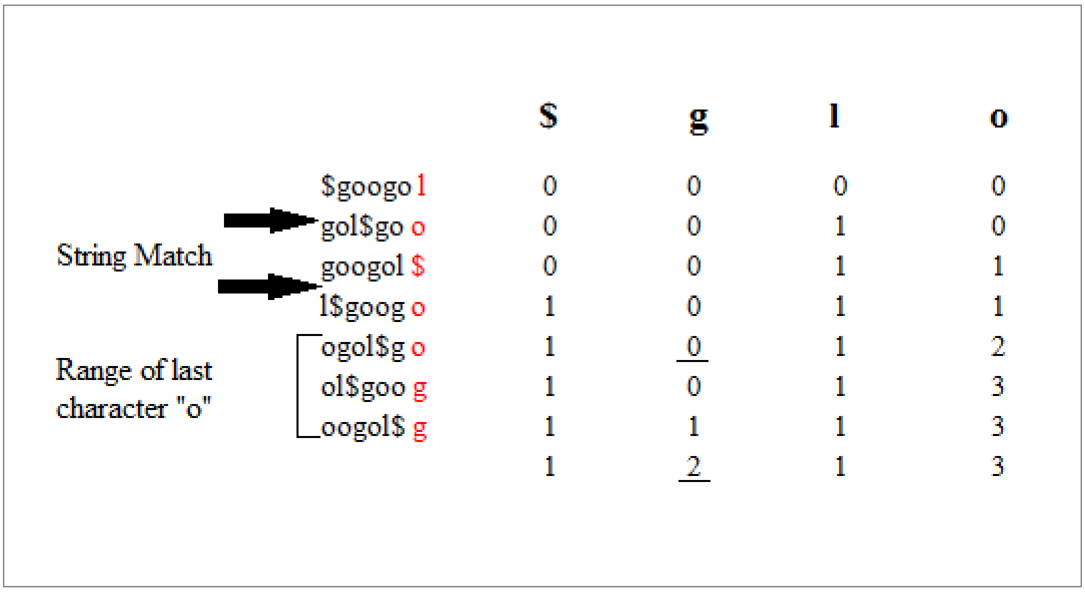


Figure 2.7 BWT of text X= googol$ (Adapted: H. Li & Durbin, 2009)

As seen in the figure BWT of text X is lo$oogg (last column of matrix resulted after string sorting), which is generated by lexicographically sorting rows according to the first characters of the matrix. The rows of matrix are made of all possible text generated by cyclic rotations of test X added character “$” at the end. “$” is not in the text and is considered lexico-graphically less than all the characters.

The text search by navigating the index is possible due the ‘last first (LF) mapping’ property of the matrix. As stated by Langmead *et al.*, 2009 the LF property is that “the ith occurrence of letter X in the last column corresponds to the ith occurrence of X in the first column.” This property is used for exact matching of substring and can also be used to un-permute BWT to recover original text.



	$	g	l	o
$googo l	0	0	0	0
gol$go o	0	0	1	0
googol $	0	0	1	1
l$goog o	1	0	1	1
ogol$g o	1	0	1	2
ol$goo g	1	0	1	3
oogol$ g	1	1	1	3
	1	2	1	3

Figure 2.8 Exact matching of substring “go” using LF mapping property of the burrows-wheeler matrix. BWT(X=googol$) is in red. The values in the matrix are the number of times the character appears before that position in the last column (starting from top to down).

For example the substring “go” is searched in following steps in main text X. Step 1. The last character in the substring i.e. “o” is selected and the range of that character in the first column of the matrix is chosen in lexicographical order as shown in the figure.

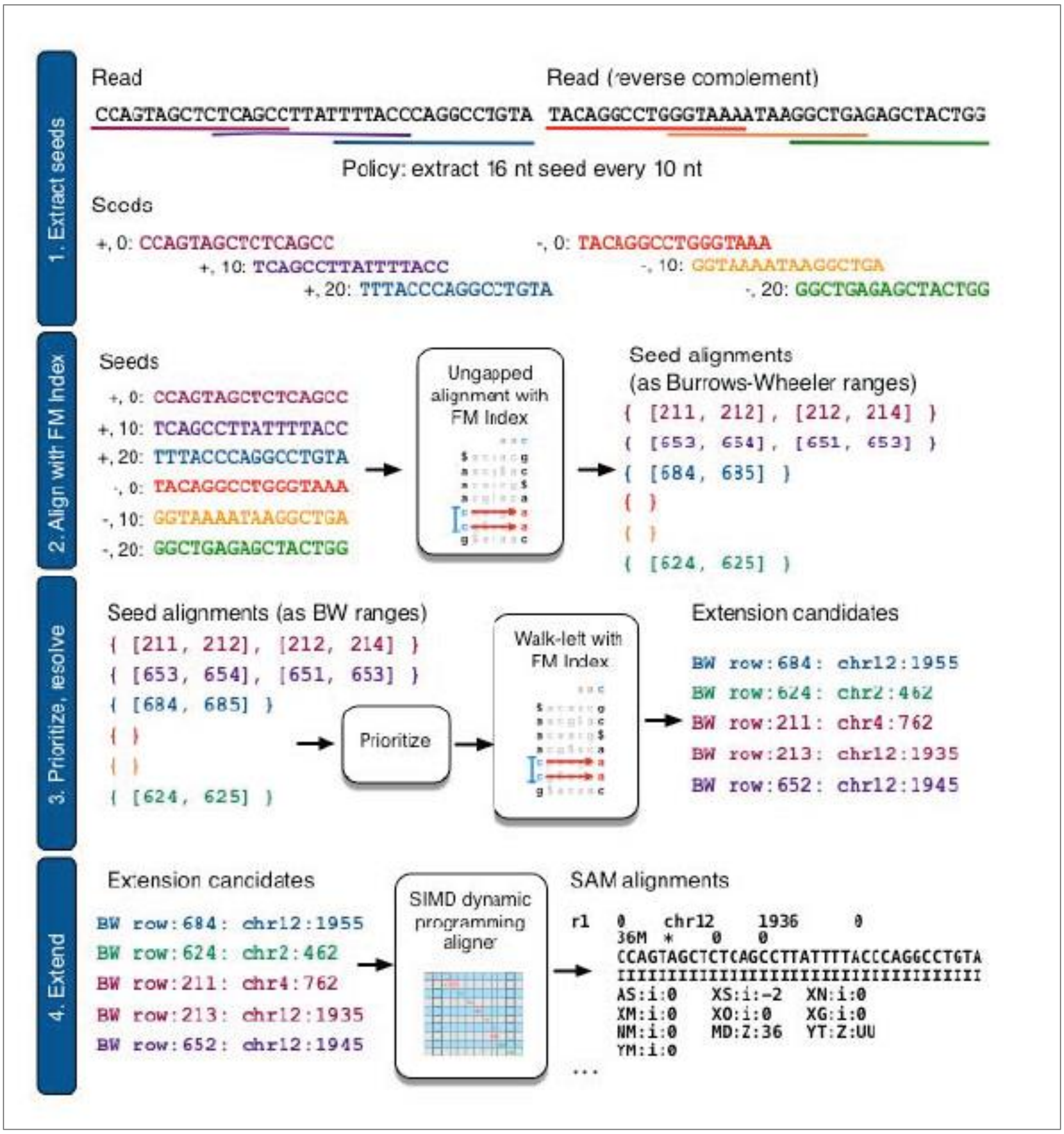
Step 2. Same range is selected in the last column.

Step 3. As “g” is the next pattern character, the values of that character in the matrix corresponding to the first row and next to last row of range (selected in previous step) are selected from the LF matrix, which are shown as 0 and 2 respectively.

Step4. The first row value in the LF of the character “g” matrix corresponds to that row before which the character has appeared as many times as the value, which is equal to the value “0”. The second value which is “2” is used to select number of rows of that character (“g”) which gives the range of the rows in which string is matched as seen in the figure above.

The EXACTMATCH algorithm in bowtie works in the same fashion to navigate or search the text in the index. As this algorithm does not allow for mismatches or inexact matches which can be due to sequencing errors or genetic variations two extension were made in addition to it, a back tracking algorithm that allows high quality alignments with mismatches and other being double indexing strategy to avoid excessive backtracking (Langmead *et al.*, 2009).

The efficiency of index based alignment in BOWTIE comes to question when alignment are permitted to contain gaps which again can occur due to sequencing errors or insertion and deletions (Langmead & Salzberg, 2012). The number of gaps increases the search space thus increasing the search time. The inability of bowtie to overcome this BOWTIE 2 was developed which extends the FM index approach to allow gapped alignment (Langmead & Salzberg, 2012).



```
r1   0    chr12    1936     0
     36M  *    0    0
     CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
     AS:i:0    XS:i:-2   XN:i:0
     XM:i:0    XO:i:0    XG:i:0
     NM:i:0    MD:Z:36   YT:Z:UU
     YM:i:0
...
```

Figure 2.9: Workflow of alignment in bowtie 2. (Adapted: supplementary data Langmead & Salzberg, 2012)

As seen in the figure 2.9 BOWTIE2 proceeds for gapped alignment in two stages. First is the alignment of the seeds generated from the reads (forward and reverse both). The alignment proceeds in an un-gapped fashion using the space and memory efficiency of full text minute (FM) index. The alignment yields burrows-wheeler ranges and rows which are prioritized according to the ranges (rows from small range given high priority). The rows prioritized are then chosen randomly by BOWTIE2 and selected row's offset is resolved (or located) onto the reference genome using last first mapping property of FM index. First stage can be seen in first 3 steps of the above figure 2.9. In second stage BOWTIE 2 through dynamic programming perform gapped alignments accelerated by efficient single-instruction multiple-data (SIMD) parallel processing of contemporary processors in the flanking neighborhood of the prioritized and resolved seed hits. The process continues until all seed hits or alignments are examined, or up to the limit of dynamic programming (Langmead & Salzberg, 2012).

Fixing the deficiency of BOWTIE in BOWTIE 2 by including gapped alignments and improving speed and number of reads aligned by using dynamic programming, makes bowtie 2 an extremely accurate and sensitive alignment tool. The alignment draw parallels in terms of data output of number of reads generated after sequencing. Specific file formats have been assigned or developed for this purpose which are considered default or close to default when analyzing NGS data.

2.4.3 SAMtools and File formats (SAM/BAM)

There are numerous sequencing platforms available (Mardis, 2013) which produce data in the formats which may vary accordingly. Generally FASTQ format (Cock *et al.*, 2010) is most common file format amongst them. Again with different technologies there are different alignment tools for read mapping again reference genome which produce data in different format making the downstream process complicated, also it makes work cumbersome if the data is to be compared and utilized from different sources (Li, H. *et al.*, 2009). Li *et al.* devised a common format called as sequence alignment map (SAM) which supports various aligners for storing alignment information generated, can be converted from other alignment formats generated, compact and supports different read types like single- , paired- end and their combination which can be handled with SAMtools software package making downstream analysis streamlined.

SAM stands for sequence alignment map, as the term suggests sequence alignment can be stored and visualized easily. A generic SAM file which is tab delimited has an optional header section followed by the alignment of reads section. These read alignments are one per line and each has mandatory fields which are 11 in number.

QNAME: query name of the read or pairs of reads. (Required for all alignments)
FLAG: a bitwise flag for tagging mate pairs, paired reads, strand etc. (Required for all alignments)
RNAME: reference sequence name. Value "*" if read is not aligned.

POS: leftmost position of the first aligned base (clipped). 1-based i.e. the first base in a reference sequence has coordinate 1.

MAPQ: phred based mapping quality describing alignment uniqueness.

CIGAR: a set of characters describing pairwise alignment. (M=match/mismatch, I=Insertion, D=Deletion, N=bases skipped on the reference, S=soft clipping, H=hard clipping, P=Padding, “=” = sequence match, X=sequence mismatch).

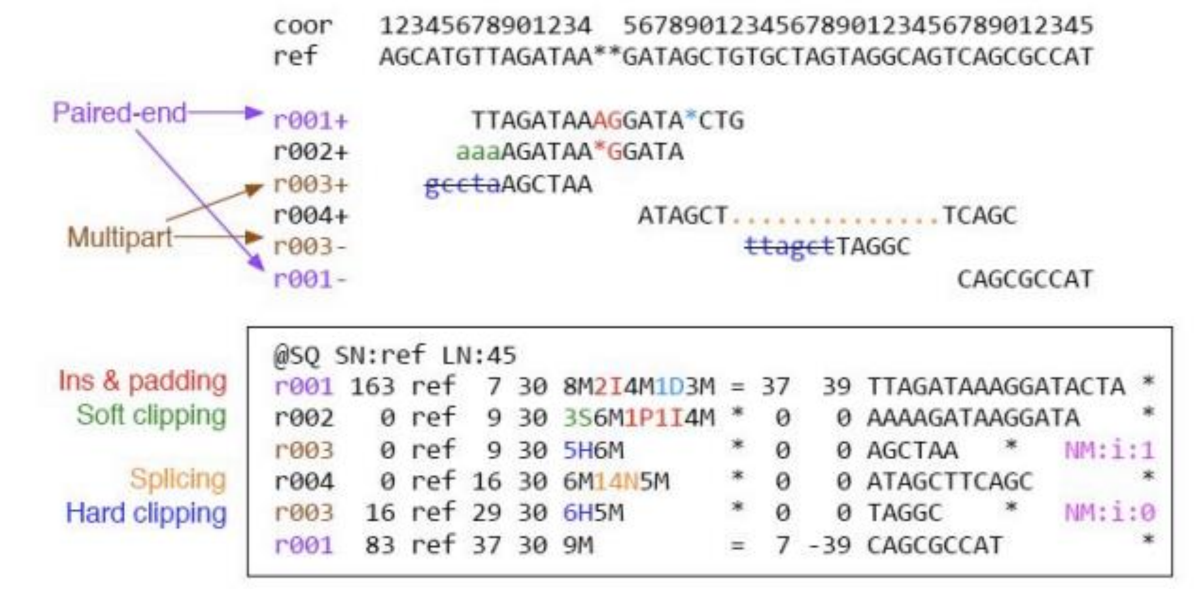
MPOS: leftmost mate position.

MRNM: mate reference name.

ISIZE: Inferred insert size.

SEQ: Segment sequence on the same strand as the reference.

QUAL: Query quality which is ASCII of phred base quality.



```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     gcctaAGCTAA
r004+                   ATAGCT..............TCAGC
r003-                          ttagctTAGGC
r001-                                        CAGCGCCAT
```

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Figure 2.10: SAM file format specifications, showing alignment of reads to the reference. (H. Li *et al.*, 2009)

SAM format can store different kinds of alignments which can be seen as different colored in the figure. Paired read alignments where reads have identical read pair name. In this mapped reads are stored in two or more alignment lines in the file. Other alignments can be clipped (hard and soft), spliced, multi-part, padded and alignment in color space (not shown in the figure 2.10). Each alignment then can be described with set of extended cigar strings. All reads mapped to the reference are represented on the forward genomic strand.

Parsing is slow in SAM, therefore there is a compressed binary format (BAM) which is generally used for intensive data processing. Both file formats can be converted into each other, performed I/O operations and can be manipulated via command line SAMtools. For example viewing, sorting, indexing and merging. SAM/BAM files can be sorted according to reference coordinate, query name or can be left unsorted. Mostly these can only be performed on BAM files sorted by left most coordinate. Sorting is necessary for stream based processing where alignment operations can be carried on in a stream without loading each alignment in the memory (H. Li *et al.*, 2009). Indexing is also necessary which provides a way of quick retrieval of alignments. One of the things to note here is the removal of duplicate reads is a part of the downstream process which can be handled with SAMtools. The generic nature of the format separating alignment step from the downstream process provides a modular approach to NGS data analysis where data can be handled accordingly to the users need and set of fixed steps can be comprised to make a pipeline for repeated analysis.

2.4.4 Variant Detection with VARSCAN 2

One of the main advantages of high-throughput sequencing is particularly to identify variants in the genome affecting human diseases (Koboldt *et al.*, 2009). When it comes to cancer studies where the disease is caused by mutations and variations, deep sequencing of normal-tumor pairs can make it feasible to detect rare variants, mutations and can provide their allele frequencies (Brockman *et al.*, 2008) in both individual and pooled samples. Although most of the tools for variant detection are dependent on single sequencing platform or aligners (Koboldt *et al.*, 2009) very few tools are available for analysis that are compatible with multiple data and aligner types. SAMtools (Li, H. *et al.*, 2009) and VARSCAN 2 (Koboldt *et al.*, 2012) fall in that category with some other tools which can be applied on generic alignment formats like SAM/BAM formats or "SAMtools pileup", but SAMtools is not a good option when comes in handling normal-tumor pairs as described by Pabinger *et al.*, 2013 in their survey of tools for somatic callers as it is only confined in detecting short INDELs and SNPs with its BCFtools set of commands.

Variant calling for short read alignments is a challenging task. Sophisticated analysis tools are required for obtaining true variant calls given the error prone data of sequenced short reads (Shigemizu *et al.*, 2013). High quality of variants is again a critical issue because of the false positive rates if we are considering whole genomes. In cancer studies generally normal-tumor pairs are studied for somatic variation as it gives a direct comparison at every position for accurate mutation detection. Though tumors not only contains somatic mutations but their genomes are heterogonous (Ding *et al.*, 2010), contains somatically acquired rearrangements (Campbell *et al.*, 2008) and copy number alterations (Beroukhim *et al.*, 2010) making variant calling even more challenging. Platform independent VARSCAN2 here provides a sensitive and specific way of detecting germline variants in addition to somatic mutations, loss of heterozygosity (LOH) and somatic CNVs in normal-tumor samples.

Somatic point mutations have been implicated in oncogenesis long before (Reddy *et al.*, 1982) and are observed in all the cancer genomes (Stratton *et al.*, 2009), and some of those which are termed as driver mutations are linked to carcinogenesis and tumor progression (Pleasance *et al.*, 2010). High quality point mutation detection can be performed by VARSCAN2. As Koboldt *et al.*, 2012 describe, the algorithm reads SAMtools pileup or mpileup files for both normal-tumor samples and make pairwise comparison of base calls and normalized sequence depth at each position. By default VARSCAN 2 requires 3x minimum coverage, with base quality of 20 (phred based) (minimum 3 reads with base quality >=20), allele frequency > 8% and P-value < 0.05. Based on these assumptions the algorithm adopts a heuristic approach for variant detection by determining normal and tumor genotype independently. VARSCAN2 categorizes variants called as germline, somatic and LOH. Mutation detection at each position is performed in several steps. First, minimum coverage requirement is checked for both the samples and genotype is determined based on the number of reads observed. If criteria are not met by the variant allele position is referred wild type (homozygous reference) and most supported variant allele is chosen if multiple variant alleles are observed. Direct comparison between normal and tumor is performed if one or both samples have a variant at the position.

When genotypes do not match their read counts are computed by one-tailed "Fisher's exact test" (Graham, 1992) in a two by two table as shown below, by comparing number of reference-supporting reads and variant supporting reads observed in tumor with those of observed in normal. Fisher's exact test (FET) of independence is used in cases where we have 2 nominal variables for each column (2 rows and 2 columns), like in our case with tumor and normal reads in reference and variant supporting categories. The null hypothesis is the relative proportions of the one variables is independent to other i.e. ratio of reference supporting reads to variant supporting reads is same for tumor and normal reads. FET uses hyper-geometric distribution to calculate the probability of the observed data and also the all data sets with extreme deviations under the null hypothesis. Variant is called somatic if *P*-value threshold is met (0.10 by default) and normal matches the reference but

		Reference supporting	Variant supporting
A.	Tumor Reads	tumor_reads1	tumor_reads2
	Normal Reads	normal_reads1	normal_reads2
B.	Observed Reads	total_reads1	total_reads2
	Expected Reads	error_free_reads	reads_with_error

Table 2.1: 2x2 tables A and B for FET.

LOH if the normal is heterozygous. Variant is termed germline if the significant threshold is not met by a different process using again one sided FET with category B 2x2 table as seen in the table. VARSCAN2 provides further filtering of somatic variants into high confidence (HC) and low confidence (LC) with the command *process somatic*, where HC somatic mutations should have variant allele frequency at least 0.1 in tumor sample and less than 0.05 in normal with statistical significance < 0.07 (Koboldt *et al.*, 2012).

Copy number alterations can also be detected by VARSCAN 2 which applies a different algorithm. Heuristic genotypic and computing statistical significance by FET are carried out simultaneously to evaluate normal and tumor samples which make VARSCAN 2 to detect the smallest of significant differences between them by exploiting the digital nature of NGS data. Though the performance of VARSCAN2 is robust as stated by Koboldt *et al.*, 2012 but the specificity and sensitivity of the algorithm applied by VARSCAN2 depends on the accuracy of the alignments i.e. performance of the sequence aligner.

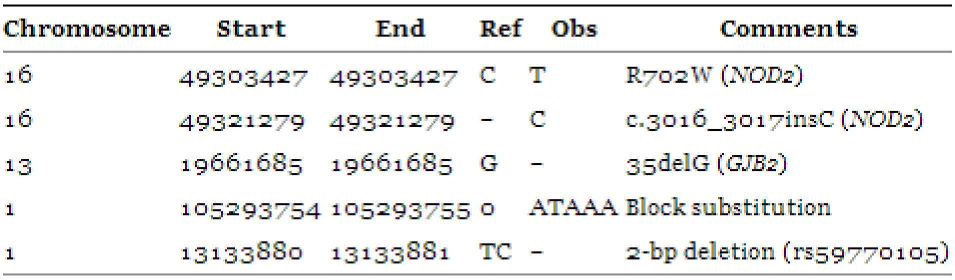
2.4.5 Variant annotation: ANNOVAR

With large number of experiments being conducted generating huge amounts data, many number of databases have been made or exists which store that data and are some of them are freely available. For example with 1000 genomes project (Abecasis *et al.*, 2012), not only functional aspects of genome variation have been categorized but it has also led to advent of many sophisticated tools and algorithms for data analysis and data mining. Variant identification is not the final step in revealing the cause for disease like cancer, the prediction of their functional aspects is also crucial. Several databases are in existence which serve this purpose of annotation such as dbSNP for SNPs, refseq for nucleotides, COSMIC for mutations in cancer and many other similar databases centered around human genomic variation (Küntzer *et al.*, 2010). As data generated by NGS is huge the need for automated annotation (prediction) is important. The annotation process then helps to filter out the variants and prioritize them for further analysis (Pabinger *et al.*, 2013).

There are very less number of methods available that can detect large number of variants (whole genome) and annotate them simultaneously. Usually sequencers are provided with their own functional annotation software which is specific to them only. This sequencing platform-specific nature does not fulfill the needs of the user (Wang *et al.,* 2010). There are many computer aided annotation tools have been developed, their annotation methods may differ but most of them provide links to the existing databases for annotation. One of such tools is ANNOVAR.

ANNOVAR (Wang *et al.*, 2010) which stands for annotate variation is a tool developed for up to date functional annotation of human genome and other genomes from high throughput sequencing data so as to examine the functional effect on the genes and to identify variants such as reported in 1000 genomes project or dbSNP. INDEL, block substitution plus CNVs can also be annotated through it. It is a command line tool which uses external databases (pre-compiled annotated databases) that can be downloaded locally and used for annotation. Different methods for annotations are provided in it which are categorized as gene-based, region-based and filter-based (Pabinger *et al.*, 2013).

ANNOVAR has additional features other than annotation of variants with respect to gene one of which is the capability to compare variants to existing. It does it by its feature of filter based annotation where it can evaluate and filter out subsets of variations which are not reported in filter based database like 1000 genomes (SNPs with >1% frequency) or non-synonymous SNPs with SIFT scores >0.05 and dbSNP. In other sense only those variants will be identified whose exact chromosome positions, start and end, and observed alleles are matched. It is helpful in cases where variations are assumed to be mutations. In gene based annotation it can identify variants in the disrupted genes or nearby them as reported in whole genome experiments. These are intronic, exonic, spliced, ncRNA, 3UTR, 5UTR, upstream or downstream TSS (1kb) and intergenic vairants whose distance with flanking gene is reported. Thus they can help identify variants in nearby promoter or enhancers though it doesn't exactly notify variants in areas as functional element of genome. With region based annotation variants in conserved regions or predicted TFBS can be annotated with many other annotation tracks (K. Wang *et al.*, 2010).



Chromosome	Start	End	Ref	Obs	Comments
16	49303427	49303427	C	T	R702W (*NOD2*)
16	49321279	49321279	-	C	c.3016_3017insC (*NOD2*)
13	19661685	19661685	G	-	35delG (*GJB2*)
1	105293754	105293755	0	ATAAA	Block substitution
1	13133880	13133881	TC	-	2-bp deletion (rs59770105)

Figure 2.11 Example of the content of an ANNOVAR input file. (Adapted: K. Wang *et al.*, 2010)

ANNOVAR takes a text based file as an input where each line corresponds to a variant which has 5 space or tab delimited columns (chromosome, start position, end position, reference nucleotide(s) and alternate nucleotide(s)). ANNOVAR package comes with a set of PERL scripts to automate the procedure which are used in the annotation pipelines, such as conversion of input file to

ANNOVAR readable file, downloading the relevant database which is important and the annotation script to perform annotations on selected databases by scanning them (K. Wang *et al.*, 2010).

2.4.6 Peaks Annotation: HOMER

HOMER (Hyper-geometric Optimization of Motif EnRichment) (Heinz *et al.*, 2010) is a command line tool for motif discovery and analyzing data from NGS. Apart from *de novo* motif discovery algorithm and many tools for functional genomic sequencing analysis like ChIP-seq, RNA-seq, DNase-seq and other data sets, it can also be used to for annotating peaks predicted in ChIP-seq or DNase-seq experiments. Annotating peaks can help in associating peaks with nearby genes. In variant analysis to identify variants in regulatory regions (peaks from ChIP-seq or DNase-seq) annotating peaks can expand the information base, for example accurate distance of peaks from TSS or TTS can be predicted. The regions annotated where peak falls include regions like TSS (by default defined from -1kb to +100bp), TTS (by default defined from -100 bp to +1kb), CDS Exons, 5' UTR Exons, 3' UTR Exons, CpG Islands, Repeats, Introns and Intergenic. CpG Islands and Repeats are the part of detailed annotation (Homer Software and Data Download).

3. AIMS

In addition to the achieve research goals the main aims of the study were:

- To understand PCa genomics and develop a pipeline for variant analysis.
- To detect, annotate somatic mutations in non-coding regions of PCa genome and find their location if they fall in regulatory regions.
- To investigate significance of the genes related to mutations in PCa or other cancers.

4. Methods
4.1 Data collection

Data obtained for variant analysis was high quality matched normal-tumor complete genomes in processed BAM files (Single file each normal and tumor sample) from PCa study done by Berger *et al.*, 2011 “The genomic complexity of primary human PCa”. These were DNA samples from the blood and the tumors of human male participants (7 samples in total) sequenced to approximately 30x haploid coverage on an Illumina GA II sequencer which were put to the processing pipeline and aligned to the reference(hg18, UCSC genome assembly) (Berger *et al.*, 2011).

Sample	SRA Identifier	Age	PathologicalStage	FusionStatus	TumorPurity
(S1)PR-0508	Normal-SRS165685 Tumor SRS165686	57	T2c	Negative	73 %
(S2)PR-0581	Normal-SRS165687 Tumor SRS165688	69	T3b	*TMPRSS2-ERG*	60 %
(S3)PR-1701	Normal-SRS165689 Tumor SRS165690	62	T3a	*TMPRSS2-ERG*	49 %
(S4)PR-1783	Normal-SRS165691 Tumor SRS165692	66	T2c	Negative	75 %
(S5)PR-2832	Normal-SRS165693 Tumor SRS165694	66	T2c	*TMPRSS2-ERG*	59 %
(S6)PR-3027	Normal-SRS165695 Tumor SRS165696	66	T3b	Negative	74 %
(S7)PR-3043	Normal-SRS165697 Tumor SRS165698	69	T2c	Negative	68 %

Table 4.1: Clinical Characteristics of 7 PCa Genomes. (Berger *et al.*, 2011)

To determine if variants are located in regulatory regions, open chromatin data by DNaseI HS from ENCODE/OpenChrom(Duke University) was used. Open chromatin regions are highly sensitive to DNaseI enzyme and its activity is used to map general chromatin accessibility and DNaseI "hypersensitivity" is a feature of active *cis*-regulatory sequences. Functional regulatory elements or Peaks obtained (DNaseI hypersensitive sites) in data were isolated using methods called DNase-seq or DNase-chip (Song and Crawford, 2010; Boyle et al., 2008; Crawford et al., 2006). Sample cells for cell lines (LNCaP) were digested with DNaseI and then DNase digested ends are captured which was sequenced using Illumina (Solexa) sequencing. After additional verification step reads were aligned to reference genome using BWA (Li *et al.*, 2009) for the GRCh37 (hg19) assembly. The tracks thus obtained represent DNaseI sensitivity as a continuous function using sequencing tag density (Raw Signal) and discrete loci of DNaseI sensitive zones (HotSpots) and hypersensitive sites (Peaks). “Peaks” are the regions of enriched signal in the DNase HS experiment. Peaks were called based on signals created using F-Seq (Boyle *et al.*, 2008b). ENCODE consortium with Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) and ChIP-seq experiments have collected locations of active regulatory elements identified as open chromatin regions for many cell types which were used to validate the assay delineating areas of open chromatin. The format of the peak data or tracks can be seen in the table below.

field	example	description
bin	589	Indexing field to speed chromosome range queries.
chrom	chr1	Reference sequence chromosome or scaffold
chromStart	526518	Start position in chromosome
chromEnd	526780	End position in chromosome
name	chr1.1	Name given to a region (preferably unique).
score	593	Indicates how dark the peak will be displayed in the browser (0-1000)
strand	.	+ or - or . for unknown
signalValue	0.0555	Measurement of average enrichment for the region
pValue	2.23	Statistical significance of signal value (-log10). Set to -1 if not used.
qValue	-1	Statistical significance with multiple-test correction applied (FDR -log10). Set to -1 if not used.
peak	126	Point-source called for this peak; 0-based offset from chromStart. Set to -1 if no point-source called.

Table 4.2: Sample Schema for Duke DNaseI HS - Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University). In BED format.

Using UCSC table browser uniform tracks for DNaseI HS for LNCaP cell lines in the form of two files was downloaded. The data was for LNCaP cell line (Horoszewicz *et al.*, 1983) with and without androgen induced which was then combined to form a single BED (UCSC BED format) file.

4.2 Data Analysis: Variant analysis by whole genome sequencing

A pipeline was setup using different tools and utilities on a pilot data for chromosome 21 which was then used to upscale analysis on all seven samples. A variant detection pipeline follows a typical workflow but tools may vary according to needs. After sequencing at a specific platform reads were assessed for quality and aligned, in the case of current study reads were obtained pre-aligned where reads were already processed with manufacturer's (sequencing platform) software and through PICARD pipeline a single BAM file for each sample was produced (Berger *et al.*, 2011). Our pipeline was imposed on these BAM files where steps involved are reads filtering and extraction with BamUtil, realignment to reference (hg19) with BOWTIE2, conversion from BAM files to pileups for variant detection through SAMtools, variant detection with VARSCAN2, variant annotation with ANNOVAR and finally determining variants in regulatory region using DNaseI HS data with annotating peaks hit with HOMER for further accuracy. In addition peaks (Overlapping variants) were further sought for predicted TFBS using ENCODE ChIP-seq data (TFBS Track, ENCODE).

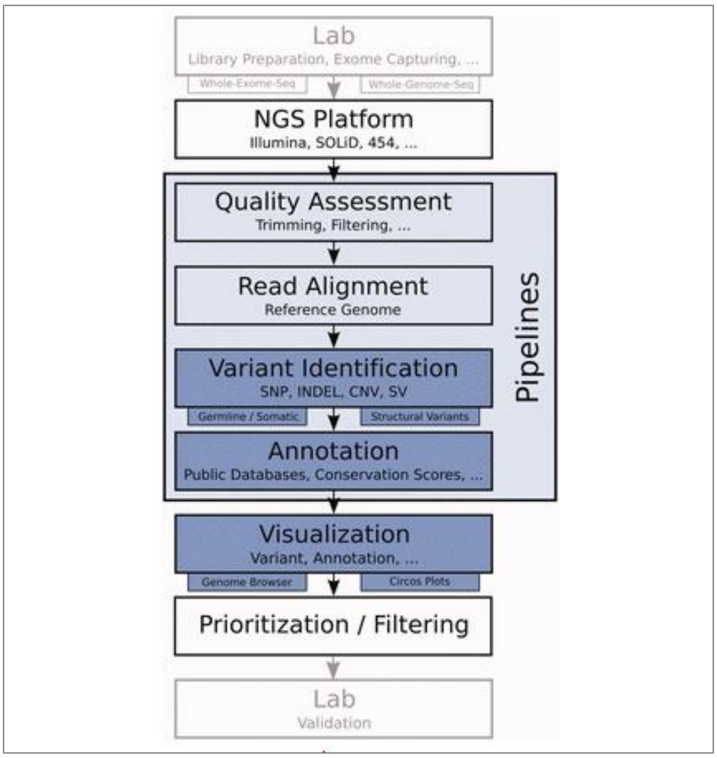


Figure 4.1: Basic steps in whole genome sequencing studies. (Adapted: Pabinger *et al.*, 2013)

4.2.1 Reads extraction from BAM file

As sequenced reads in BAM file obtained were aligned to hg18 (UCSC genome assembly) the reads first have to be extracted before realignment to the newer version hg19. BamUtil, a repository containing several programs to perform operations on SAM/BAM file, was used for extraction of reads from pre-aligned BAM files. Reads were extracted via bam2Fastq utility of BamUtil as both paired and unpaired reads (single end) also with unmapped reads were extracted in two different files. Before extracting, reads were filtered to remove low quality and duplicated reads by using dedup utility of BamUtil with phred based quality score of 20 (q20). BamUtil was preferred over SAMtools and PICARD for extracting reads due to its feature of producing mapped paired and unpaired reads separately in single run, also the ability to handle mates on different chromosomes which Picard specially does not handle (BamUtil, Abecasis Group).

4.2.2 Reads re-alignment

Filtered reads paired and unpaired (single end plus unmapped) in FASTQ format (Cock *et al.*, 2010) were aligned to newer reference genome build hg19 using BOWTIE2. Alignment was done separately for the two different files i.e. paired and unpaired reads. Default parameters were used and resulting output was piped to SAMtools (Li, H. *et al.*, 2009) to get a sorted BAM file each for paired and unpaired reads.

4.2.3 Sequence alignment and file operations

SAM/BAM format is most common file formats for downstream sequencing analysis. Most of the file processing and operation were carried out by program SAMtools (Li, H. *et al.*, 2009). SAMtool view-, sort- and index- option were used for conversion for SAM/BAM files, sorting and indexing respectively. To get alignment statistics option flagstat- was used on BAM files. In end to get normal and tumor files per sample as an input for variant detection SAMtool merge- option was utilized to merge the sorted paired and unpaired BAM files resulted after alignement BOWTIE2 which was then piped to another SAMtool option mpileup- to get normal and tumor pileups ready for variant detection.

4.2.4 Variant Detection

To detect somatic variation between obtained normal and tumor samples, VARSCAN2 program was utilized (Koboldt *et al.*, 2012). The program has an outstanding feature of producing variants with high confidence. After getting normal tumor pileups, each sample pair in pileups was fed to VARSCAN 2 somatic- command to get variant output in different files according to somatic status to report germline, somatic, and LOH events at positions where both normal and tumor samples have sufficient coverage (default is 8) as stated in VarScan user’s manual. Options --strand-filer and --output-vcf were set to 1 to remove variants with strand bias > 90% and variant file in vcf format which is a general format to store gene sequence variations. Other options were left to default. To remove clusters of false positives and SNV calls near indels somaticFilter- command was used with -min-avg-qual option 30 (phred base minimum average base quality for variant supporting reads) on somatic mutation file (VarScan User's Manual). Finally isolating somatic mutations by type and confidence the command processSomatic was utilized to yield low and high confidence mutations ready for annotation (VarScan: variant detection in massively parallel sequencing data).

4.2.5 Variant Annotation

Variants were annotated based using program ANNOVAR (Wang et. al. 2010). Files first were converted to ANNOVAR readable files by convert2annovar.pl perl script in ANNOVAR package. Finally gene based and filer based annotations were simultaneously carried out using table_annovar.pl script. The database utilized were refGene, 1000g2012apr_all, snp137, cosmic65 and esp6500_all, only refGene being the only gene base and rest filter based. Databases where initially downloaded using annotate_variation.pl and -downdb option for hg19 build.

4.2.6 Peak detection and mutation filtering

Peaks or DNaseI HS sites obtained as tracks from Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University) for LNCaP cell lines with both androgen induced and without it were combined to get a single peak file. Peaks or chromosome intervals were searched if mutations fall in them. Mutation filtering according to their annotation and all other operations for producing relevant results were performed with R scripts.

4.2.7 Peaks Annotation with HOMER

Finally after getting hits for mutations falling in peaks, peaks were annotated using HOMER software's annotatePeaks.pl perl script for better accuracy of finding promoter or enhancer regions of candidate genes.

The candidate genes determined with mutations in their promoter and enhancer regions were then searched extensively in published literature and selected relevantly as final results of the study.

4.2.8 Identifying TFBS

TFBS clusters (V3) from ENCODE ChIP-seq experiments for 161 transcription factors were downloaded to identify variants detected (in overlapping peaks) in previous steps which are also TFBSs (TFBS Track, ENCODE).

5. Results

Berger *et al.*, 2011 have reported in their study mutations in the gene *SPOP* and *SPTA1* in 2/7 samples. In accordance with the finding we have also detected mutations in 2/7 samples for *SPOP* and *SPTA1* (only 1 sample for *SPTA1*) delineating the pipeline and variant analysis of the current study to be appropriate.

Sample	chromosome	position	reference	alternate	gene	mutation type
S1(PR-0508)	chr17	47696426	A	C	SPOP	non-synonymous
S7(PR-3043)	chr17	47696643	A	C	SPOP	non-synonymous
S7(PR-3043)	chr1	158605757	C	T	SPTA1	non-synonymous

Table 5.1: Mutations detected in SPOP and SPTA1 in accordance with reported in Berger *et al.*, 2011.

After filtering the SNVs and their annotation a total of 54679 SNVs were detected across 7 samples with mean of 7811 SNVs. With respect to refSeq gene annotations following are the described mutational landscape across 7 samples.

refSeq Annotations	S1 (PR0508)	S2 (PR0581)	S3 (PR1701)	S4 (PR1783)	s5 (PR2832)	s6 (PR3027)	s7 (PR3043)
downstream	57	51	44	44	48	39	27
exonic	36	85	49	44	30	51	30
intergenic	5550	8881	4725	5319	5059	4315	5051
intronic	1635	2222	1612	2177	1805	1562	1286
ncRNA-exonic	24	47	26	15	20	26	13
ncRNA-intronic	323	485	301	268	245	254	256
upstream	26	68	43	37	41	30	21
upstream/ downstream	3	0	1	1	1	0	0
ncRNA_UTR3	0	0	0	1	1	0	0
splicing	0	1	0	0	0	1	0
UTR3	35	57	35	31	33	27	25
UTR5	2	7	3	2	6	1	2

Table 5.2: Mutational Landscape across 7 samples. Mutations annotated both "upstream/downstream" were flanked by two genes whereas upstream and downstream regions are within 1 kb region from TSS and TTS respectively and mutation annotated "splicing" is variant that is within 2-bp away from an exon/intron boundary by default. (ANNOVAR, refSeq output).

5.1 Mutations in Promoters/Near Promoter regions

A total of 31SNVs were detected and filtered out in promoter regions (regions overlapping DNaseI HS sites within 1000 bp upstream of TSS or TTS) across all samples. SNVs were filtered according to annotations done with ANNOVAR and HOMER which also includes mutations in introns, UTR, and TTS.

A table describing SNVs near promoter regions overlapping DNaseI HSs is shown below where distance to TSS is negative for upstream SNV and positive for downstream. Only 4 out of them were found relevant in literature as most of them are related to pseudogenes and RNAs.

	Chromosome	**Distance to TSS**	**Gene Name**	**Gene Type**	**Position**	**ref**	**alt**	**TFs**
S1	*chr1*	*-695*	*ALDH4A1*	*protein-coding*	*19229941*	*G*	*C*	*TCF7L2* *HDAC8* *GATA2* *GATA3* *FOXA1*
	chr9	-73	LOC100132352	pseudo	68726487	G	A	
S2	chr3	-132	LOC401074	miscRNA	75721264	C	T	
	chr3	-132	LOC401074	miscRNA	75721270	C	A	
	chr3	-132	LOC401074	miscRNA	75721366	C	T	
	chr5	-911	MIR4461	miscRNA	134262767	T	C	
	chr5	-911	MIR4461	miscRNA	134262778	C	T	
	chr7	-144	*PRSS1*	protein-coding	142457244	T	C	
	chr9	-585	LOC642236	pseudo	68454897	C	G	
	chr9	-585	LOC642236	pseudo	68454907	G	A	
	chr11	-74	*RAB30*	protein-coding	82782850	C	T	
	chr17	27	*ZNF286A*	protein-coding	15602950	G	T	
S3	chr3	-201	FLJ20518	pseudo	75713238	C	A	
	chr3	-201	FLJ20518	pseudo	75713241	A	G	
	chr3	-201	FLJ20518	pseudo	75713264	G	T	
	chr5	4	*GFM2*	protein-coding	74062991	C	A	
	chr9	-73	LOC100132352	pseudo	68726463	C	G	
	chr12	-904	*RPAP3*	protein-coding	48100700	G	C	
	chr17	96	LOC100130581	miscRNA	41466073	T	C	
S4	*chr1*	*465*	*CREB3L4*	*protein-coding*	*153941196*	*T*	*G*	*MYC*
	chr17	-438	*KCNJ18*	protein-coding	21307973	G	A	
	chr2	-399	*TIA1*	protein-coding	70476210	C	A	
	chr2	-38	ANKRD30BL	pseudo	133015653	G	A	
	chr8	-132	*RP1*	protein-coding	55528477	G	A	
S5	chr2	282	*ANKRD30BL*	pseudo	133015252	G	A	
	chr3	-245	*ZNF717*	protein-coding	75834451	C	G	
	chr8	40	*SLA*	protein-coding	134115205	C	G	
	chr9	-333	LINC00094	miscRNA	136890317	C	A	

S6	chr11	-352	*MTRNR2L8*	protein-coding	10531001	A	G	
	chr17	*-681*	*ACACA*	*protein-coding*	*35716787*	*C*	*G*	*POLR2A* *MAX* *E2F1* *KDM5B* *E2F6* *REST* *CBX3*
S7	*chr2*	*682*	*CAD*	*protein-coding*	*27440974*	*G*	*A*	*POLR2A* *CTCF* *TBP* *MAX* *USF2*

Table 5.3: SNVs in promoters or near promoter regions overlapping with DNaseI HSs. Only 4 SNVs shown in red are found relevant in literature. In the table SNVs are described with Chromosome, distance to TSS, respective gene, its type, SNV position, reference nucleotide, alternate nucleotide and in the final column TFs found binding at that location

5.2 Distal SNVs

A total of 621 distal SNVs were found overlapping the DNase peaks across all the samples. These SNVs were detected both upstream and downstream within 1Mb distance from the respective genes. Only those SNVs were selected which are found relevant in literature and filtered according to annotations by ANNOVAR and HOMER, for example pseudogenes and non-protein coding RNAs were not considered. These were 21 in total across all samples.

	Chromosome	**Distance to TSS**	**Gene Name**	**Gene Type**	**Position**	**ref**	**alt**	**TFs**
S1	chr14	-22644	*GPR68*	protein-coding	91742928	C	A	EP300 ZNF217 MYC FOXA1 RXRA ESR1
	chr17	-10155	*NCOR1*	protein-coding	16108182	A	T	
S2	chr11	51557	*ARHGEF12*	protein-coding	120259203	T	C	ARIDA GATA3 FOXA1 FOXA2
S3	chr7	-11426	*NRCAM*	protein-coding	107892079	G	A	
S4	chr11	9627	*FOLH1*	protein-coding	49220622	G	A	
	chr13	66179	*CDK8*	protein-coding	26895001	C	A	
	chr21	34593	*BTG3*	protein-coding	18950621	G	T	
	chr21	16128	*SON*	protein-coding	34931410	T	G	ZKSCAN1

								GATA3
	chr6	76542	*MYB*	protein-coding	135578967	G	T	ARID3A CTCF SMC3 RAD21
	chr7	16421	*PDGFA*	protein-coding	543104	C	T	
	chr8	406490	*POTEA*	protein-coding	43554102	G	A	
	chr8	-17405	*MTSS1*	protein-coding	125758202	C	T	
S5	chr1	24929	*HSD3B1*	protein-coding	120074796	G	T	
	chr5	-328520	*IRX4*	protein-coding	2211334	T	C	
	chr10	112114	*MGMT*	protein-coding	131377630	C	G	E2F1
	chr11	6789	*TP53AIP1*	protein-coding	128800871	T	G	
	chr21	3744	*TMPRSS2*	protein-coding	42876172	C	G	
S6	chr17	21446	*IGF2BP1*	protein-coding	47096228	T	G	ZNF263
	chr5	-4123	*DOCK2*	protein-coding	169060126	T	G	
S7	chr17	-56982	*DLX4*	protein-coding	47989648	G	A	
	chr3	148694	*CHL1*	protein-coding	510059	A	T	CTCF ZNF143 SMC3 RAD21

Table 5.4: Relevant SNVs in Enhancer regions. In the table SNVs are described with Chromosome, distance to TSS, respective gene, its type, SNV position, reference nucleotide, alternate nucleotide and in the final column TFs found binding at that location.

6. Discussion

Somatic mutations in noncoding regions of the genome may relate to tumorigenesis. This hypothesis was supported by two separate studies by Horn *et al.*, 2013 and Huang *et al.*, 2013 as in their findings somatic mutations were reported to be occurring in TERT promoters in human melanoma elevating the transcriptional activity. In the current study conducted we have also tried to identify non-coding mutations in published WGS data and found mutations in promoter and enhancer regions of genes somehow relevant in PCa or related to genes significant in PCa. Most of the research work till now is or has been carried out in protein coding genome due to cost of WGS or because there is certain level of difficulty in interpreting non-coding mutations. But with large projects like ENCODE which has accumulated large amounts of data sets and annotated most of the non-coding genome which is freely available has made it easier to understand functions of elements in non-coding genome. Though this search in non-coding genome can be divided into many regions of interest like chromatin structures, conserved, epigenetic, regulatory etc. we have tried to cover the regulatory aspects of PCa genome. We believe findings of the study may prove significant in PCa. Mutations distances from the genes were carefully selected to categorize them (< 1kb for promoters and < 1MB for enhancers). We have also included the mutations in introns as DNaseI HSs were also detected engulfing them which implicate they may play role in regulation of the respective gene. Annotations of peaks containing mutations were done by HOMER as it annotate peaks with distance from TSS or TTS of the gene which is important in our case in categorizing and selecting relevant mutations. After carefully filtering SNVs, their genes were then looked in published literature which may implicate their role in PCa and as we report significant mutations have emerged.

6.1 Mutations in Promoters/near Promoter regions

A total of four genes relevant to our study were identified across four different samples whose promoters or downstream regulatory (intronic) regions were found to be mutated. *ALDH4A1* and *ACACA* were two genes having mutation in promoter region (in samples S1 and S6 respectively) whereas *CREB3L4* and *CAD* were mutated in their intron (in samples S4 and S7 respectively) which may act as a region of downstream regulation. *ALDH4A1* or *ALDH4* (aldehyde dehydrogenase 4 family, member A1) has been reported by researchers as a direct target of *p53* which prevents cell damage through transcriptional activation of *ALDH4* (Yoon *et al.,* 2004). Though binding site for p53 have been shown to occur at intron 1 of ALDH4, it is unclear how promoter mutation can affect ALDH4 functioning. Interestingly, HDAC8 and FOXA1 TFs were found to have binding site in that region of somatic mutation. HDAC8 inhibition has been shown to suppress both wild type and mutant *p53* transcription (W. Yan et al., 2013) citing an anti-cancer property of HDAC inhibitors. Various classes of HDACs were seen related to PCa progression (Abbas & Gupta, 2008) but specific role for HDAC8 is unclear. FOXA1 on other hand is well known factor and plays as crucial role PCa progression (Jin *et al.,* 2013) and (Imamura *et al.*, 2012). Another gene *ACACA* or *ACCA* (acetyl-CoA carboxylase alpha) was detected with a somatic mutation in its upstream promoter region. *ACCA* has been shown by researchers as a *BRCA1* associated protein which interacts specifically with *BRCA1* by its tandem of *BRCT* domain which is crucial for *BRCA1* function (Magnard *et al.*, 2002). *BRCA1* has been identified as tumor suppressor in PCa as reported by (Rosen *et al.*, 2001) and risk of PCa is higher for individual carrying mutated (germline) *BRCA1*. *ACCA* as fatty acid synthase (FAS) have been shown to have increased expression in PCa which is known to occur in early phase of PCa development (Swinnen *et al.*, 2002). *ACCA* knock down has also been reported by researchers in inhibiting LNCaP cell proliferation (Brusselmans *et al.*, 2005). One of the hypotheses that can be generated by the given facts this somatic mutation in *ACCA* promoter may play a significant role in overexpression of the respective gene in PCa or its deregulation resulting in loss of interaction with *BRCA1* which is known to be conserved in humans (Magnard *et al.*, 2002) but then role of *BRCA1* in PCa could be point of discussion.

Other somatic mutations which fall in DNaseI HS were in the introns of genes *CREB3L4* and *CAD* in sample S4 and S7 respectively. *CREB3L4* gene which encodes a CREB (cAMP responsive element binding) protein also known by its homolog *AIbZIP*, has been shown to be expressed at higher levels in PCa. As it is androgen regulated *AIbZIP* may have a significant role in PCa and AR signaling as stated by Qi *et al.*, 2002. Another somatic mutation of this type was seen in *CAD* 682bp downstream from TSS. *CAD* was identified to be one of the partners of AR to which it interacts in PCa and was shown to assist AR translocation in the nucleus and its transcription stimulation (Morin *et al.*, 2012).

Interestingly all the somatic mutations described above were seen in the samples which are *TPRSS2-ERG* fusion negative and two of them in same chromosome (chromosome 1, Table 4) but in different samples. For most of the somatic mutation TFs were identified related to the mutated sites but it is a difficult task to accurately predict the actual binding sequence of the TFs at that site. Therefore it is unclear how all these factors combine to characterize these promoter somatic mutations as driver mutations.

6.2 Mutations in Enhancer regions

In distal DNaseI HSs detected containing somatic mutations, out of 621 SNVs only 21 SNVs were selected on the basis of relevant literature which may prove significant in PCa. Out of 7 samples sample S4 (PR1783) was found to have most number of distal SNVs or enhancer SNVs. Enhancer SNVs were located both upstream and downstream of their respective gene (see Table 5). None of the samples were seen to have common mutation/s in the DNaseI HSs detected. This fact somehow slightly weakens the credibility of the results but the mutations and related genes the enhancer site cannot be neglected which have been seen related to PCa in literature. Though none of samples were found to harbor same mutations but chromosomes 11, 17 and 21 were found be mutated on maximum occasions (3 times) each across different samples, only chromosome 21 was mutated twice in S4 (PR1783).

Samples S2 (PR0581), S4 (PR1783) and S5 (PR2832) harbored mutations in chromosome 11 enhancers of genes *ARHGEF12*, *FOLH1* and *TP53AIP1* respectively. *ARHGEF12* has been reported by Robbins *et al.*, 2011 harboring a candidate somatic mutation in their study of metastatic prostate tumors which among other candidate mutations reported may contribute to lethal PCa. *ARHGEF12* (also known as LARG- leukemia-associated Rho guanine-nucleotide exchange factor) has also been reported as a candidate tumor suppressor gene by Ong *et al.*, 2009 in human breast and colorectal cancer. The mutation or any change in expression of *ARHGEF12* may have an effect in PCa. *FOLH1* also known as *GCPII* and *PSM*, when expressed as prostate-specific membrane antigen (PSMA) and folate hydrolase (FOLH1) has been shown to be related to breast and PCa risk (Divyya *et al.*, 2013). *PSMA* overexpression in PCa has been used for targeted therapy (Cao *et al.*, 2007). In addition *PSM-E* which is a spliced variant of *PSMA*, has been shown in a research study that could suppress proliferation, migration and invasiveness of PCa cells (Cao *et al.*, 2012). Another gene *TP53AIP1* (tumor protein p53 regulated apoptosis inducing protein 1) has been associated in p53 signaling pathway which is one of most commonly disrupted pathways in PCa (Osman *et al.*, 1999). The literature findings supporting the significance of mutations may be just implicate or create a hypothesis but give a certain evidence to proceed in another direction to understand PCa genomics. Furthermore, samples S1(PR0508), S6(PR3027) and S7(PR3043) harbored mutations in chromosome 17 in enhancers of genes *NCOR1*, *IGF2BP1* and *DLX4* respectively. *NCOR1* when phosphorylated (PKA-dependent) act as a suppressor of a transcriptional activity of AR in PCa (Choi *et al.*, 2013). Thus negative regulation of *NCOR1* will affect the transcriptional activity of AR in opposite fashion i.e. de-repression. *PTEN* has been found mutated by many research studies and its effect in progression of PCa, the gene *IGF2BP1* was reported to enhance the *PTEN* expression in a study by Stöhr *et al.*, 2012. Thus *IGF2BP1* can indirectly affect PCa. *DLX4* gene which is widely expressed in many cancers including PCa was reported in a study to promote epithelial to mesenchymal transition of cancer cells, cancer migration, invasion and metastasis (Zhang *et al.*, 2012). The mechanism and pathways involved is unclear. It can be seen clearly that each gene could affect in a certain way directly or indirectly which may have an effect in PCa. In chromosome 21 where the significant fusion between *TMPRSS2-ERG* genes occurs harbored 3 enhancer mutation across samples S4 (PR1783) and S5 (PR2832). Two mutations in same sample S4 and other one in S5 which is also a fusion positive as seen in table 4.1. Two genes in S4 whose enhancers contain a mutation were *BTG3* and *SON*. *SON* is a MAPK modulated gene which has been seen to play major role in the proliferation, survival, and tumorigenicity of pancreatic cancer cells (Furukawa *et al.*, 2012). "Signal transduction via mitogen activated protein (MAP) kinases plays a key role in a variety of cellular responses, including proliferation, differentiation, and cell death" (Maroni *et al.*, 2004). Study done by Mukherjee *et al.*, 2011 have shown that MAP kinase (MAPK) pathway may promote development of castrate-resistant PCa (CRPC). On the contrary MAPK/ERK activation inhibits PCa cell proliferation (Moro *et al.*, 2007). As MAPKs regulate diverse cellular programs, the role of the MAPK-associated *SON* in PCa is not evident yet. "B-cell translocation gene 3 (*BTG3*) is a member of the BTG family which inhibits cell proliferation, metastasis, and angiogenesis, and also regulates cell-cycle progression and differentiation in a variety of cell types" (Deng *et al.*, 2013). In many cancer types like renal cancer, breast cancer, ovarian cancer and lung cancer decreased expression, down-regulation, epigenetic silencing or inactivation (which inhibits its tumor suppressing property) of the gene has been related to carcinogenesis and subsequent progression of cancer (Chen *et al.*, 2013; Deng *et al.*, 2013; Majid *et al.*, 2009; Jingwei Yu *et al.*, 2008). In sample S5 the gene central to PCa *TMPRSS2* had a mutation in its enhancer region. Interesting fact is that the tumor sample is also *TMPRSS2-ERG* fusion positive. This fusion has been discussed elaborately in previous sections.

Chromosomes chr5, chr7 and chr8 were seen to harbor mutations on two occasions across different samples. Chr5 mutation was seen in S5 (PR2832) and S6 (PR3027) in enhancers of genes *IRX4* and *DOCK2* respectively. Similarly Chr7 mutation was seen in S3 (*NRCAM* gene) and S4 (*PDGFA* gene) while Chr8 enhancer was mutation twice in S4 (genes *MTSS1* and *POTEA*). *IRX4* is expressed in prostate cells and its expression is related to tumor suppressive effect in PCa (Nguyen *et al.*, 2012). El-Haibi *et al.*, 2011 found a positive role for dedicator of cytokinesis 2 (*DOCK2*) in PCa growth. Neuron-glia-related cell-adhesion molecule (*NRCAM*) is frequently expressed in PCa, and a recent study by Tsourlakis *et al.*, 2013 conclude association of high *NRCAM* expression with favorable tumor phenotype and reduced risk of prostate specific antigen recurrence. *PDGF* has been reported to assist tumor progression through PI3K pathway activation in a study by Werth *et al.*, 2008. Metastasis suppressor 1 (*MTSS1*) or *MIM* as the name suggest has anti-metastatic properties which was shown in breast and bladder cancer by Parr & Jiang, 2009 and Lee *et al.*, 2002 respectively. The role of before-mentioned gene still is not evident in PCa. *POTE* ankyrin domain family member A is a member of *POTE* gene family which is selectively expressed in prostate, testis, ovary, and placenta, as well as in PCa (Bera *et al.*, 2002). Though its expression is observed in LnCAP PCa cell lines and some PCas the expression is almost undetectable in normal essential tissues.

Some mutations were also observed occurring in the respective chromosome just once in different samples. GPR68 (also known as Ovarian cancer G protein-coupled receptor 1, *OGR1*) in sample S1 in chr14 was revealed as metastasis suppressor gene like *MTSS21* in PCa (Singh et al., 2007; Yan *et al.,* 2014). In sample S4 where most number of mutations were seen Chr6 and Chr13 harbored enhancer mutation in *MYB* and *CDK8* genes. *MYB* in a study by Srivastava *et al.*, 2012 showed its overexpression resulting in PCa cells malignancy and its role in castration resistance. The upregulation was many folds in castration-resistant cells as compared with androgen-dependent (LNCaP) cells. Gu *et al.*, 2013 in their research show oncogenic (melanoma and colorectal cancers) *CDK8* in a contrasting role as a tumor-suppressor gene in endometrial cancers. In other study Adler *et al.*, 2012 state role of *CDK8* in tumor growth and maintenance of tumor dedifferentiation. The role of *CDK8* in PCa is still to be elucidated. “The *MGMT* gene is an important DNA repair gene and plays a central role in the pathogenesis of cancer” (Du *et al.*, 2013). In their study Du *et al.*, 2013 have reported polymorphic factors in *MGMT* gene associated with cancer risk. *Leu84Phe* polymorphism was shown to be associated with increased risk of PCa with other types of cancers also. This gene was detected in sample S5 of the current study in enhancer of Chr10. Another mutation in the same sample was seen in Chr1 related to *HSD3B1* gene. 3β-hydroxysteroid dehydrogenase type 1 was reported to have gain of function resulting from mutation in CRPC which catalyzes conversion of the adrenal-derived steroid dehydroepiandrosterone to DHT. PCa growth is dependent on androgen stimulation, and DHT is one the most potent androgen. *HSD3B1* gene provides an alternate mechanism for DHT synthesis (Chang *et al.*, 2013). *CHL1* gene promotes cancer development (Senchenko *et al.*, 2011). A mutation related to *CHL1* was detected in sample S7 in Chr3 in our study. Region where gene is located on Ch3 was shown to have PCa susceptibility in Finnish PCa families (Rökman *et al.*, 2005).

The results we have discussed to a certain extent imply that direction of research in which we have proceeded is still unexplored and there is need to further look into the role of the mutations for their exact role in the PCa. We believe the research has contributed to some extent to at least attract some attention in the role of regulatory regions in PCa but there is need for more dedicated and exclusive research in this aspect of PCa.

7. Conclusion

As the cost of sequencing has considerably decreased which has been the case for past few years, PCa genomics which was concentrated only to exomes in various studies has now expanded to whole genomes. Though sequencing was achievable of whole genomes, interpreting mutation in vast non-coding genome was still not trivial. In this study we have made an effort to expand and explore biological insights of PCa as described by Berger *et al.*, 2011 and shifted the objective to interpreting and analyzing mutations in coding region to non-coding region. With justification of participation of non-coding region in cellular machinery as supported by various published studies we have achieved to some extent in our objective in focusing and analyzing mutations in regulatory regions of PCa genome. Satisfactory results were produced through our pipeline of selectively chosen independent tools which formed the base of our analysis. Though tools were selected according to their performances and external data utilized was generic (DNase HS was not from the samples utilized in the study) it is difficult to conclude that the pipeline was optimal but accuracy and precision of mutation detection or whole analysis could give much improved results if different approach is followed in selecting tools and their utilization. This can only be possible through experience. Mutations observed support our hypothesis but there is further need for specific analysis of each significant mutation and its validation. No direct evidence was found in the literature relating mutations to PCa but indirectly satisfactory information have been found in already published literature about the genes linked to the mutations reflecting their role in PCa. In addition if expression profile of those genes could be mapped the overall picture will become much clear if suspected genes are showing aberrant expression, but for further validation of suspected mutations further expansion of work may be needed which will encompass the molecular biology involving TFs and their binding sites. PCa has such complex genomics of having subtypes characterized by type specific aberrations and in addition more complications resulting during the course of disease, our research is a minute step to attract the attention towards our hypothesis but there is need for more specific work (region based) to be done and with novel statistical techniques and bigger sample size it may result in further expansion of understanding about PCa.

8. References

Abbas, A. & Gupta, S. The role of histone deacetylases in prostate cancer. *Epigenetics* **3,** 300–9 (2008).

Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

Adler, A. S. *et al.* CDK8 maintains tumor dedifferentiation and embryonic stem cell pluripotency. *Cancer Res.* **72,** 2129–39 (2012).

Amit, I. *et al.*Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* **326,** 257–63 (2009). (Cited by Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–55 (2011))

Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38,** 652–8 (2006).

ANNOVAR, refSeq output.
At http://www.openbioinformatics.org/annovar/annovar_gene.html#output1

Arber, S., Ladle, D. R., Lin, J. H., Frank, E. & Jessell, T. M. ETS gene Er81 controls the formation of functional connections between group Ia sensory afferents and motor neurons. *Cell* **101,** 485–98 (2000).

BamUtil, Abecasis Group. At http://genome.sph.umich.edu/wiki/BamUtil

Banine, F. *et al.* SWI/SNF chromatin-remodeling factors induce changes in DNA methylation to promote transcriptional activation. *Cancer Res.* **65,** 3542–7 (2005).

Barbieri, C. E. & Tomlins, S. A. The prostate cancer genome: perspectives and potential. *Urol. Oncol.* **32,** 53.e15–22 (2014).

Barbieri, C. E. *et al.* Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **44,** 685–9 (2012).

Barbieri, C. E. *et al.* The mutational landscape of prostate cancer. *Eur. Urol.* **64,** 567–76 (2013).

Beke, L., Nuytten, M., Van Eynde, A., Beullens, M. & Bollen, M. The gene encoding the prostatic tumor suppressor PSP94 is a target for repression by the Polycomb group protein EZH2. *Oncogene* **26,** 4590–5 (2007).

Beltran, H. *et al.* Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. *Eur. Urol.* **63,** 920–6 (2013).

Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–9 (2008). (Cited by Metzker, 2010)

Bera, T. K. *et al.* POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16975–80 (2002).

Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–20 (2011).

Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010). (Cited by Koboldt et al., 2012)

Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

Bohlander, S. K. ETV6: a versatile player in leukemogenesis. *Semin. Cancer Biol.* **15,** 162–74 (2005).

Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–22 (2008).

Brockman, W. *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* **18,** 763–70 (2008). (Cited by Koboldt et al., 2009)

Brown TA. Accessing the Genome. (2002). At http://www.ncbi.nlm.nih.gov/books/NBK21137/

Brusselmans, K., De Schrijver, E., Verhoeven, G. & Swinnen, J. V. RNA interference-mediated silencing of the acetyl-CoA-carboxylase-alpha gene induces growth inhibition and apoptosis of prostate cancer cells. *Cancer Res.* **65,** 6719–25 (2005).

Burkhardt, L. *et al.* CHD1 is a 5q21 tumor suppressor required for ERG rearrangement in prostate cancer. *Cancer Res.* **73,** 2795–805 (2013). (Cited by Barbieri & Tomlins, 2014)

Burrows, M. A block-sorting lossless data compression algorithm.
At http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.6177

Cairns, B. R. The logic of chromatin architecture and remodelling at promoters. *Nature* **461,** 193–8 (2009).

Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40,** 722–9 (2008). (Cited by Koboldt et al., 2012)

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455(7216)**:1061-8. doi: 10.1038/nature07385 (2008).

Cao, K.-Y. *et al.* New alternatively spliced variant of prostate-specific membrane antigen PSM-E suppresses the proliferation, migration and invasiveness of prostate cancer cells. *Int. J. Oncol.* **40,** 1977–85 (2012).

Chang, K.-H. *et al.* A gain-of-function mutation in DHT synthesis in castration-resistant prostate cancer. *Cell* **154,** 1074–84 (2013).

Chen, X. *et al.* Downregulation of BTG3 in non-small cell lung cancer. *Biochem. Biophys. Res. Commun.* **437,** 173–8 (2013).

Choi, H.-K. *et al.* Protein kinase A phosphorylates NCoR to enhance its nuclear translocation and repressive function in human prostate cancer cells. *J. Cell. Physiol.* **228,** 1159–65 (2013).

Christoforides, A. *et al.* Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics* **14,** 302 (2013).

Clancy, S. Copy number variation. *Nature Education* **1(1)**:95 (2008).

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38,** 1767–71 (2010).

Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–8 (2008). (Cite by Taylor *et al*., 2010)

Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **3,** 503–9 (2006).

Demichelis, F. *et al.* TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene* **26,** 4596–9 (2007). (Cited by Tomlins et al., 2007)

Deng, B. *et al.* Decreased expression of BTG3 was linked to carcinogenesis, aggressiveness, and prognosis of ovarian carcinoma. *Tumour Biol.* **34,** 2617–24 (2013).

Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999–1005 (2010). (Cited by Koboldt et al., 2012)

Ding, L. *et al*. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–75 (2008). (Cite by Taylor *et al*., 2010)

Divyya, S. *et al.* Association of glutamate carboxypeptidase II (GCPII) haplotypes with breast and prostate cancer risk. *Gene* **516,** 76–81 (2013).

Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012). (Cited by Bernstein, B. E. *et al.* (2012))

Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** e105 (2008). (Cited by Metzker, 2010)

Du, L. *et al.* The polymorphisms in the MGMT gene and the risk of cancer: a meta-analysis. *Tumour Biol.* **34,** 3227–37 (2013).

El-Haibi, C. P., Singh, R., Sharma, P. K., Singh, S. & Lillard, J. W. CXCL13 mediates prostate cancer cell proliferation through JNK signalling and invasion through ERK activation. *Cell Prolif.* **44,** 311–9 (2011).

Felsenfeld, G., Boyes, J., Chung, J., Clark, D. & Studitsky, V. Chromatin structure and gene expression. *Proc. Natl. Acad. Sci.* **93,** 9384–9388 (1996).

Ferragina, P. & Manzini, G. Opportunistic data structures with applications. 390 (2000). At http://dl.acm.org/citation.cfm?id=795666.796543

Furukawa, T. *et al.* Targeting of MAPK-associated molecules identifies SON as a prime target to attenuate the proliferation and tumorigenicity of pancreatic cancer cells. *Mol. Cancer* **11,** 88 (2012).

Genome Bioinformatics group (UCSC). Comparison of UCSC and NCBI human assemblies. At https://genome.ucsc.edu/FAQ/FAQreleases.html#release4
(Cited by Pabinger, S. *et al*. 2013)

Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012). (Cited by Bernstein, B. E. *et al.* (2012))

Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1513–8 (2011). (Cited by Mardis, 2013)

Graham J. G. Upton. Fisher's Exact Test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* , Vol. 155, No. 3 (1992) , pp. 395-402 Published by: Wiley for the Royal Statistical Society Article DOI: 10.2307/2982890. At: http://www.jstor.org/stable/2982890

Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487,** 239–43 (2012). (Cited by Barbieri & Tomlins, 2014)

Griffiths AJF, Gelbart WM, Miller JH, *et al*. Chromosomal Rearrangements. (1999). At http://www.ncbi.nlm.nih.gov/books/NBK21367/

Guyton, A.C. and J.E. Hall, Textbook of medical physiology. 11th ed. 2006, Philadelphia: *Elsevier Saunders*. xxxv, **1116 p**. (Cited in Doctoral thesis by Thellenberg Karlsson, Camilla, 1972-. - Prostate cancer [Elektronisk resurs] : epidemiological studies of risk factors. - 2008. - ISBN: 978-91-7264-594-3)

Gu, W. *et al.* Tumor-suppressive effects of CDK8 in endometrial cancer cells. *Cell Cycle* **12,** 987–99 (2013).

Haffner, M. C. *et al.* Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat. Genet.* **42,** 668–75 (2010).

He, H. H. *et al.* Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome Res.* **22,** 1015–25 (2012).

Heinz, S., Benner, C., Spann, N. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell, 38*(4), 576–89 (2010). doi:10.1016/j.molcel.2010.05.004

Helgeson, B. E. *et al.* Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. *Cancer Res.* **68,** 73–80 (2008).

Hershko, D. D. Oncogenic properties and prognostic implications of the ubiquitin ligase Skp2 in cancer. *Cancer* **112,** 1415–24 (2008).

Homer Software and Data Download. At http://homer.salk.edu/homer/ngs/annotation.html

Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339,** 959–61 (2013).

Horoszewicz, J. S. *et al.* LNCaP model of human prostatic carcinoma. *Cancer Res.* **43,** 1809–18 (1983).

Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339,** 957–9 (2013).

Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat. Genet.* **46,** 126–35 (2014).

Imamura, Y. *et al.* FOXA1 promotes tumor progression in prostate cancer via the insulin-like growth factor binding protein 3 pathway. *PLoS One* **7,** e42456 (2012).

Iwamoto, M. *et al.* Transcription factor ERG variants and functional diversification of chondrocytes during limb long bone development. *J. Cell Biol.* **150,** 27–40 (2000).

Jin, H.-J., Zhao, J. C., Ogden, I., Bergan, R. C. & Yu, J. Androgen receptor-independent function of FoxA1 in prostate cancer metastasis. *Cancer Res.* **73,** 3725–36 (2013).

Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316,** 1497–502 (2007).

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008). (Cite by Taylor *et al*., 2010)

Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–76 (2012).

Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–5 (2009).

Kumar-Sinha, C., Tomlins, S. A. & Chinnaiyan, A. M. Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer* **8,** 497–511 (2008).

Kumar-Sinha, C., Tomlins, S. A. & Chinnaiyan, A. M. Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer* **8,** 497–511 (2008).

Küntzer, J., Eggle, D., Klostermann, S. & Burtscher, H. Human variation databases. *Database (Oxford).* **2010,** baq015 (2010).

Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–31 (2012).

Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13,** 233–45 (2012).

Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–8 (2008).

Li, J. *et al.* PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275,** 1943–7 (1997).

Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24,** 713–4 (2008).

Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343,** 78–85 (2000).

Lin, C. *et al.* Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139,** 1069–83 (2009).

Lindberg, J. *et al.* The mitochondrial and autosomal mutation landscapes of prostate cancer. *Eur. Urol.* **63,** 702–8 (2013).

Lodish H, Berk A, Zipursky SL, *et al*. Chromosomal Organization of Genes and Noncoding DNA. (2000). At http://www.ncbi.nlm.nih.gov/books/NBK21571/

Lodish H, Berk A, Zipursky SL, *et al.* Mutations: Types and Causes. (2000). At http://www.ncbi.nlm.nih.gov/books/NBK21578/

Ludlow, L. B. *et al.* Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome. *J. Biol. Chem.* **271,** 22076–80 (1996).

Magnard, C. *et al.* BRCA1 interacts with acetyl-CoA carboxylase through its tandem of BRCT domains. *Oncogene* **21,** 6729–39 (2002).

Majid, S. *et al.* BTG3 tumor suppressor gene promoter demethylation, histone modification and cell cycle arrest by genistein in renal cancer. *Carcinogenesis* **30,** 662–70 (2009).

Mani, R.-S. *et al.* Induced chromosomal proximity and gene fusions in prostate cancer. *Science* **326,** 1230 (2009).

Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **6,** 287–303 (2013).

Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.* **11,** 285–96 (2010).

Maroni, P. D., Koul, S., Meacham, R. B. & Koul, H. K. Mitogen Activated Protein kinase signal transduction pathways in the prostate. *Cell Commun. Signal.* **2,** 5 (2004).

Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7,** 29–59 (2006).

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, *7*, 29–59. doi:10.1146/annurev.genom.7.080505.115623

Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

MIM, a potential metastasis suppressor gene in bladder cancer. *Neoplasia* **4,** 291–4

Mohrs, M. *et al.* Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat. Immunol.* **2,** 842–7 (2001).

Morin, A. *et al.* Identification of CAD as an androgen receptor interactant and an early marker of prostate tumor recurrence. *FASEB J.* **26,** 460–7 (2012).

Moro, L., Arbini, A. A., Marra, E. & Greco, M. Constitutive activation of MAPK/ERK inhibits prostate cancer cell proliferation through upregulation of BRCA2. *Int. J. Oncol.* **30,** 217–24 (2007).

Mukherjee, R. *et al.* Upregulation of MAPK pathway is associated with survival in castrate-resistant prostate cancer. *Br. J. Cancer* **104,** 1920–8 (2011).

Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012). (Cited by Bernstein, B. E. *et al.* (2012))

Nguyen, H. H. *et al.* IRX4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin D receptor, conferring prostate cancer susceptibility. *Hum. Mol. Genet.* **21,** 2076–85 (2012).

Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010). (Cited by Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–55 (2011))

Oh WK, Hurwitz M, D'Amico AV, *et al.* Biology of Prostate Cancer. (2003). At http://www.ncbi.nlm.nih.gov/books/NBK13217/

Ong, D. C. T. *et al.* LARG at chromosome 11q23 has functional characteristics of a tumor suppressor in human breast and colorectal cancer. *Oncogene* **28,** 4189–200 (2009).

Osman, I. et al. Inactivation of the p53 pathway in prostate cancer: impact on tumor progression. Clin. Cancer Res. 5, 2082–8 (1999).

Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. Brief. *Bioinform* doi:10.1093/bib/bbs086 (2013).

Park BH, V. B. Candidate Tumor-Suppressor Genes. (2003).
At: http://www.ncbi.nlm.nih.gov/books/NBK13782/

Parr, C. & Jiang, W. G. Metastasis suppressor 1 (MTSS1) demonstrates prognostic value and anti-metastatic properties in breast cancer. *Eur. J. Cancer* **45,** 1673–83 (2009).

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008). (Cite by Taylor *et al*., 2010)

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–95 (2013).

Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–6 (2010).

Popescu, N. C. & Zimonjic, D. B. Chromosome-mediated alterations of the MYC gene in human cancer. *J. Cell. Mol. Med.* **6,** 151–9. (Cited by Maston *et al*., 2006)

Qi, H. *et al.* AIbZIP, a novel bZIP gene located on chromosome 1q21.3 that is highly expressed in prostate tumors and of which the expression is up-regulated by androgens in LNCaP human prostate cancer cells. *Cancer Res.* **62,** 721–33 (2002).

Raab, J. R. & Kamakaka, R. T. Insulators and promoters: closer than we think. *Nat. Rev. Genet.* **11,** 439–46 (2010).

Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300,** 149–52 (1982).

Robbins, C. M. *et al.* Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. *Genome Res.* **21,** 47–55 (2011).

Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4,** 651–7 (2007).

Rökman, A. *et al.* Hereditary prostate cancer in Finland: fine-mapping validates 3p26 as a major predisposition locus. *Hum. Genet.* **116,** 43–50 (2005). (Cited by Senchenko et al., 2011)

Rosen, E. M., Fan, S. & Goldberg, I. D. BRCA1 and prostate cancer. *Cancer Invest.* **19,** 396–412 (2001).

Saglio, G. & Cilloni, D. Abl: the prototype of oncogenic fusion proteins. *Cell. Mol. Life Sci.* **61,** 2897–911 (2004). (Cited by Maston *et al*., 2006)

Scott Freeman, Kim Quillin, Lizabeth Allison. Biological Science. San Fransicso, CA : *Benjamin Cummings*,©2011.
At http://wps.prenhall.com/esm_freeman_biosci_1/7/1949/499159.cw/index.html

Senchenko, V. N. *et al.* Differential expression of CHL1 gene during development of major human cancers. *PLoS One* **6,** e15612 (2011).

Shigemizu, D. *et al.* A practical method to detect SNVs and indels from whole genome and exome sequencing data. *Sci. Rep.* **3,** 2161 (2013).

Singh, L. S. *et al.* Ovarian cancer G protein-coupled receptor 1, a new metastasis suppressor gene in prostate cancer. *J. Natl. Cancer Inst.* **99,** 1313–27 (2007).

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, *et al*. The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006). (Cite by Taylor *et al*., 2010)

Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5384 (2010).

Srivastava, S. K. *et al.* Myb overexpression overrides androgen depletion-induced cell cycle arrest and apoptosis in prostate cancer cells, and confers aggressive malignant traits: potential role in castration resistance. *Carcinogenesis* **33,** 1149–57 (2012).

Stöhr, N. *et al.* IGF2BP1 promotes cell migration by regulating MK5 and PTEN signaling. *Genes Dev.* **26,** 176–89 (2012).

Strachan T, R. A. in (Wiley-Liss, 1999). At http://www.ncbi.nlm.nih.gov/books/NBK7585/

Strachan T, R. A. Instability of the human genome: mutation and DNA repair. (1999). At http://www.ncbi.nlm.nih.gov/books/NBK7566/

Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–24 (2009). (Cited by Pleasance et al., 2010)

Sun, X. *et al.* Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *Nat. Genet.* **37,** 407–12 (2005).

Swinnen, J. V *et al.* Overexpression of fatty acid synthase is an early and common event in the development of prostate cancer. *Int. J. Cancer* **98,** 19–22 (2002).

Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18,** 11–22 (2010). (Cited by Barbieri & Tomlins, 2014)

TFBS Track, ENCODE.
At http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegTfbsClusteredV3

The Genome Reference Consortium. (Cited by Pabinger, S. *et al*. 2013)
At http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012). (Cited by Bernstein, B. E. *et al.* (2012))

Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448,** 595–9 (2007).

Tomlins, S. A. *et al.* ETS gene fusions in prostate cancer: from discovery to daily clinical practice. *Eur. Urol.* **56,** 275–86 (2009).

Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–8 (2005).

Tomlins, S. A. *et al.* TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. *Cancer Res.* **66,** 3396–400 (2006).

Trojanowska, M. Ets factors and regulation of the extracellular matrix. *Oncogene* **19,** 6464–71 (2000).

Tsourlakis, M. C. *et al.* High Nr-CAM expression is associated with favorable phenotype and late PSA recurrence in prostate cancer treated by prostatectomy. *Prostate Cancer Prostatic Dis.* **16,** 159–64 (2013).

UCSC BED format. At http://genome.ucsc.edu/FAQ/FAQformat.html#format1

Understanding Prostate Cancer, Prostate Cancer risk factors.
At http://www.pcf.org/site/c.leJRIROrEpH/b.5802027/k.D271/Prostate_Cancer_Risk_Factors.htm

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–63 (2009).

Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419,** 624–9 (2002).

VarScan User's Manual. At http://varscan.sourceforge.net/using-varscan.html#v2.3_somatic

VarScan, Variant detection in massively parallel sequencing data.
At http://varscan.sourceforge.net/somatic-calling.html

Wang, J., Cai, Y., Ren, C. & Ittmann, M. Expression of variant TMPRSS2/ERG fusion messenger RNAs is associated with aggressive prostate cancer. *Cancer Res.* **66,** 8347–51 (2006).

Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14,** 703–18 (2013).

Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, *et al*. Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893–898 (2007). (Cite by Taylor *et al*., 2010)

Werth, C. *et al.* Stromal resistance of fibroblasts against oxidative damage: involvement of tumor cell-secreted platelet-derived growth factor (PDGF) and phosphoinositide 3-kinase (PI3K) activation. *Carcinogenesis* **29,** 404–10 (2008).

Witte, J. S. Prostate cancer genomics: towards a new understanding. *Nat. Rev. Genet.* **10,** 77–82 (2009).

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, *et al*. The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007). (Cite by Taylor *et al*., 2010)

Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8,** 206–16 (2007). (Cited by Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–55 (2011).)

Wu, C. *et al.* Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer. *J. Pathol.* **227,** 53–61 (2012).

Yan, W. *et al.* Histone deacetylase inhibitors suppress mutant p53 transcription via histone deacetylase 8. *Oncogene* **32,** 599–609 (2013).

Yoon, K.-A., Nakamura, Y. & Arakawa, H. Identification of ALDH4 as a p53-inducible gene and its protective role in cellular stresses. *J. Hum. Genet.* **49,** 134–40 (2004).

Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17,** 443–54 (2010).

Yu, J. *et al.* Methylation-mediated downregulation of the B-cell translocation gene 3 (BTG3) in breast cancer cells. *Gene Expr.* **14,** 173–82 (2008).

Zhang, L. *et al.* DLX4 upregulates TWIST and enhances tumor migration, invasion and metastasis. *Int. J. Biol. Sci.* **8,** 1178–87 (2012).

Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12,** 7–18 (2011).

Yan, L., Singh, L. S., Zhang, L. & Xu, Y. Role of OGR1 in myeloid-derived cells in prostate cancer. *Oncogene* **33,** 157–64 (2014).