Genome-Wide Comparison of Copy Number Variations in Finnish Familial Prostate and Breast Cancers

Master's thesis

Oyediran Olulana, AKINRINADE
Institute of Biomedical Technology
University of Tampere

June 2012

DEDICATION

This work is dedicated to GOD ALMIGHTY, the source of my wisdom; to all restless and adventurous souls in the world; to all curious souls ever seeking to acquire knowledge to salvage the world from ruin; and lastly, to humanity.

ACKNOWLEDGEMENT

This research work was done in the Genetic Predisposition to Cancer group, Institute of Biomedical Technology, University of Tampere.

My heart goes out in profound appreciation to God for His mercies and grace. I want to thank Him for good health, sound mind and strength that kept me on in the face of seemingly insurmountable odds.

I cannot but thank my supervisors Professor Johanna Schleutker and Professor Mauno Vihinen for their meticulous supervision, support, guidance and interest in my research. Special thanks to Professor Johanna Schleutker for the great privilege of working in her group. I also like to express my appreciation to Tiina Wahlfors, PhD for her support during and after the research. Special thanks to other members of the Prostate cancer Investigator Group (PIG); Kirsi Kuusisto and Virpi Laitinen among others, it was nice working with you.

Special appreciation goes to my tutor Martti Tolvanen for his kind gestures and support. To all other staffs and my colleagues in the Bioinformatics group, I say thank you. I also would like to express my appreciation to Mira Pihlström for her support and kind gestures.

I want to in a special way express my heart-felt gratitude to my parent and siblings for their love and understanding; their unflinching support and confidence in me have been a lingering source of strength and motivation in a foreign land. I love you all.

Finally, I want to say a big thank you to all teaching staffs of the Masters Degree Program (MDP) in Bioinformatics, Tampere and Turku. Special thanks to Ortutay Csaba, Acting Professor, PhD for his kind and prompt review of my thesis.

Oyediran Akinrinade

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE Institute of Biomedical Technology
Author:	Oyediran Olulana, Akinrinade
Title:	Genome-Wide Comparison of Copy Number Variations in Finnish Familial Prostate and Breast Cancers
Pages:	69pp + appendices 9pp
Supervisors:	Professor Johanna Schleutker; Professor Mauno Vihinen
Reviewer:	Prof. Johanna Schleutker; Ortutay Csaba, Acting Prof., PhD
Time:	June 2012

Abstract

Background and aims: Prostate and breast cancers are the most prevalent types of cancer in all Western countries including Finland. For both cancer types, there is a "missing heritability" - the genetic defects predisposing individuals to the cancers remain unknown. Copy number variations (CNVs) have recently been implicated in predisposition to complex diseases including cancer. To this end, this genome-wide association study was aimed at evaluating the role of CNV in prostate and breast cancer susceptibility in the Finnish population.

Methods: Four algorithms were used in identifying CNVs in Illumina genotyped prostate and breast cancer samples; called CNVs were compared with CNVs published in the Database of Genomic Variants (DGV) in order to identify novel CNVs. Genes located within or close to the regions of CNVs were queried against the genes listed in OMIM to identify CNVs which warrant further investigation. Case-control association test was carried out using Fisher's exact test to identify CNVs associated with the cancer types in question.

Results: A total of 359 and 764 CNVs were identified in the breast and prostate cancers datasets, respectively; while the average number of CNVs per sample is higher in the prostate cancer (male genome), the size of CNVs in breast cancer dataset is double the size in prostate cancer. Three susceptibility loci were associated to prostate cancer: 2p25.3, 3p26.1 and 10q11.22. While 3p26 has previously been reported, 2p25.3 and 10q11.22 are novel. Several of the genes affected by CNVs in the datasets had already been implicated in different cancers.

Conclusion: This study is the first to compare CNVs in male and female genomes. The data suggests that several genes located within the identified CNVs may contribute to cancer predisposition in this small cohort of samples, and this trend needs to be confirmed in larger population samples.

CONTENTS

Abbreviations

BAC	Bacterial Artificial Chromosome
BAF	B Allele Frequency
BF	Bayes Factor
BIR	Break-Induced Replication
BPH	Benign Prostate Hyperplasia
CNP	Copy Number Polymorphism
CNV	Copy Number Variant/Variation
DGV	Database of Genomic Variants
DSB	Double-Stranded Break
EM	Expectation Maximization
FoSTeS	Fork Stalling and Template Switching
GWAS	Genome-Wide Association Study
HMM	Hidden Markov Model
HR	Homologous Recombination
HWE	Hardy-Weinberg Equilibrium
INDEL	Insertion or a Deletion
LCR	Low Copy Repeat
LOH	Loss of Heterozygosity
LRR	Log R Ratio
MHC	Major Histocompatibility Complex
MMBIR	Microhomology-Mediated Break-Induced Replication
MMEJ	Microhomology-Mediated End Joining
NAHR	Non-Allelic Homologous Recombination
NHEJ	Nonhomologous End Joining
OB-HMM	Objective Hidden Markov Model
OMIM	Online Mendelian Inheritance in Man
PSA	Prostate Specific Antigen
Rh	Rhesus
ROMA	Representational Oligonucleotide Microarray Analysis
SD	Standard Deviation
SDSA	Synthesis Dependent Strand-Annealing
SNP	Single Nucleotide Polymorphism
SSA	Single Strand Annealing

1 Introduction

Rare mutations have been found underlying about two thousand Mendelian diseases; more recently, it has become possible to assess the contribution of common single nucleotide polymorphisms (SNPs) to complex diseases (McCarroll and Altshuler 2007). The known role of copy-number alterations in sporadic genomic disorders, combined with emerging information about inherited copy-number variation, indicate the importance of systematically assessing copy-number variants (CNVs), including common copy-number polymorphisms (CNPs), in disease. In addition to such sporadic diseases, inherited CNVs have been found to underlie Mendelian diseases in several families (Padiath et al. 2006; Le et al. 2006, and Lee and Lupski 2006).

Chromosomal abnormalities such as germ line and somatic alterations are the leading causes of developmental defects and cancer respectively. The presence of CNVs in humans has been reported by several large scale studies, suggesting that CNVs may account for a significant proportion of human phenotypic variation, including disease susceptibility (Feuk et al. 2006; Freeman et al. 2006; and Eichler et al. 2007).

Furthermore, the role of CNVs in complex diseases, such as autism, rheumatoid arthritis and cancer to mention but a few, has been successfully evaluated by applying high throughput analysis at genome-wide level (Bae et al. 2008; Bassett et al. 2008 and Ionita-Laza et al. 2008).

Like other types of genetic variation, some CNVs have been associated with susceptibility or resistance to disease. Copy number variation has recently been implicated in predisposition to complex diseases including cancer; gene copy number can be elevated in cancer cells. For instance, the *EGFR* copy number can be higher than normal in non-small cell lung cancer (Cappuzzo et al. 2005). In addition, a higher copy number of *CCL3L1* has been associated with lower susceptibility to HIV infection (Gonzalez et al. 2005).

Breast and prostate cancers are the most prevalent female and male cancers respectively in all western countries. In Finland, the number of reported cases is constantly increasing according to the Finnish Cancer Registry's statistics (www.cancerregistry.fi). For both cancer types, there is "missing heritability", which means that although many susceptibility loci have been identified, for the majority of the cases even with strong familial background of the disease the genetic defect is still unknown.

In Finnish familial breast cancer, research has shown that only 10% and 11% of cases are attributable to mutations at *BRCA1* and *BRCA2* loci as against 52% and 32% respectively in other parts of the world (Vehmanen et al. 1997). Similar statistics have been observed in Southern Sweden (Håkansson et al. 1997). Whereas mutations in high-penetrance susceptibility genes have been identified in familial breast cancer and several single nucleotide polymorphisms (SNPs) have been shown to be associated with both familial and sporadic breast cancer risk, the impact of genomic copy number variants (CNVs) on breast cancer risk has so far poorly been studied.

Copy number variations are duplications or deletions of chromosomal segments that are greater than 1kb (Feuk et al. 2006; Itsara et al. 2009). CNV identification algorithms differ in their sensitivities and accuracies; while some identify CNVs on an individual level, others combine information from multiple samples hence the need to use different methods in CNV identification.

To this end, the present genome-wide association study (GWAS) was aimed at using different CNV identification algorithms, PennCNV, QuantiSNP, cnvPartition and CNstream, to assess genome-wide CNVs in familial breast and prostate cancers, in a bid to unraveling the contribution of CNVs to the cancer types in question.

2 Review of Related Literature

2.1 Genetic Variations

Genetic variation refers to variation in the alleles of genes, occurring both within and among populations. Genetic variation among individuals within a population can be identified at a variety of levels. It is possible to identify genetic variation from observations of phenotypic variation in either quantitative traits or discrete traits. There are three primary sources of genetic variation: mutations, gene flow and recombination in sexual reproduction; however, the ultimate source of new genetic variation in populations is via mutations, new mutations give rise to new allele. This could be point mutation or chromosomal mutation. While point mutations affect only one or a few nucleotides within a gene, chromosomal mutations change the number of chromosomes or the number or arrangement of genes in a chromosome (change in chromosome structure).

Although studies carried out by Lander et al (2001) show that any two humans are 99.9% identical at the nucleotide sequence level, many phenotypic differences are apparent in individuals within the same and from distinct human populations. Genetic diversity underlying the remaining 0.1% nucleotide differences has been postulated to contribute to phenotypic diversity among humans, and to population-specific susceptibility to disease and variability in the response to pharmacological treatments (Bamshad et al. 2004; Daar and Singer 2005). The different prevalence of Mendelian diseases reflects variability in allele frequencies for specific genes and haplotypes, and the relevance of ethnic background in the susceptibility to disease is recognized for several disorders (Botstein and Risch 2003), including cystic fibrosis (Bobadilla et al. 2002), sickle cell anaemia, and deafness (Gasparini et al. 2000), among many others. Similarly, there are differences in the prevalence of common disorders and associated genetic variants in human populations, such as the factor V Leiden (venous thromboembolic disease) (Ridker et al. 1997), variants in the *CARD15* gene (Crohn's disease) (Hugot et al. 2001), the CCR5–Δ32 variant (human immunodeficiency virus (HIV) infection and progression) (Stephens et al. 1998), and *APOE* e4 (Alzheimer's disease) (Farrer et al. 1997). There may be multiple variants of any given gene in the human population (alleles), leading to polymorphism; however, many genes have only one allele present in the population.

2.1.1 Polymorphisms

Polymorphism, defined as a genetic variant that occurs in at least 1% of a population, differs from mutations, a heritable genetic variant present in <1% of the population. The term mutation is usually used for a rare deleterious genetic change that can cause disease. Some polymorphisms could contribute to predisposition to disease; examples include human ABO blood groups, human Rh factor, and human major histocompatibility complex (MHC).

Polymorphic sequence variants usually do not cause overt debilitating diseases. Many are found outside of genes and are completely neutral in effect. Others may be found within genes, but may influence characteristics such as height and hair colour rather than characteristics of medical importance. However, polymorphic sequence variation does contribute to disease susceptibility and can also influence drug responses.

2.1.2 Single Nucleotide Polymorphisms

The most common type of variation in the human genome is the single nucleotide polymorphism (SNP), where a single base differs between individuals (Figure 1). SNPs occur about once every 1000 base pairs in the genome, making up the bulk of the three million variations found in the genome, and the frequency of a particular polymorphism tends to remain stable in the population. Unlike the other, rarer, kinds of variations, many SNPs occur in genes and in the surrounding regions of the genome that control their expression.

Single nucleotide polymorphisms (SNPs) are common biallelic variations that are widely used as genetic markers in linkage analyses and association studies (Sachidanandam et al. 2001). Most human SNPs satisfy the Hardy-Weinberg equilibrium (HWE), the condition of allelic independence, in which allele frequencies and genotype frequencies do not change over generations (Hardy 1908 cited in Lee et al. 2008).

2.1.3 Copy Number Variations

Copy number variation (CNV), one of the recently discovered classes of genetic variation,

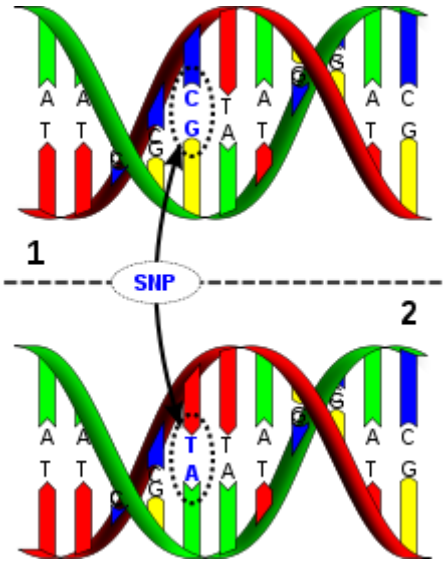

Figure 1 Diagrammatic representation of Single Nucleotide Polymorphism. DNA molecule 1 differs from DNA molecule 2 at a single base-pair location (a C/T polymorphism) (modified from en.wikipedia.org/wiki/Single-nucleotide_polymorphism).

refers to alterations in DNA fragments larger than 1kb in length when compared to a reference genome. When such alterations (insertions and deletions) are less than 1kb, they are called INDELs. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. CNVs differ from SNPs, which affect only one single nucleotide base. A vast majority of CNVs are inherited; however, some are caused by de novo mutations (Lee et al. 2007). Cytogenetic techniques such as fluorescent in situ hybridization, comparative genomic hybridization, array comparative genomic hybridization, and virtual karyotyping with SNP arrays have been effectively applied in the detection of CNVs. Moreover, recent advances in DNA sequencing technology have further enabled the identification of CNVs by next-generation sequencing (Korbel et al. 2007; Mills et al. 2011).

CNVs can be limited to a single gene or include a contiguous set of genes. It can result in having either too many or too few dosage-sensitive genes, which may be responsible for a substantial amount of human phenotypic variability, complex behavioural traits and disease susceptibility (Redon et al. 2006; Freeman et al. 2006).

First discovered in 2004 (Iafrate et al. 2004; Sebat et al. 2004), CNVs have since received much attention because of their potential implication in common disease susceptibility. When

the variant frequency is larger than 1% in a population, it is called a copy number polymorphism (CNP). In some contexts, CNV stands for copy number variants (Korbel et al. 2007), which refers to individuals whose copy number is different from the majority in a population. Copy number polymorphisms (CNPs) are of interest as they segregate at an appreciable frequency in the general population (> 1%) and are potentially implicated in the genetic basis of common diseases.

Studies have shown that about 12 - 15% of the human genome is covered by copy number variations (Redon et al. 2006; Sebat et al. 2004). Furthermore, about 56% of the CNVs identified by Iafrate et al. (2004) and Zogopoulos et al. (2007) were in known genes. The large proportion of CNVs in the genome indicates that a significant number of SNPs may fall in these regions. Nguyen et al. showed that SNPs are significantly enriched in known human CNVs (Nguyen et al. 2006).

Copy number variants (CNVs) can arise both meiotically and somatically, as shown by the finding that identical twins can have different CNVs (Bruder et al. 2008) and that repeated sequences in different organs and tissues from the same individual can vary in copy number (Piotrowski et al. 2008). The non-homologous recombination events that underlie changes in copy number also allow generation of new combinations of exons between different genes by translocation, insertion or deletion (Rotger et al. 2007; Feng et al. 2009), so that proteins might acquire new domains, and hence new or modified activities.

2.2 Mechanisms of CNV formation

CNVs form at rates far outstripping other kinds of mutagenesis, and appear to do so by similar mechanisms in bacteria, yeast, and human. Change in copy number involves change in the structure of the chromosomes and the mechanisms of all structural changes are the same as those that cause CNV. There are two general mechanisms that produce changes in the structure of chromosomes: homologous recombination (HR) and nonhomologous recombination.

2.2.1 Homologous Recombination

HR requires extensive DNA sequence identity (about 50bp in *E. coli* (Lovett et al. 2002), to as many as 300bp in mammalian cells and human (Liskay et al. 1987)) and most mechanisms also require a strand exchange protein, RecA in prokaryotes and its orthologue Rad51 in eukaryotes. HR underlies many DNA repair processes, and is also responsible for ordered segregation of chromosomes and for generating new combinations of linked alleles at meiosis. HR is used in repair of DNA breaks and gaps. The best studied mechanism of HR is double-strand break (DSB)-induced recombination. DSB repair can take place when either two double-stranded ends are present, or when there is only one. Either of these can lead to or avoid generation of copy-number variation. When two double-stranded ends are involved, DSB repair can happen either by double Holliday junction model or by synthesis-dependent strand annealing model. While double Holliday junction DSB repair leads to gene conversion and crossing over, synthesis-dependent strand-annealing (SDSA) does not generate crossovers. Crossing-over between homologous chromosomes can lead to loss of heterozygosity (LOH) if the chromatids carrying the same alleles segregate together at mitosis. Duplication and deletion of sequence result from the formation of crossovers between homologies in non-allelic positions on the same chromosome (NAHR).

CNVs can result from HR either via non-allelic homologous recombination (NAHR), break-induced replication (BIR) or single-stranded annealing. NAHR is a recombination repair event that uses a direct repeat as homology. A crossover outcome from this event leads to products that are reciprocally duplicated and deleted for sequence between the repeats. These might segregate from each other at the next cell division, thus changing the copy number in both daughter cells (Figure 2A (i)).

Another form of NAHR is BIR. In BIR, broken molecule uses ectopic homology to restart the replication fork. BIR forms duplications and deletions in separate events (Figure 2A (ii)).

SSA does not require RecA/Rad51, but requires the annealing protein Rad52. SSA happens when neither end at a two-ended double-strand break invades homologous sequence. In this case, erosion of the 5′ ends continues, exposing substantial lengths of single-stranded 3′ ends. If this process exposes complementary sequences in the two single strands, annealing can occur. After removal of the flaps and ligation, the broken molecule has been repaired, but all

sequence between the repeat sequence and one of the repeats themselves have been deleted (Figure 2B).

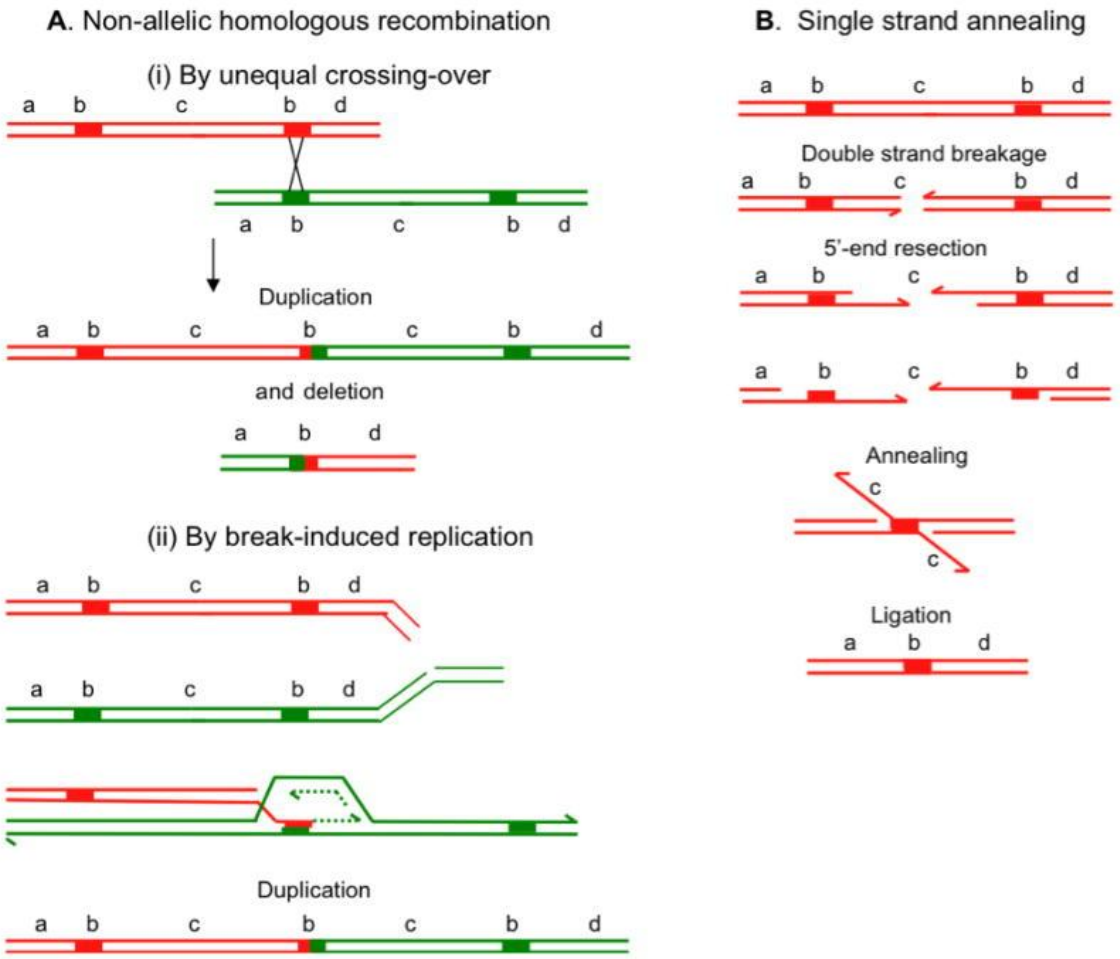

Figure 2 Change in copy number by homologous recombination. **A (i)** Non-allelic homologous recombination (NAHR). **A (ii)** Break-induced replication (BIR). **B** Single-strand annealing (SSA) (Hastings et al. 2009).

2.2.2 Nonhomologous Recombination

In contrast to HR, nonhomologous recombination mechanisms use only microhomology of a few complementary base pairs or no homology. Nonhomologous mechanisms can be divided into replicative and non-replicative mechanisms.

2.2.2.1 Non-Replicative Mechanisms

Non-replicative mechanisms do not require homology or need very short microhomologies for repair. There are two sub-categories: Nonhomologous end joining and Breakage-fusion-bridge cycle. Nonhomologous end joining is a type of DSB repair that either does not require homology or requires very short microhomologies for repair. There are two variants of this: nonhomologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) (Lieber 2008). While NHEJ rejoins DSB ends accurately or leads to small 1-4 bp deletions, and also in some cases to insertion of free DNA, often from mitochondria or retrotransposons (Haviv-Chesner et al. 2007), MMEJ uses 5 to 25 bp long homologies to anneal to ends of DSBs and, like SSA, leads to deletions of sequences between annealed microhomologies. The second distinction between these pathways is that they require different proteins. Key proteins involved in NHEJ (Ku70/Ku80) are not required for MMEJ. Also the strand-annealing protein Rad52 is not required for MMEJ, which distinguishes this pathway from SSA.

In the Breakage-fusion-bridge cycle however, an unreplicated chromosome suffers a double-strand break so that it loses a telomere. Upon replication, both sister chromatids lack telomeres. These two ends are proposed to fuse, forming a dicentric chromosome. At anaphase, the two centromeres of the dicentric chromosome are pulled apart, initially forming a bridge between the telophase nuclei. Eventually the bridge is broken in a random position. This inevitably leads to the formation of a large inverted duplication and process is repeated until the end acquires a telomere from another source.

2.2.2.2 Replicative Mechanisms

They are nonhomologous mechanisms that depend solely on replicative stress. Several studies (Kuo et al. 1994; Coquelle et al. 1997; Rozier et al. 2002) have shown that aphidicolin, an inhibitor of replicative DNA polymerases, induces CNV both at chromosomal fragile sites, and throughout the genome. Replicative mechanisms include: Replication slippage or template switching; Fork stalling and template switching; and Microhomology-mediated break-induced replication.

In replicative slippage, the length of lagging-strand template becomes exposed as a single strand during replication. Whether or not due to secondary structures in the lagging-strand

template, the 3′ primer end can move to another sequence showing a short length of homology on the exposed template and continue synthesis after having failed to copy part of the template. This results in a deletion (Figure 3A). Several variations on this mechanism can also produce a duplication of a length of DNA sequence with or without sister chromatid exchange (Lovett 2004).

In Fork stalling and template switching (FoSTeS) (Lee et al. 2007) however, exposed single-stranded lagging strand template might acquire secondary structures which can block the progress of the replication fork. The 3′ primer ends then become free from their templates, and might then alight on other exposed single-stranded-template sequence on another replication fork that shares microhomology, thus causing duplication, deletion, inversion or translocation depending on the relative position of the other replication fork. Fork stalling can be caused by other situations, such as lesions in the template strand or shortage of deoxynucleotide triphosphates (Figure 3B).

Microhomology-mediated break-induced replication (MMBIR) occurs when replication fork collapse, and there is a break off of one arm of the fork. The collapse can be as result of the fork encountering a nick on a template strand, or can be caused by endonuclease. When this happens, the 5′ end of the broken molecule will be recessed from the break, exposing a 3′ tail. When insufficient RecA or Rad51 is present to allow invasion of homologous duplex, the 3′ tail will anneal to any exposed single stranded DNA that shares microhomology (Figure 3C).

2.3 Copy Number Variations and Diseases

Like other types of genetic variation, some CNVs have been associated with susceptibility or resistance to disease. Several studies have reported and assessed the role of CNVs in complex diseases including neurological disorders and leukaemia (Iafrate et al. 2004; Sebat et al. 2004), autism, rheumatoid arthritis and idiopathic learning disability (Knight et al. 1999; Bae et al. 2008; Bassett et al. 2008; Ionita-Laza et al. 2008). Copy number of genes can be elevated in cancer cells. Studies carried out by Cappuzzo et al. (2005) reported a higher copy number of *EGFR* in non-small cell lung cancer. In addition, a higher copy number of *CCL3L1* has been associated with lower susceptibility to HIV infection, and a low copy

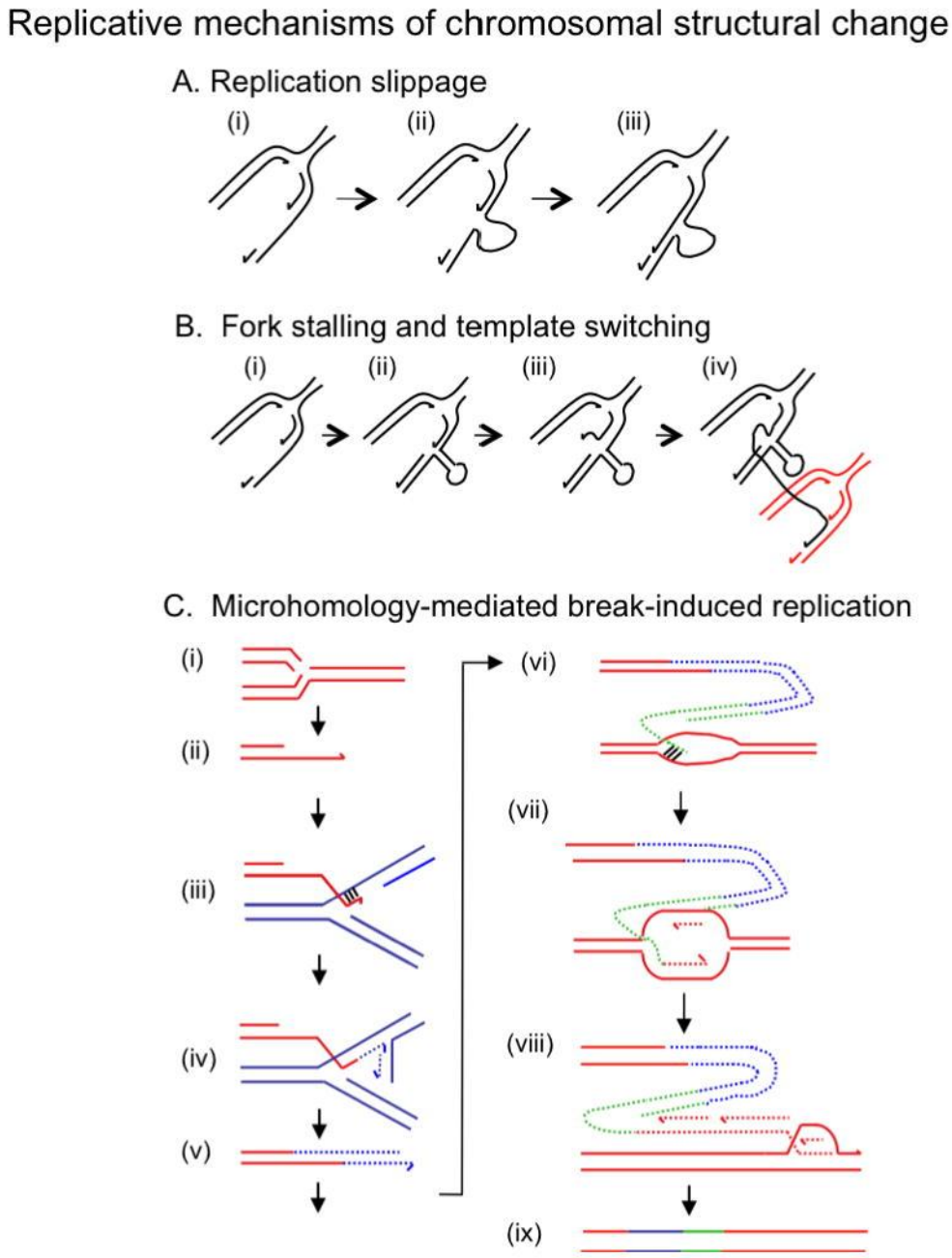

Figure 3 Replicative mechanisms for nonhomologous structural change. **A.** Replication slippage. **B.** Fork stalling and template switching (FoSTeS). **C.** Microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009).

number of *FCGR3B* (the CD16 cell surface immunoglobulin receptor) can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders (Gonzalez et al. 2005; Aitman et al. 2006).

Although most copy number variants exist in healthy individuals, these variants are however hypothesized to cause diseases through several mechanisms. First, copy number variants can directly influence gene dosage through insertions or deletions, which can result in altered gene expression and potentially cause genetic diseases. Gene dosage describes the number of copies of a gene in a cell, and gene expression can be influenced by higher and lower gene dosages. Deletions can also result in the unmasking of a recessive allele that would normally not be expressed. Structural variants that overlap a gene can reduce or prevent the expression of the gene through inversions, deletions, or translocations. Variants can also affect a gene's expression indirectly by interacting with regulatory elements (position effect). For instance, if a regulatory element is deleted, a dosage-sensitive gene might have a lower or higher expression than normal. Sometimes, the combination of two or more copy number variants can produce a complex disease, whereas individually the changes produce no effect. Some variants are flanked by homologous repeats, which can make genes within the copy number variant susceptible to non-allelic homologous recombination and can predispose individuals or their descendants to a disease (Freeman et al. 2006). Additionally, complex diseases might occur when copy number variants are combined with other genetic and environmental factors (Feuk et al. 2006).

2.4 CNV Identification Algorithms

Detection of chromosomal copy number changes in the human genome has been conducted using array-based technologies. Recent studies have identified numerous copy number variants (CNV) and some are common polymorphisms that may play a role in disease susceptibility. CNVs have been identified as a potential factor responsible for a significant proportion of human phenotypic variability that remains unexplained (McCarroll and Altshuler 2007; Ionita-Laza et al. 2009). Several technologies such as multiple ligation-dependent probe amplification and array comparative genomic hybridization can be used to characterize CNV genotypes. There are several CNV detection methods; however, differences in CNV call threshold and characteristics exist.

Experimental techniques used in detecting CNVs include bacterial artificial chromosome (BAC) arrays, paired end mapping, fluorescent in situ hybridization, representational oligonucleotide microarray analysis (ROMA) and whole genome single nucleotide polymorphism (SNP) arrays (Iafrate et al. 2004). As the use of genome-wide association

studies (GWAS) increase, SNP arrays with high density (>300,000 SNPs) have become a convenient tool for studying CNVs. Accurate CNV detection in SNP arrays require sophisticated algorithms or statistical methods. Several factors including robustness of the statistical method, batch effect, population stratification and differences between experiments influence the accuracy of CNV boundaries derived from SNP arrays.

To date, there are several detection methods available for identifying CNVs from genome-wide SNP array data, and the statistical methods underlying these approaches include Hidden Markov Models (HMMs) (Colella et al. 2007; Wang et al. 2007), segmentation algorithms (Hupé et al. 2004; Olshen et al. 2004), t-tests and standard deviation (SDs) of the log R Ratio (LRR) (Fiegler et al. 2006).

2.4.1 Sample-based CNV Calling Algorithms

Sample-based algorithms otherwise referred to as non-segmenting algorithms, perform CNV identification on an individual level. Most algorithms in this category are based on the differences in the Log R Ratio (LRR), a measure of the normalised total signal intensity, and B-Allele frequency (BAF), a measure of the normalised allelic intensity ratio measurements between samples and a model learned from a reference set. They perform well for large deletions and amplifications but are sensitive to intensity noise.

2.4.1.1 PennCNV

PennCNV is an integrated Hidden Markov Model (HMM) algorithm for detecting CNVs with high resolution using the Illumina Infinium assay. Hidden Markov Model is a statistical technique that models a Markov process, where the probability of observing a particular state at a particular time point only depends on the states at previous time points. HMM provides a natural statistical framework for modelling dependence structures between copy numbers at nearby SNPs. To detect CNVs, PennCNV uses the first-order HMM that assumes that the hidden copy number state at each SNP depends only on the copy number state of the most preceding SNP (Wang et al. 2007).

PennCNV models LRR and BAF, and developed more realistic models for state transition between different copy number states. Additionally, PennCNV incorporates the population

allele frequency for each SNP and the distance between adjacent SNPs. Unlike many other algorithms that use conventional "loss", "normal", and "gain" to model CNV states, PennCNV uses a six-state definition (Colella et al. 2007) to model CNV events more precisely (Table 1). GenomeStudio software from Illumina displays two summary measures for a genotype signal at each SNP: LRR and BAF. The patterns of LRR and BAF in regions with copy number changes are demonstrated in Figure 4. The combination of LRR and BAF can be used together to determine several different copy numbers and to differentiate copy-neutral LOH (loss of heterozygosity) regions from normal state regions, supporting the utility of six distinct copy number states in the modelling strategy.

PennCNV however provides quality filtering criteria for removing unreliable calls so as to reduce false discovery rate. A confidence score of 10 with the CNV spanning at least 3SNPs have been suggested. Moreover, family information could be incorporated in the analysis and helps to eliminate CNVs that are incompatible with Mendelian inheritance, thus improve the accuracy of the CNV calling and boundary prediction. Figure 5 outlines the procedure for CNV identification in PennCNV.

Table 1 Hidden states, copy numbers, and their description

Copy no. State	Total copy no.	Description (for autosome)	CNV genotypes
1	0	Deletion of two copies	Null
2	1	Deletion of one copy	A, B
3	2	Normal state	AA, AB, BB
4	2	Copy-neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

Each state has a different distribution of CNV genotypes

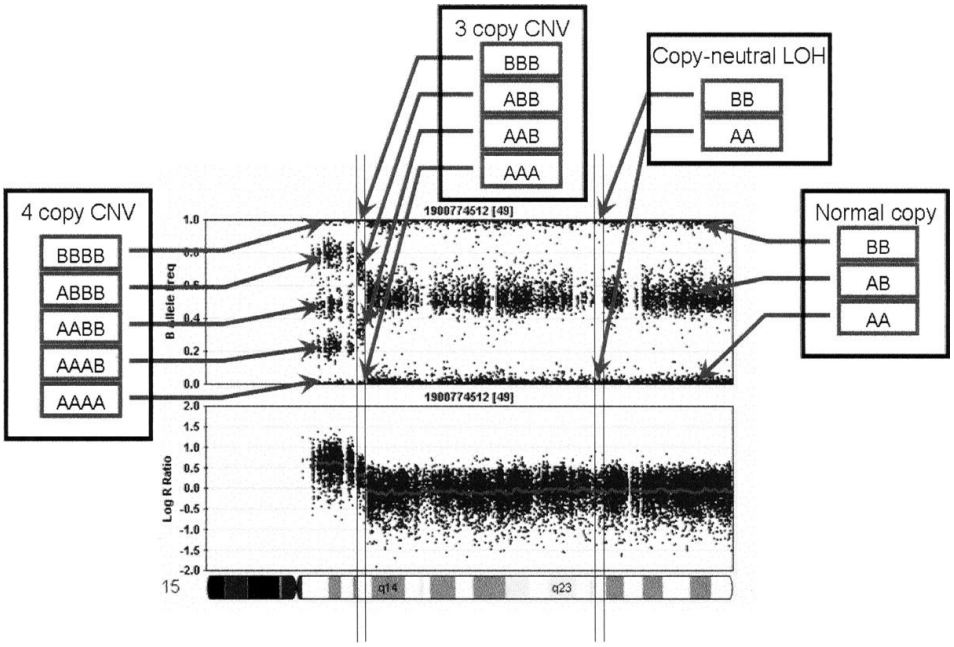

Figure 4 An illustration of log R Ratio (LRR) and B Allele Freq (BAF) values for the chromosome 15 q-arm of an individual. A normal chromosome region has three BAF genotype clusters, represented as AA, AB, and BB genotypes in boxes and with LRR values centred around zero. The copy-neutral LOH region has normal LRR values, but without the AB genotype cluster. The increased copy number for a CNV region can be detected based on an increased number of peaks in the BAF distribution, as well as increased LRR values. The patterns of LRR and BAF for different CNV regions, normal regions, and copy-neutral LOH regions are distinct from each other, thus the combination of LRR and BAF can be used to generate CNV calls (Wang et al. 2007).

2.4.1.2 QuantiSNP

QuantiSNP is a novel computational framework for detecting regions of copy number variation from Bead Array™ SNP genotyping data using an Objective Bayes Hidden-Markov Model (OB-HMM) (Colella, et al. 2007). Objective Bayes measures are used to set certain hyperparameters in the priors using a novel re-sampling framework to calibrate the model to a fixed Type I (false positive) error rate. Other parameters are set via maximum marginal likelihood to prior training data of known structure. QuantiSNP provides probabilistic quantification of state classifications and significantly improves the accuracy of segmental aneuploidy identification and mapping. QuantiSNP and PennCNV use different HMMs. While PennCNV uses a first-order HMM, QuantiSNP uses an Objective Hidden Markov Model (OB-HMM) to infer copy number variation. In the model, the hidden states are the (unknown) copy number at each SNP. The states are inferred in terms of LRR and BAF for each SNP. Table 2 lists the hidden states used in QuantiSNP's HMM. In Bayesian inference,

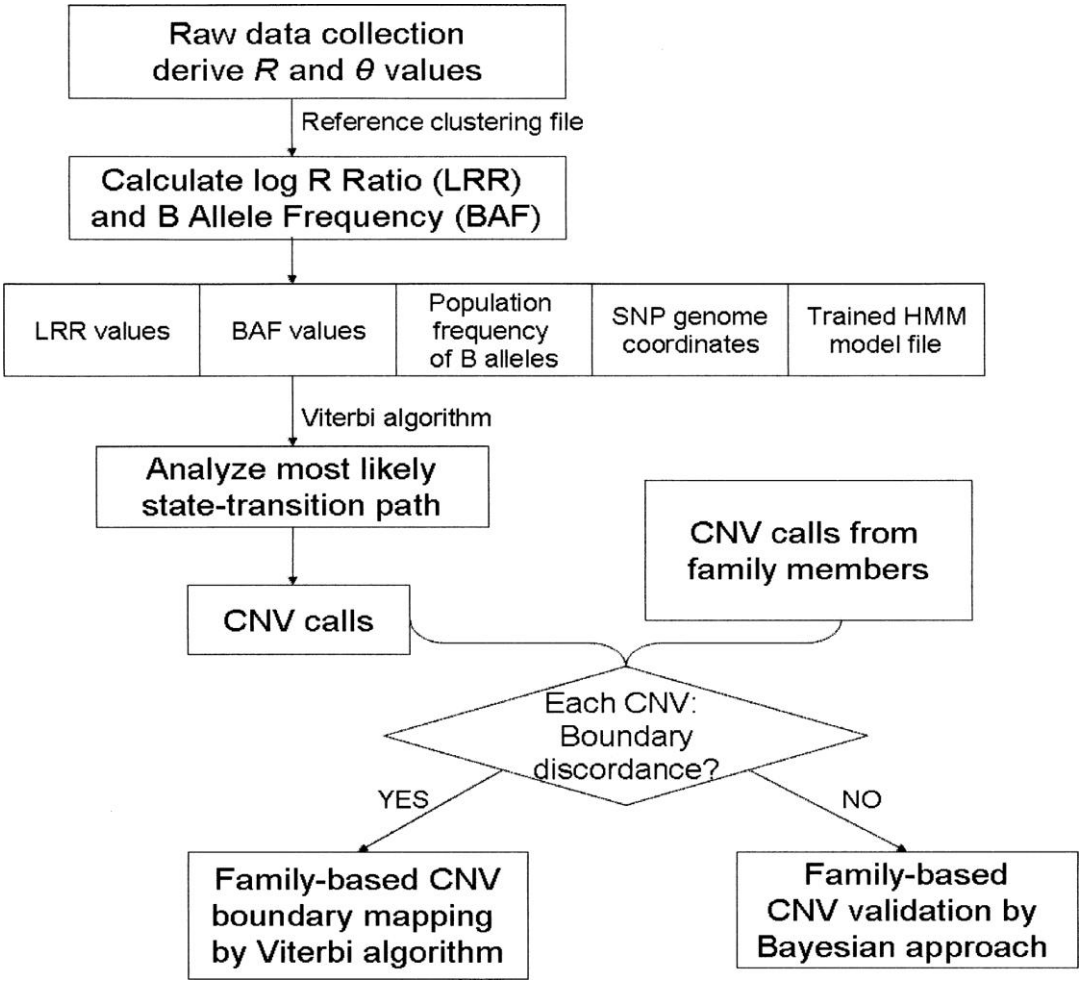

Figure 4 A flowchart outlining the procedure for CNV calling from genotyping data. The first step for LRR and BAF calculation can be alternatively performed by the BeadStudio software, given a clustering file containing canonical genotype cluster positions. The HMM integrates several sources of information to give CNV calls. When genotype data are available for family members, the pedigree information can be incorporated to model CNV events more accurately (Wang et al. 2007).

prior probability models are developed for unknown parameters and these prior beliefs are then updated in light of new data, using Bayes' Rule, to give posterior probability distributions for the parameters. In a subjective Bayesian approach, prior distributions are elicited using expert knowledge or personal beliefs, and the Bayesian framework provides a powerful means by which to incorporate such information into an inference problem.

In QuantiSNP, model parameters are learnt from the data using an Expectation Maximization (EM) algorithm. QuantiSNP assigns a Bayes Factor (BF) to each region of copy number

variation detected. Bayes Factor provides a probability measure of the strength of evidence from the data for the presence of copy number variant in a region versus the null hypothesis that there is no variant. The greater the value of BF, the stronger the evidence for the existence of a copy number variant. Table 3 shows the relationship between log Bayes Factor and false discovery. To reduce the number of false positive calls, a BF value of 10 has been suggested appropriate for filtering identified CNVs by QuantiSNP

.

Table 2 Hidden states, associated copy numbers and biological interpretation in QuantiSNP

Hidden state, z	Copy number, c(z)	Number of genotypes, K(z)	Description
1	0	0	Full deletion
2	1	1	Single copy deletion
3	2	2	Normal (heterozygote)
4	2	3	Normal (homozygote)
5	3	4	Single copy duplication
6	4	5	Double copy duplication

While QuantiSNP uses a fixed rate of heterozygosity for each SNP, PennCNV generates a hidden state for copy neutral loss of heterozygosity (LOH) and uses each population-based BAF of the SNP to infer CNVs.

Table 3 Log Bayes Factor and False Discovery used in filtering identified CNVs

Log Bayes Factor	False CNVs per/sample
0	10.84
5	0.84
10	<1

2.4.1.3 cnvPartition

cnvPartition, developed by Illumina, is available as a plug-in in the GenomeStudio software. It is based on the assumption that majority of the CNV vary between 0 and 4 copies, thus yielding five options: homozygous deletion, heterozygous deletion, dizygous (normal state), trizygous (one extra copy), and tetrazygous (two extra copies) (Table 4). cnvPartition models

LRR and BAF as a simple bivariate Gaussian distribution for each of the 14 possible genotypes. cnvPartition implemented in the Illumina GenomeStudio software uses an undocumented method of CNV detection. However, the program provides confidence score (threshold) for filtering the identified CNVs. Based on experience, a score of 10 is appropriate for the filtering.

Table 4 Copy numbers and states used in cnvPartition

Copy Number	State	Description
0	Homozygous	Single copy deletion
1	Heterozygous	Double copy deletion
2	Dizygous	Normal state
3	Trizygous	Single copy duplication
4	Tetrazygous	Double copy duplication

2.4.2 Segment-based CNV Identification Algorithms: CNstream

Segment-based algorithms, otherwise called segmentation algorithms, take information at the same locus from multiple samples to perform CNV identification. Although the primary raw data used for detecting CNVs from SNP arrays are the SNP intensity measured by LRR, some methods also use BAF to enhance detection. Circular Binary Segmentation (CBS) (Olshen et al. 2004), Nexus 4.1 Rank and Nexus 4.1 SNPRank use the same segmentation algorithm that recursively divides chromosomes into segments of common intensity distribution functions. However, while CBS has no inherent method of determining segment significance, Nexus uses an undocumented equation to compute segment significance.

CNstream is an R-statistical software package for whole-genome CNV discovery and genotyping specifically adapted for Illumina arrays, thus, the required data for the analysis can be directly extracted from GenomeStudio without any formatting step. It is an algorithm for detecting whole-genome copy number polymorphisms (CNP) (Alonso et al. 2010). It is based on a robust single locus scoring algorithm followed by a segment-based genotyping algorithm – multilocus calling. It performs a joint calling at each probe of multiple samples

and increases the accuracy of the CNP calls by considering the scores of nearby and consecutive markers. Fully implemented of R statistical software, the CNstream method is publicly available. The algorithm takes the X and Y channel intensities (measuring A and B alleles, respectively) as arguments.

CNstream performs CNP identification in four major steps: Pre-processing, SNP genotyping, Single-locus scoring, and Segment-based calling (Figure 6). Pre-processing is an optional step as the data are normalised by Illumina genotyping software. However, in order to reduce inter-plate variability, CNstream performs per-plate normalisation if the user includes the plate number of each sample. By applying a clustering algorithm, the algorithm determines the SNP genotype. Copy number scores for each probe are computed by combining the estimated number of copies of each intensity channel for each sample using a single-locus scoring approach. After single-locus scoring, CNstream analyses the scores obtained for each sample along a set of consecutive and nearby probes, referred to as segments. It outputs all the probe segments in which copy number frequency exceeds the frequency threshold (default=1%).

2.5 Prostate Cancer

Prostate cancer is cancer that starts in the prostate gland. The prostate is a small, walnut-sized structure that makes up part of a man's reproductive system. It wraps around the urethra, the tube that carries urine out of the body. Although most prostate cancers are slow growing, there are cases of aggressive prostate cancer. In 2009, Sam Lister reported that about two-thirds of cases are slow growing, the other third more aggressive and fast developing (Lister 2009). Studies carried out by Siegel et al. (2011) shows that the more aggressive prostate cancers account for more cancer-related mortality than any other cancer except lung cancer. The cancer cells may spread from the prostate to other parts of the body, and it may cause pain, difficulty in urinating, problems during sexual intercourse, or erectile dysfunction. The development of prostate starts before birth and it grows rapidly during puberty, accelerated by male hormones (androgens) in the body. Prostate cancer is the most common cause of death from cancer in men over age 75 and it is rarely found in men younger than 40. People who are at higher risk include: African-American men, who are also likely to develop cancer at every age, men who are older than 60 and men who have a father or brother affected with prostate cancer.

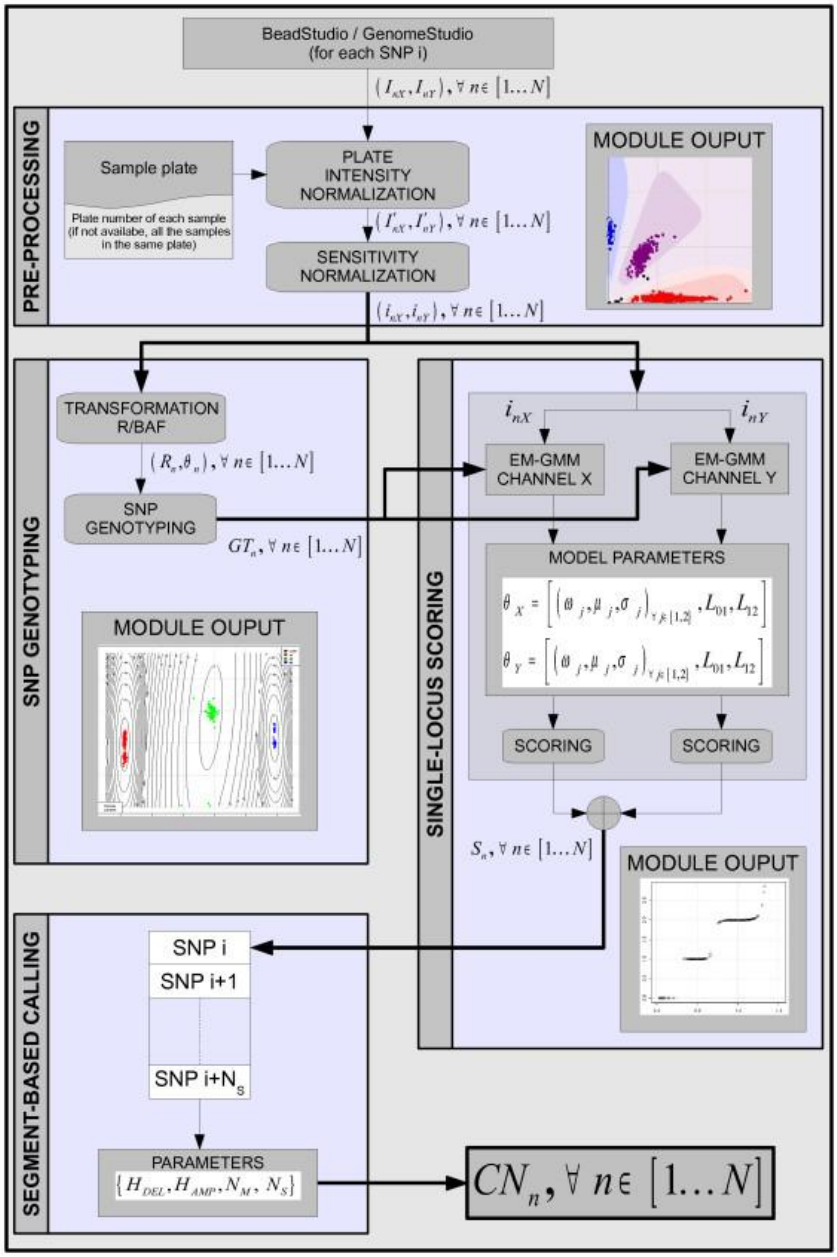

Figure 6 Flowchart of CNstream processing steps. Data processing with CNstream is organized into four functional modules: Pre-processing, SNP genotyping, Single-locus scoring and Segment-based calling (Alonso et al. 2010).

As of 2011, prostate cancer is the second most frequently diagnosed cancer and the sixth leading cause of cancer death in males worldwide (Jemal et al. 2011). Rates of prostate cancer vary widely across the world. Although the rates vary widely between countries, it is least common in South and East Asia, more common in Europe, and most common in the United States (Parkin et al. 1997).

Epidemiological and genetic studies have reported familial clustering of prostate and breast cancers. Causes of prostate cancer are essentially unknown. Several epidemiological studies have suggested various factors that might play a role in prostate cancer risk. These factors include history of benign prostate hyperplasia (BPH) (Chokkalingam et al. 2003; Guess 2001), history of high-grade prostatic intraepithelial neoplasia (PIN) (Bostwick and Qian 2004), and alcohol consumption (Sommer et al. 2004) to mention but a few. However, a lot of inconsistencies have been observed in these results. Age, ethnicity and family history are the only well-documented risk factors for prostate cancer (Crawford 2003). Furthermore, many factors, including genetics and diet, have been implicated in the development of prostate cancer. The presence of prostate cancer may be indicated by symptoms, physical examination, prostate-specific antigen (PSA), or biopsy.

Defining the full range of molecular genetic alterations in prostate cancer should provide improved understanding and new targets for prevention and treatment. The molecular alterations underlying prostate cancer are partially understood (DeMarzo et al. 2003; Shen and Abate-Shen 2010). Common events include deletion of tumor suppressors, including *CDKN1B* (p27/KIP1), *RB1*, *TP53*, *PTEN* and the prostate-specific homeobox transcription factor NKX3-1. Amplification of the MYC oncogene is also frequent. In addition, oncogenic fusions driving ETS-family oncogenic transcription factors (*ERG*, *ETV1*, *ETV4* and *ETV5*), most commonly as *TMPRSS2-ERG*, have been identified in approximately half of prostate cancers. Androgen receptor alterations, including amplification and rearrangement, can also occur in castration-recurrent prostate cancer. More recently, genomic profiling by comparative genomic hybridization (CGH) and single-nucleotide polymorphism arrays have provided comprehensive views of DNA copy number alterations in prostate cancer (Lapointe et al. 2004; Kim et al. 2007; Robbins et al. 2011), and have led to the nomination of new prostate cancer genes, for example, *NCOA2* (Taylor et al. 2010). Next-generation genome sequencing is also now beginning to reveal the full landscape of somatic rearrangements (Berger et al. 2011). Although several susceptibility loci have been identified by various research groups worldwide, yet the genetic defect underlying prostate cancer is unknown. Over the years, there have been many published reports of possible linkage of prostate cancer susceptibility to different chromosomes, but the results have not always been reproducible between studies (Table 5).

2.6 Breast Cancer

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can invade surrounding tissues or spread (metastasize) to distant areas of the body. The disease occurs almost entirely in women, but men can get it, too. The female breast is made up mainly of lobules (milk-producing glands), ducts (tiny tubes that carry the milk from the lobules to the nipple), and stroma (fatty tissue and connective tissue surrounding the ducts and lobules, blood vessels, and lymphatic vessels). Most breast cancers begin in the cells that line the ducts (ductal cancers). Some begin in the cells that line the lobules (lobular cancers), while a small number start in other tissues (Sariego 2010). Studies have shown that breast cancer comprises 22.9% of all cancers (excluding non-melanoma skin cancers) in women. In 2008, breast cancer caused 458,503 deaths worldwide (13.7% of cancer deaths in women). Breast cancer is more than 100 times more common in women than

Table 5 Putative prostate cancer susceptibility loci (modified from omim.org).

Location	Gene/Locus	Ensembl Gene ID	References
1q25.3	*RNASEL*	ENSG00000135828	Smith et al. 1996
1q42.2-q43	PCAP	NA	Berthon et al. 1998
3p26	HPC5	NA	Rokman et al. 2005
7p22.3	*MAD1L1*	ENSG00000002822	Tsukasaki et al. 2001
7p11-q21	HPC4	NA	Friedrichsen et al. 2004
7q11.23	*HIP1*	ENSG00000127946	Rao et al. 2002
8p22	*MSR1*	ENSG00000038945	Xu et al. 2001
10p15.1	*KLF6*	ENSG00000067082	Narla et al. 2001
10q23.31	*PTEN*	ENSG00000171862	Cairns et al. 1997
10q25.2	*MXI1*	ENSG00000119950	Eagle et al. 1995
11p11.2	*CD82*	ENSG00000085117	Dong et al. 1995
13q13.1	*BRCA2*	ENSG00000139618	Gronberg et al. 2001
16q22.1	*CDH1*	ENSG00000039068	Jonsson et al. 2004
16q22.2-q22.3	*ZFHX3*	ENSG00000140836	Sun et al. 2005
17p12	*ELAC2*	ENSG00000006744	Rokman et al. 2001
19q	HPCQTL19	NA	Witte et al. 2000
20q13	HPC3	NA	Berry et al. 2000
22q12.1	*CHEK2*	ENSG00000183765	Dong et al. 2003
22q12.3	HPC6	NA	Xu et al. 2005
Xq12	*AR*	ENSG00000169083	Gaddipati et al. 1994

in men, although males tend to have poorer outcomes due to delays in diagnosis (World cancer report).

The first noticeable symptom of breast cancer is typically a lump that feels different from the rest of the breast tissue. Research has shown that more than 80% of breast cancer cases are discovered when the woman feels a lump. The earliest breast cancers are detected by a mammogram. Lumps found in lymph nodes located in the armpits can also indicate breast cancer (Merck Manual of Diagnosis and Therapy, 2003). Indications of breast cancer other than a lump may include changes in breast size or shape, skin dimpling, nipple inversion, or spontaneous single-nipple discharge. Pain is an unreliable tool in determining the presence or absence of breast cancer, but may be indicative of other breast health issues.

The primary risk factors for breast cancer are female sex (Giordano et al. 2004), age, lack of childbearing or breastfeeding (Collaborative Group on Hormonal Factors in Breast Cancer 2002), higher hormone levels (Yager and Davidson 2006; Santoro et al. 2009), race, economic status and dietary iodine deficiency (Venturi 2001; Aceves et al. 2005). The incidence of breast cancer varies greatly around the world: it is lowest in less-developed countries and highest in the more-developed countries. The number of reported cases has witnessed a significantly increase since the 1970s, a phenomenon partly attributed to the modern lifestyles. Breast cancer is strongly related to age with only 5% of all breast cancers occurring in women under 40 years old (Breast Cancer, 2006).

To date, 22 common breast cancer susceptibility loci have been identified accounting for ~8% of the heritability of the disease (Table 6). Only a very small fraction of cases in the general population, however, can be explained by high-penetrance breast cancer susceptibility genes, such as *BRCA1* and *BRCA2*, besides, little mutations have been found at these loci in the Finnish population.

In 2002, Thompson et al evaluated the contribution of the *BRCA3* locus on 13q21 to breast cancer susceptibility in 128 high-risk breast cancer families of western European ancestry with no identified *BRCA1* or *BRCA2* mutations. No evidence of linkage was found. They therefore concluded that, if a susceptibility gene does exist at 13q21, it can account for only a small proportion of non-*BRCA1/2* families with multiple cases of early-onset breast cancer (Thompson et al. 2002).

Table 6 Putative breast cancer susceptibility loci (modified from omim.org).

Location	Gene/Locus	Ensembl Gene ID	References
1p34.1	*RAD54L*	ENSG00000085999	Matsuda et al. 1999
2q33.1	*CASP8*	ENSG00000064012	MacPherson et al. 2004
2q35	*BARD1*	ENSG00000138376	Thai et al. 1998
3q26.32	*PIK3CA*	ENSG00000121879	Lee et al. 2005
5q34	*HMMR*	ENSG00000072571	Pujana et al. 2007
6p25.2	*NQO2*	ENSG00000124588	Yu et al. 2009
8q11.23	*RB1CC1*	ENSG00000023287	Chano et al. 2002
11p15.4	*SLC22A1L*	ENSG00000110628	Gallagher et al. 2006
11p15.1	*TSG101*	ENSG00000074319	Steiner et al. 1997
11q22.3	*ATM*	ENSG00000149311	Broeks et al. 2000
12p12.1	*KRAS*	ENSG00000133703	Yanez et al. 1987
13q13.1	*BRCA2*	ENSG00000139618	Healey et al. 2000
14q32.33	*XRCC3*	ENSG00000126215	Kuschel et al. 2002
14q32.33	*AKT1*	ENSG00000142208	Carpten et al. 2007
15q15.1	*RAD51A*	ENSG00000051180	Wang et al. 1999
16p12.2	*PALB2*	ENSG00000083093	Erkko et al. 2007
16q22.1	*CDH1*	ENSG00000039068	Berx et al. 1995
17p13.1	*TP53*	ENSG00000141510	Borresen et al. 1992
17q21.33	*PHB*	ENSG00000167085	Jupe et al. 2001
17q23.2	*PPM1D*	ENSG00000170836	Li et al. 2002
17q23.2	*BRIP1*	ENSG00000136492	Cantor et al. 2001
22q12.1	*CHEK2*	ENSG00000183765	Walsh et al. 2006

3 Study Objectives

The present genome-wide association study (GWAS) was carried out with the following aims and objectives:

1. To study CNVs' contribution to genetic diseases using familial prostate and breast cancers as case studies
2. To assess and compare the performance of different CNV calling algorithms with the aim of knowing the degree of agreement between the different algorithms used in the study
3. To study heritability of CNVs in a bid to unravelling the genetic predisposition to cancer types in question.

4 Materials and Methods

4.1 Study Objects

4.1.1 HPC Families

In the present GWAS, 31 Hereditary Prostate Cancer (HPC) families consisting 102 cases, 33 controls, and 7 mothers from different families, were genotyped using Illumina HumanOmniExpress-12v1 genotyping microarray with approximately 700,000 markers (SNPs) per sample covering the entire genome. The families were evaluated by the number of affected individuals in the family and the number of relatives from whom blood sample was available for genotyping.

The families selected had at least three first or second-degree relatives affected and at least two affected individuals were genotyped from each family. Table 7 shows the characteristics of the 31 HPC families.

4.1.2 Breast Cancer Families

In the case of the breast cancer (BrCa) dataset, 84 and 36 cases and controls samples respectively were genotyped using Illumina HumanCytoSNP-12 genotyping microarray (HUMAN_CYTO_SNP-12V2) with approximately 300,000 markers (SNPs) per sample covering the entire genome. The samples have been tested negative for *BRCA1/BRCA2* mutation. Each of the cases and controls was selected from different families and as such, the BrCa data was treated as a case-control study.

LRR, BAF, channel X and Y intensities from each sample were exported from the normalized Illumina data through the GenomeStudio software (GSGTv1.7.4) to perform CNV identification.

4.2 Methods

4.2.1 Quality Control Measures

The genotypings using Illumina HumanOmniExpress (OmniExpress) and HumanCytoSNP-12 BeadChips for PrCa and BrCa samples respectively were done according to the manufacturer's protocol. Filtering criteria adopted by Singapore Genome Variation Project (SGVP) were applied to remove unsuitable samples (Teo et al. 2009b).

4.2.1.1 Pre-CNV Identification

This was carried out to remove unsuitable samples based on genotype call rates before CNV calling. All samples have call rates greater than 99.5% and thus were all included in the CNV calling.

4.2.1.2 Post-CNV Identification

Applying a set of filtering criteria as recommended by PennCNV, individuals not meeting at least one of the CNV specific quality control metrics were excluded from further analysis: LRR-Standard Deviation>0.25, 0.45>BAF-median>0.55, BAF-drift>0.002, and –0.04>Wave Factor >0.04 (Wang et al. 2007). Consequently, 81 cases and 35 controls samples from the BrCa data and all the samples from the PrCa data were suitable for this analysis.

4.2.2 CNV Identification and Construction of CNV Loci

Three sample-based algorithms and one segment-based algorithms were applied: cnvPartition [v3.1.6], PennCNV [2009Aug27] (Wang et al. 2007), QuantiSNP [v2.3] (Colella et al. 2007), and CNstream [v1.0] (Alonso et al. 2010).

The underlying statistical models for the four CNV identification algorithms differ by varying degrees. The primary raw data used for detecting CNVs from SNP arrays are the SNP intensity measured by LRR. Some methods also use BAF to enhance detection. CNstream on the other hand uses the channel X and Y intensities of each sample.

Table 7 Characteristics of the 31 HPC families

FamilyID	TotalAffected[a]	TotalGenotyped[b]	NumberCases[c]	NumberControls[d]
2015	7	4	3	1
2017	-	2	-	2
2033	4	3	3	-
*2062	7	5	4	-
2074	4	4	2	2
2145	4	4	3	1
2232	3	4	2	2
*2241	3	4	3	-
2248	4	3	2	1
2275	5	4	3	1
*2279	4	5	4	-
2283	7	3	2	1
2292	4	5	4	1
2308	6	5	5	-
2374	3	4	3	1
2375	6	7	5	2
2386	6	4	3	1
2394	5	4	3	1
2396	5	5	5	-
2399	4	5	3	2
2401	4	5	3	2
2414	3	5	3	2
2421	4	5	3	2
2427	6	6	5	1
*2429	4	6	3	2
2431	5	3	3	-
2435	4	3	3	-
2442	5	4	3	1
2449	8	8	7	1
*2450	4	5	3	1
**2455	5	8	4	2

[a]TotalAffected column (Total number of individuals affected in the family), [b]TotalGenotyped column (Total number of individuals genotyped from the family), [c]NumberCases column (Number of affected individuals genotyped), [d]NumberControls column (Number of unaffected individuals genotyped). Asterisked families are the families with genotyped mothers. The number of asterisks corresponds to the number of mother genotyped in the family.

cnvPartition, developed by Illumina, is available as a plug-in in the GenomeStudio software. It is based on the assumption that majority of the CNV vary between 0 and 4 copies, thus

yielding five options: homozygous deletion, heterozygous deletion, dizygous (normal state), trizygous (one extra copy), and tetrazygous (two extra copies). cnvPartition models LRR and BAF as a simple bivariate Gaussian distribution for each of the 14 possible genotypes.

PennCNV and QuantiSNP algorithms use different Hidden Markov Models (HMMs). The PennCNV uses the combined LRR and BAF of the SNP to infer CNVs, while the QuantiSNP treats them independently.

All the three sample-based algorithms utilized in this GWAS provide confidence score to allow the filtering of CNV and limit false positive calls. In the case of QuantiSNP, the confidence score is the Log Bayes Factor (LBF). The score is a measure of the likelihood that the region harbours an abnormal copy number. A score of 10 or larger has been suggested as a threshold to classify reliable CNV calls (Colella et al. 2007). PennCNV also provides similar score in term of confidence threshold. Consequently, a score of 10 was used as threshold for PennCNV and the same value was used for cnvPartition and QuantiSNP (from experience).

CNV regions identified by the sample-based algorithms were merged into discrete, non-overlapping loci with boundaries of each locus determined by the union of all CNV regions that belong to that particular locus, using "anyOverlap" criterion (Redon et al. 2006). In the event that both duplications and deletions were observed in a particular locus, two separate loci were identified for each form of CNV.

4.2.3 Case-Control Association Test

Having constructed the CNV loci, the proportion (number) of cases and controls harbouring a CNV at a particular locus was estimated and a fisher's exact test was carried out to obtain the p-value and odds ratio. In order to handle the exception of non-numerical p-value as a result of zero denominators in calculating odds ratio, VCD package was loaded and implemented on R to estimate numerical values of odds ratio. VCD handles this by adding a factor of 0.5 to both the denominators and numerators thus yielding a numerical value of odds ratio as against "Inf" returned by fisher's exact test.

In the case of CNstream, a status file was supplied and the association test was carried out. A status file is a plain text file where each line corresponds to the status of one sample (0 for

controls and 1 for cases). The samples must be sorted in the same way as the input signal intensity file. With this file supplied, CNstream performs a chi-square test for each CNP segment and includes in the results file some informative fields such as the p-value and odds ratio.

4.2.4 Novel CNV Loci

In order to identify novel CNV loci, non-overlapping CNV loci obtained in this study were compared with CNV loci published in the Database of Genomic Variants [DGV] (Iafrate et al. 2004). Latest version of DGV in Build 36 of the human genome – hg18 (variation.hg18.v10.nov.2010.txt) was downloaded from the DGV website (http://projects.tcag.ca/variation/). This was done because the genotyping was done with Build 36 of the human genome. A CNV locus was declared novel CNV if it does not share at least 50% of its length with any established CNV loci in the DGV database. The "scan_region.pl" script of the PennCNV was very resourceful in accomplishing this task. The novel CNVs identified in this study were used in the comparison to return a list of putative novel loci common to the two datasets.

4.2.5 Mapping against Annotated Genes and Disease-Associated CNV Loci

In order to identify genes that are located within or partially overlap with the CNV loci, the CNV loci were queried for overlap against Refseq genes annotation. In the case of intergenic CNVs, the loci were expanded both upstream and downstream to identify the neighbouring genes with the corresponding distance.

To identify loci that warrant further investigation for their roles in complex disease, identified loci were queried for overlap against the genes listed in the Online Mendelian Inheritance in Man (OMIM) Morbid Map (http://www.ncbi.nlm.nih.gov/omim/).

4.2.6 Enrichment Analysis

Enrichment analysis was done using Gene Set Analysis ToolkitV2 – *webGestalt* (http://bioinfo.vanderbilt.edu/webgestalt/) provided by Bing Zhang (Zhang 2005; Duncan 2010). The following enrichment analysis were carried out: Gene Ontology (GO), KEGG pathways, Pathway Commons, and Wikipathways.

The Gene Ontology project (The Gene Ontology Consortium 2008) is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. The Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains: cellular component, the parts of a cell or its extracellular environment; molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

The KEGG Pathway Analysis component can be used to find clusters of co-expressed genes sharing the same pathway. KEGG, which stands for Kyoto Encyclopedia of Genes and Genomes, has become a major resource for pathway analysis and contains a wealth of data associated with pathways, genes, genomes, chemical compounds and reaction information, in addition to links to outside resources such as PubMed (Kanehisa et al. 2006).

The Pathway Commons (Cerami et al. 2006) ontology contains data on pathways from multiple sources. Pathways include biochemical reactions, complex assembly, transport and catalysis events, and physical interactions involving proteins, DNA, RNA, small molecules and complexes.

WikiPathways (Pico et al. 2008) is an open, collaborative platform dedicated to the curation of biological pathways It provides a graphical pathway editing tool and integrated databases covering major gene, protein, and small-molecule systems.

Due to the size and nature of the data, the gene lists used in gene ontology were generated using varying criteria: odds ratio greater than 1 and 2. This was done in order not to lose

meaningful information since statistical significance does not guarantee biological significance.

4.2.7 CNV Mapping

Ensembl's karyograph tool was used in generating the karyograph of the identified CNVs (http://may2009.archive.ensembl.org/Homo_sapiens/Location/Genome/). This was the latest version in Build 36 of the human genome.

In addition to the methods described above, a family-based analysis was carried out on the PrCa dataset. In this analysis, CNVs overlapping genomic regions were analysed for enrichment in certain families. Using the total number of cases in the family, number of cases (affected individuals) genotyped in the family and number of cases harbouring the variation (CNV), the percentage of case CNV in each family was estimated. Enrichment was declared if at least 50% of the total number of cases in the family and/or cases genotyped harbours the variation.

5 Results

5.1 Characteristics CNV Regions and Loci

In this genome-wide analysis study, both CNVs and INDELs (insertions and deletions < 1kb) were obtained in PrCa dataset. The study focuses only on CNVs; however, the INDELs were also noted.

After applying series of CNV quality filtering criteria as recommended by the different algorithms on the PrCa dataset, a total of 544, 639, 509 and 385 CNVs with a median size of 15.0kb, 13.8kb, 23.8kb and 22.1kb were detected by PennCNV, QuantiSNP, cnvPartition and CNstream respectively (Figures 9). Majority of the CNVs detected in this GWAS were distributed within 10-50kb, however, the individual-based CNV calling programs (PennCNV, QuantiSNP) with the exception of cnvPartition have a larger number of their variants within 1-10kb. Figure 5 and Table 8 show the frequency distribution of CNV sizes as detected by the various algorithms in PrCa dataset.

Merging the CNV loci detected by the sample-based CNV calling algorithms yielded a total of 764 non-overlapping loci of which 51.3% (392/764) overlap with RefSeq genes. CNstream detected only 32.1% (245/764) and 30.4% (119/392) of the total CNVs and CNVs overlapping RefSeq genes respectively.

In the case of BrCa dataset however, a total of 273, 295, 211 and 404 CNVs with a median size of 50.5kb, 55.5kb, 90.4kb and 35.3kb were detected by PennCNV, QuantiSNP, cnvPartition and CNstream respectively (Figure 10). Majority of the CNVs identified in this GWAS were distributed within 10 - 50kb. Figure 6 and Table 9 show the distribution of CNV sizes detected by various algorithms in the BrCa dataset.

Comparing and merging the CNV loci from PennCNV, QuantiSNP and CNstream, a total of 359 non-overlapping loci were obtained with 59.6% (214/359) overlapping with Refseq genes. CNstream detected only 27.3% (98/359) and 25.2% (54/214) of the total merged CNVs and CNVs overlapping Refseq genes respectively. The ratio of deletions to duplications is approximately 3:2 and 1:1 for PrCa and BrCa datasets respectively (see Table 10, Figures 11 and 12). The majority of the individuals have 17 - 22 and 4 - 7 CNVs in the PrCa and BrCa datasets respectively (Table 10).

Table 8 Frequency distribution of CNV sizes in prostate cancer dataset

	PennCNV	QuantiSNP	cnvPartition	CNstream
Size (kb)	(%)	(%)	(%)	(%)
>1-10kb	218 (40.1)	270(42.3)	150(29.5)	70(18.8)
>10-50kb	189(34.7)	212(33.2)	192(37.8)	255(66.2)
>50-100kb	59(10.8)	65(10.2)	55(10.8)	43(11.2)
>100-150kb	29(5.3)	26(4.1)	31(6.1)	11(2.9)
>150-200kb	16(2.9)	17(2.7)	18(3.5)	3(0.8)
>0.2-1Mb	31(5.7)	47(7.4)	56(11.0)	3(0.8)
>1Mb	2(0.4)	2(0.3)	7(1.4)	0(0.0)
TOTAL	544	639	509	385

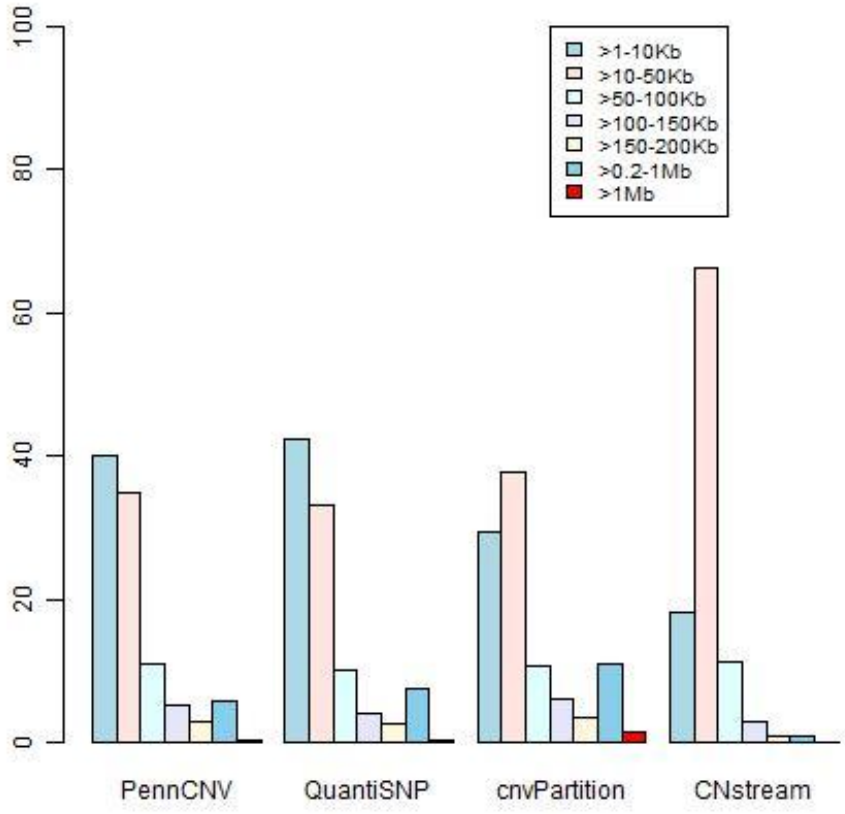

Figure 7 Bar chart of CNVs size distribution in prostate cancer

Table 9 Frequency distribution of CNV sizes in breast cancer dataset

	PennCNV	QuantiSNP	cnvPartition	Cnstream
Size (kb)	(%)	(%)	(%)	(%)
>1-10kb	14(5.1)	32(10.8)	6(2.8)	9(2.2)
>10-50kb	122(44.7)	106(35.9)	63(29.9)	273(67.6)
>50-100kb	62(22.7)	57(19.3)	40(18.9)	104(25.7)
>100-150kb	24(8.8)	22(7.5)	17(8.1)	16(3.9)
>150-200kb	16(5.9)	17(5.8)	13(6.2)	2(0.5)
>0.2-1Mb	33(12.1)	57(19.3)	66(31.3)	0(0.0)
>1Mb	2(0.7)	4(1.4)	6(2.8)	0(0.0)
TOTAL	273	295	211	404

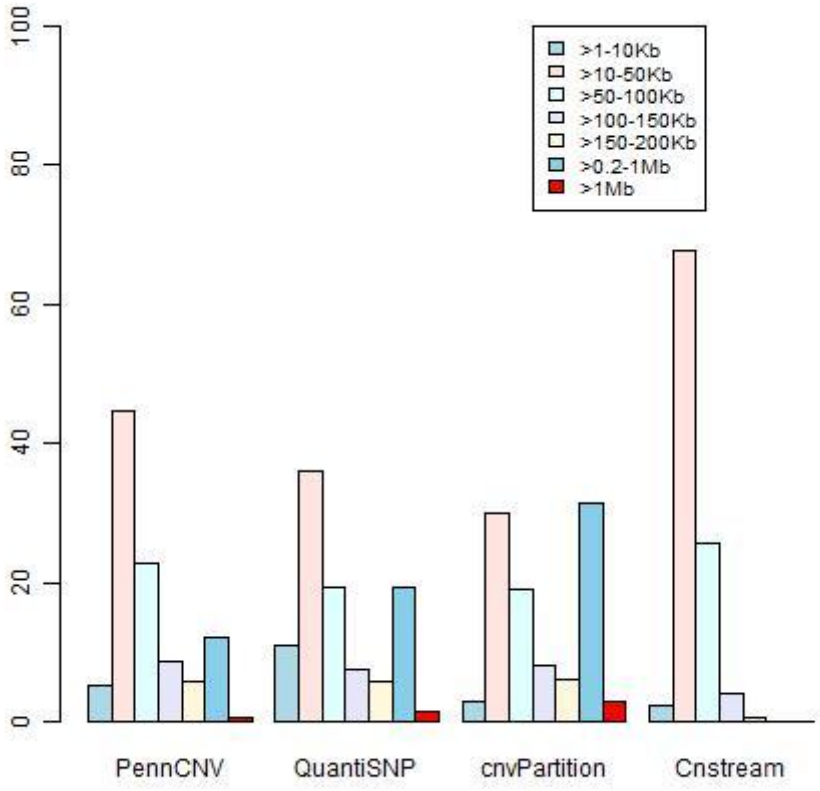

Figure 9 Bar chart of CNVs size distribution in prostate cancer

Table 10 Summary statistics of CNVs as detected by the different algorithms in the two datasets.

	ALGORITHM					
PARAMETER	PennCNV		QuantiSNP		cnvPartition	
CNV	PrCa	BrCa	PrCa	BrCa	PrCa	BrCa
Total Number	2724	545	3033	802	2397	472
Avg. No of CNVs/sample	19.2	4.7	21.4	6.9	16.9	4.1
Avg. Size of CNV (Kb)	44.1	86.5	57.4	114.6	79.2	187.6
Median size of CNVs (Kb)	12.4	51.5	13.8	57.3	19.8	89.3
Number of gain	846	245	888	345	652	206
Number of loss	1878	300	2145	457	1745	266
Ratio (Loss/Gain)	2.2	1.2	2.4	1.3	2.6	1.3
CNV region						
Total number	544	273	639	295	509	211
Avg. No of CNVs/sample	3.8	2.4	4.5	2.5	3.5	1.8
Avg. Size of CNVs (Kb)	57.8	99.9	64.3	135.2	77.4	206.9
Median size of CNVs (Kb)	15.1	50.5	13.8	55.5	23.8	90.4
Number of gain	205	142	211	144	191	107
Number of loss	339	131	428	151	318	104
Ratio (Loss/Gain)	1.7	0.92	2	1.04	1.7	0.97
Common CNVs (frequency)						
Freq.>1% (%)	371(68.2)	72(26.4)	394(61.7)	98(33.2)	327(64.2)	65(30.8)
Freq.>2.5% (%)	197(36.2)	42(15.4)	209(32.7)	65(22.0)	180(35.4)	42(19.9)
Freq.>5% (%)	96(16.4)	18(6.6)	105(16.4)	28(9.5)	89(17.5)	21(9.9)
CNVs (OR>1.5) (%)	138(21.8)	31(11.4)	139(21.8)	36(12.2)	118(23.2)	23(10.9)

CNV, copy number variations. PrCa, prostate cancer (n=142). BrCa, breast cancer (n=116)

5.2 Novel CNVs

59.03% (451/764) and 14.79% (113/764) of the CNVs identified in PrCa overlap with CNVs reported in DGV at "anyOverlap" and 50% query-database-ratio criteria (-minquerydbratio 0.5) respectively. Swapping these values implies that 40.94% (313/764) and 85.21% (651/764) are novel CNVs at "anyOverlap" and 50% query-database ratio criteria respectively. 47.28% (148/313) and 47.31 (308/651) of the "anyOverlap" and 50% query-database ratio novel CNV loci respectively are genomic.

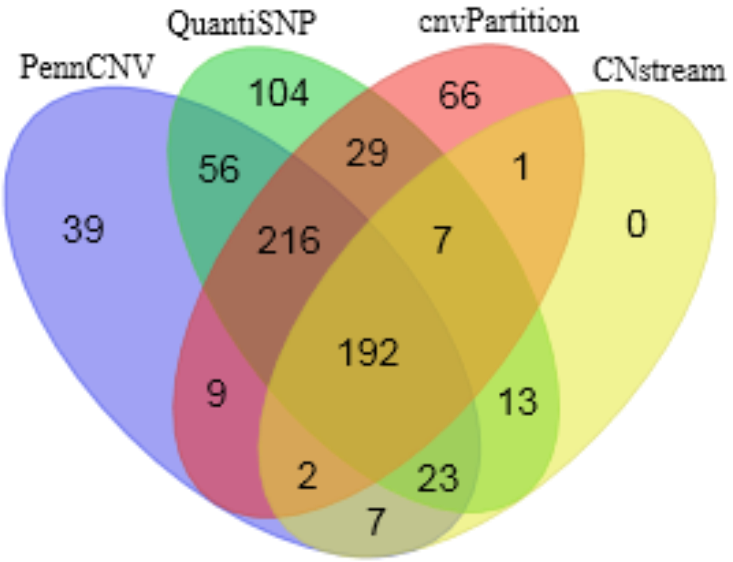

Figure 9 Venn diagram illustrating the agreement of the algorithms in detecting CNVs in the prostate cancer dataset.

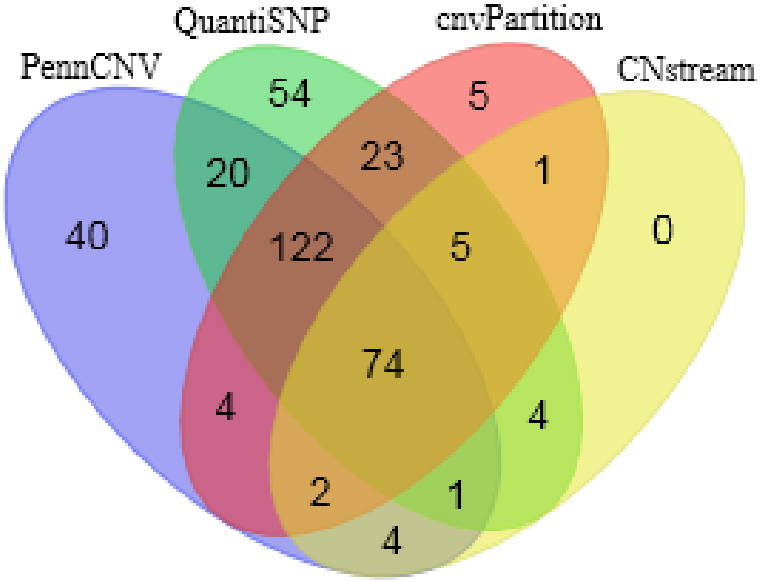

Figure 10 Venn diagram illustrating the agreement of the algorithms in detecting CNVs in the breast cancer dataset.

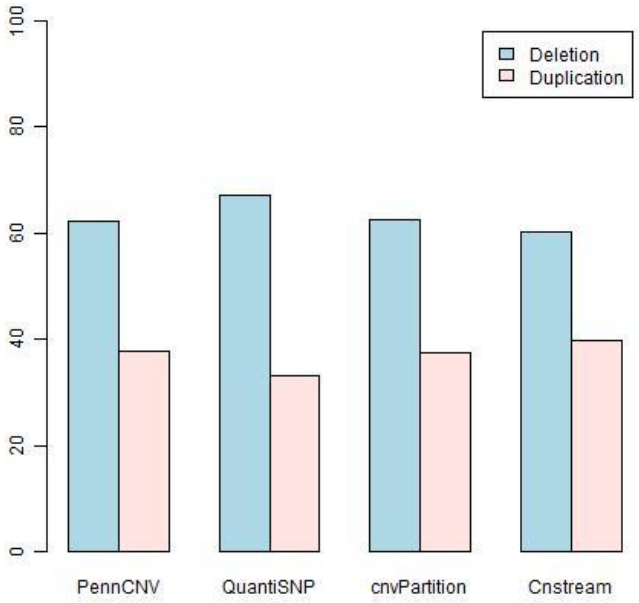

Figure 11 Proportion of deletions and duplications as detected by the different algorithms in prostate cancer

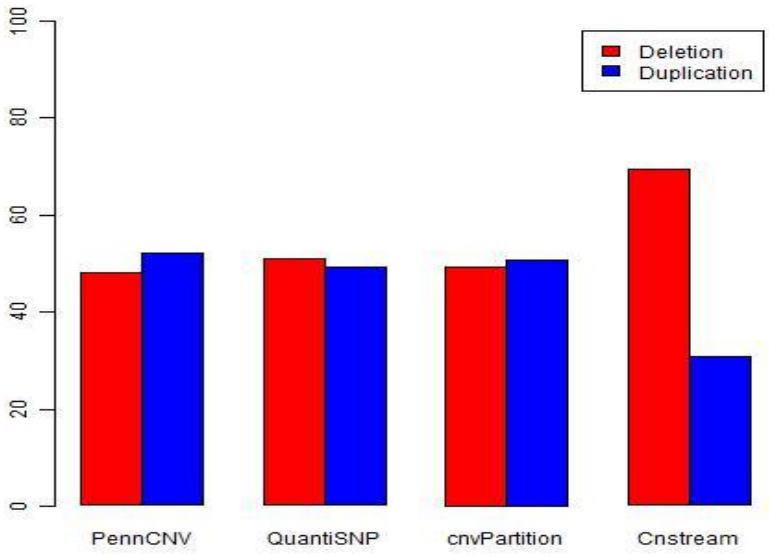

Figure 12 Proportion of deletions and duplications as detected by the different algorithms in breast cancer

For the BrCa dataset, however, 85.5% (307/359) and 23.1% (83/359) of the identified CNVs overlap with DGV loci at "anyOverlap" and 50% query-database ratio criteria respectively. Swapping these values implies that 14.5% (52/359) and 76.88% (276/359) are novel CNV loci at the two criteria respectively. 53.8% (28/52) and 57.2% (158/276) of the "anyOverlap" and 50% query-database ratio novel loci respectively are genomic. Figures 13 and 14 show the proportion of the overlap of the identified CNV loci with the DGV loci as detected by the different algorithms for the two datasets.

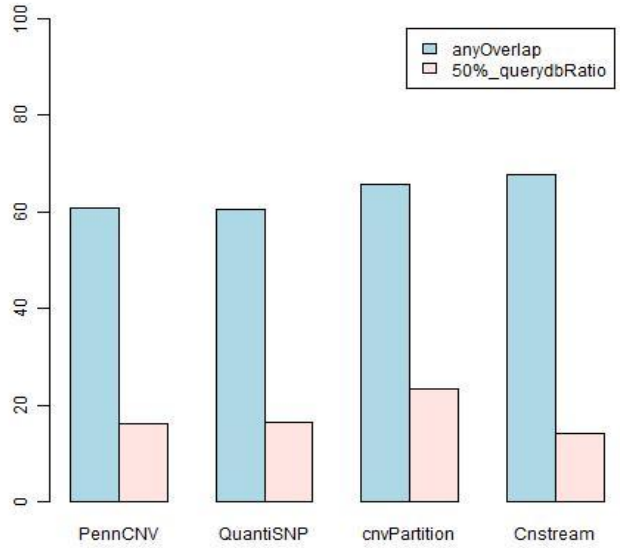

Figure 13 Proportion of CNVs in PrCa dataset already reported in DGV

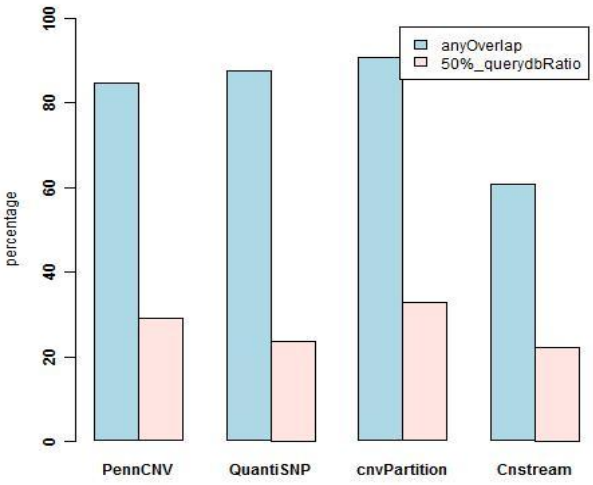

Figure 14 Proportion of CNVs in BrCa dataset already reported in DGV

Table 11 Common Novel CNV Region in the two datasets

Region	Locus	RefSeq Gene	Event	Dataset
chr1:1647686-1670079	1p36.33	*SLC35E2*	Dup	Both
chr1:1647686-1661642	1p36.33	*SLC35E2*	Del	BrCa only
chr1:12598434-12615712	1p36.22-36.21	*DHRS3*	Dup	Both
chr1:12698721-12743160	1p36.21	*AADACL3,C1orf158*	Dup	Both
chr1:12871327-12874361	1p36.21	Intergenic	Del	Both
chr1:49930749-49964737	1p33	*AGBL4*	Del	Both
chr2:49535856-49537795	2p16.3	Intergenic	Del	Both
chr2:76941049-76949101	2p12	*LRRTM4*	Del	Both
chr2:87428677-87730750	2p11.2	MIR4435-1,MIR4435-2,*NCRNA00152*	Dup	Both
chr2:87482067-87633978	2p11.2	*NCRNA00152*	Del	BrCa only
chr2:212404169-212412738	2q34	*ERBB4*	Del	Both
chr3:8826396-8832963	3p25.3	Intergenic	Dup	Both
chr5:32144879-32159517	5p13.3	*PDZD2*	Dup	Both
chr5:97075236-97099320	5q15	Intergenic	Del	Both
chr6:95424156-95578465	6q16.1	Intergenic	Dup	BrCa only
chr6:95424156-95578465	6q16.1	Intergenic	Del	Both
chr7:9097947-9102563	7p21.3	Intergenic	Del	Both
chr7:12781543-12788046	7p21.3	Intergenic	Del	Both
chr7:29678716-29687522	7p15.1	LOC646762,MIR550A3	Del	PrCa only
chr7:29678716-29687522	7p15.1	LOC646762,MIR550A3	Dup	Both
chr7:57495829-57524352	7p11.1	*ZNF716*	Dup	Both
chr7:61852895-62326882	7q11.21	Intergenic	Dup	Both
chr7:62035570-62047108	7q11.21	Intergenic	Del	Both
chr7:62154874-62159926	7q11.21	Intergenic	Del	Both
chr7:76432653-76453285	7q11.23	*PMS2P11*	Dup	Both
chr8:16010913-16021468	8p22	*MSR1*	Del	PrCa only
chr8:16010913-16021468	8p22	*MSR1*	Dup	BrCa only
chr9:196132-234457	9p24.3	*C9orf66,DOCK8*	Dup	Both
chr10:20850624-20857365	10p12.31	Intergenic	Dup	PrCa only
chr10:20850624-20857365	10p12.31	Intergenic	Del	Both
chr10:68078481-68091312	10q21.3	*CTNNA3*	Del	Both
chr10:90944216-90945756	10q23.31	Intergenic	Del	Both
chr12:8000336-8014573	12p13.31	Intergenic	Dup	Both
chr12:31266287-31292645	12p11.21	Intergenic	Dup	Both
chr15:22299434-22320561	15q11.2	Intergenic	Del	Both
chr15:32509892-32514341	15q14	*GOLGA8A*	Dup	PrCa only
chr15:32509892-32595143	15q14	*GOLGA8A*	Del	Both
chr16:16098032-16162264	16p13.11	*ABCC1,ABCC6*	Dup	Both
chr16:28733106-28825145	16p11.2	*ATP2A1,ATXN2L,MIR4721,RABEP2,SH2B1,TUFM*	Dup	Both
chr16:32137965-32165782	16p11.2	Intergenic	Del	Both
chr16:32511914-32648969	16p11.2	LOC653550,*TP53TG3,TP53TG3B*	Del	Both
chr16:32511914-32648969	16p11.2	LOC653550,*TP53TG3,TP53TG3B*	Dup	BrCa only
chr19:20664930-20715228	19p13.12	Intergenic	Del	Both

5.3 Mapping against Annotated Genes and Disease-Associated CNV Loci

51.3% (392/764) of the total CNVs identified in the PrCa dataset overlap with RefSeq genes. For the BrCa dataset on the other hand, 59.6% (214/359) overlap with RefSeq genes. 126 and 108 genes overlapping with RefSeq genes in the PrCa and BrCa datasets respectively are found in the OMIM genes and produce 55 and 36 disorders (phenotypes) respectively (Tables S1 and S2 in the supplementary attachments). Figure 15 shows a deletion at *MSR1* locus found in a family.

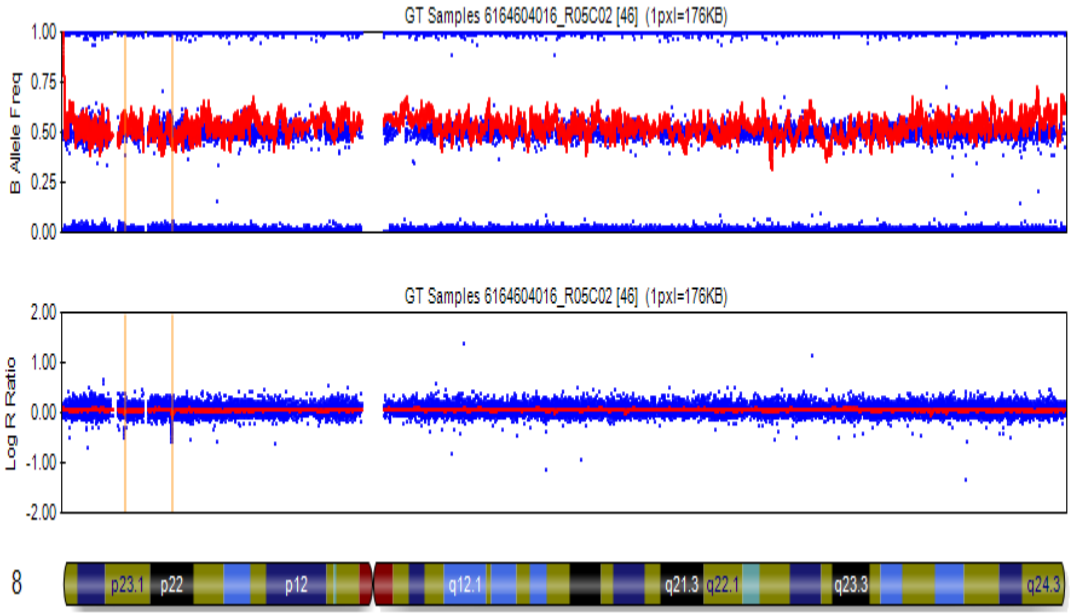

Figure 15 Deletion at MSR1 (8p22) locus found in a family

5.4 Enrichment Analysis

Various GO terms and pathways were found enriched in the two datasets with varying criteria used in generating the gene lists for the ontology. Some ontologies with p-values greater than 0.05 were included in the table because of their role in cancer pathways.

With these criteria, a total of 224 and 85 RefSeq genes with odds ratio greater than 1 and 2 respectively were included in the GO analysis from the PrCa dataset. Out of these, *WebGestalt* identified a total of 195 and 79 genes with unique user Entrez IDs for the two criteria respectively; consequently, enrichment analyses were based on 195 and 79 genes respectively (see Tables S3 and S4 in the supplementary attachments).

For the BrCa dataset, however, a total of 220 out of 407 Refseq genes were included in the GO analysis. Out of these, *WebGestalt* identified a total of 194 Unique User Entrez IDs; consequently, enrichment analyses were based on 194 genes (see Tables S3 and S4 in the supplementary attachments).

5.5 Case-control Association Test

Different algorithms yielded different association results in term of coordinates and p-value; however, there are similarities in the cytogenetic bands or loci.

PennCNV detected 10q11.22 (chr10:47543322-47703869) with a p-value of 0.0156; QuantiSNP detected 3p26.1 (chr3:6649648-6654060), 10q11.22 (chr10:47049547-47940417) and 2p25.3 (chr2:4213378-4222144) with a p-value of 0.016, 0.021 and 0.026 respectively; cnvPartition detected 3p26.1 (chr3:6649648-6654060) and 10q11.22 (chr10:47109571-47703869) with a p-value of 0.0213 and 0.0213 respectively.

A case-control association test with CNstream yielded five consecutive segments with significant p-values. These segments correspond to 2p25.3 (chr2:4211781-4228747) with a p-value and odds ratio of 0.035835 and 6.857143 respectively. This same locus was detected by QuantiSNP but with a different p-value. Mention must however be made that majority of the associated loci are intergenic with the exception of 10q11.22 that harbours some genes.

In the case of the BrCa dataset, 9.88% (8/81) of cases have intronic deletions at *EPHA3* locus (chr3:89485137-89499754- 3p11.1) with a p-value of 0.050628, which is slightly higher than 0.05. Mention must however be made that none of the controls have alterations at this locus.

Table 12 Summary of association result

	p-value			
locus	PennCNV	QuantiSNP	cnvPatition	Cnstream
2p25.3	-	0.026	-	0.035
3p26.1	-	0.016	0.0213	-
10q11.22	0.015	0.021	0.0213	-

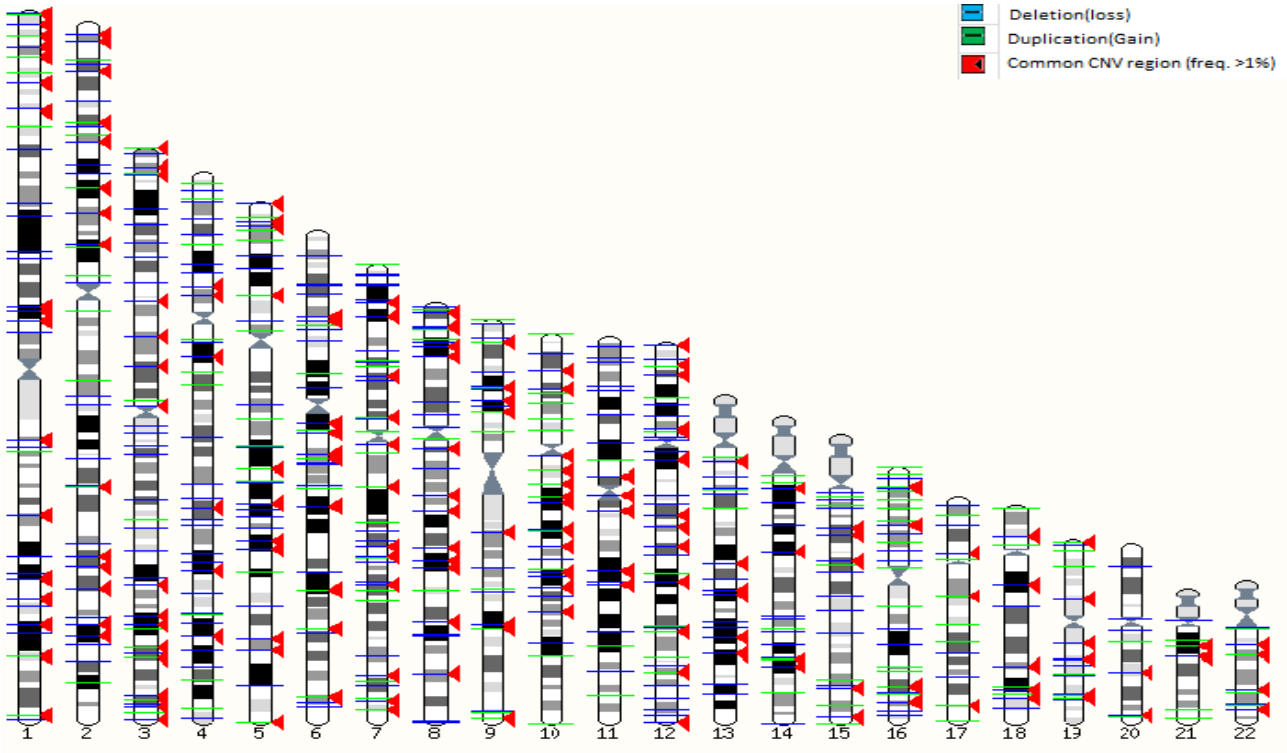

Figure 16 Map of Identified CNV Loci in Prostate Cancer Dataset

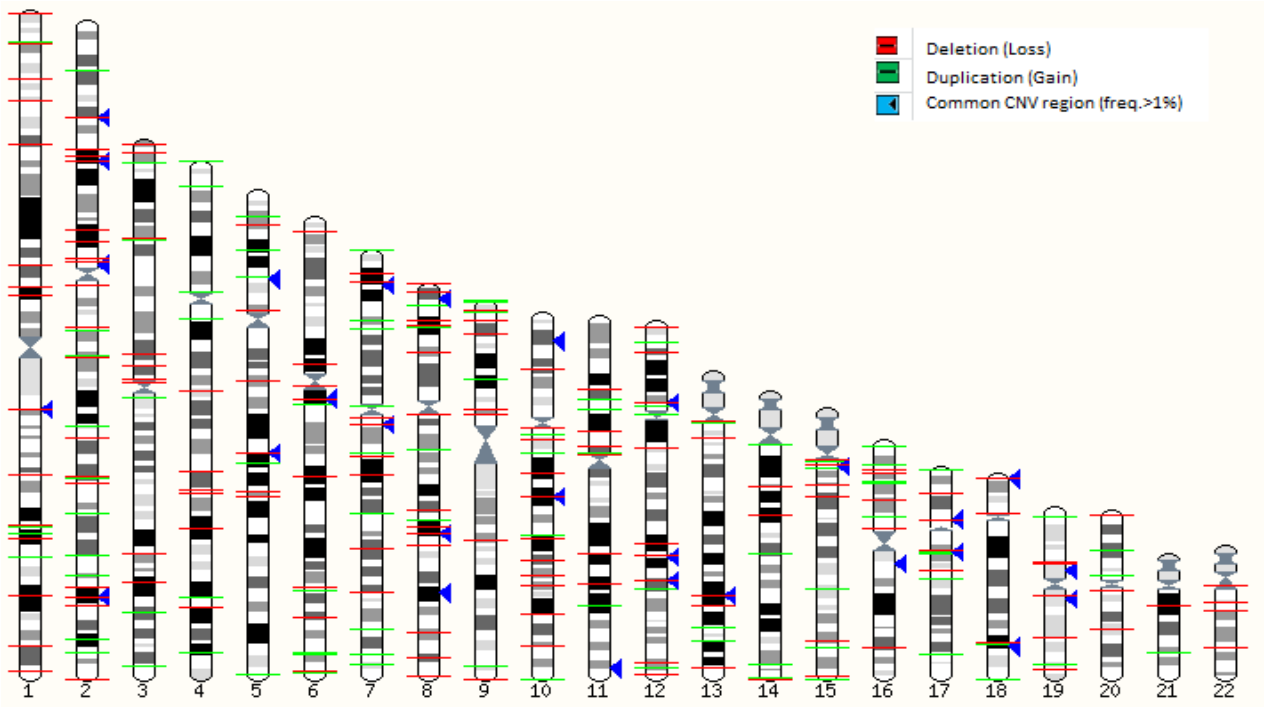

Figure 17 Map of Identified CNV Loci in Breast Cancer Dataset

5.6 CNV maps

Few CNPs were observed in the BrCa dataset as against the PrCa dataset that contains several common CNVs on different chromosomes. Although most studies on heritability of CNVs were done on trios comprising the father, mother and a child from the family; this probably suggests that closely related individuals have higher heritability than distant individuals. The BrCa families (cases) were chosen from different unrelated family members as against the PrCa families (cases) with at least two members from the same family. Consequently, more CNPs were observed in the PrCa dataset (see Figures 16 and 17).

The family-wise analysis on the PrCa dataset shows significant enrichment of some CNVs in certain families (Table 13).

Table 13 Family-wise analysis.

Locus	Gene	Event	% TAF	% TCGF	Family ID
1p36.22	*SPSB1*	Del	83.3	100	2427
1p36.11	*RHCE,TMEM57*	Del	50 - 66.7	66.7 - 80	2308, 2399,
1q21.2	*ARNT,CTSK,SETDB1*	Del	40 - 75	66.7 - 100	2062, 2374, 2442, 2145,2279
1q31.3	*PTPRC*	Dup	50	60	2308
1q32.1	*SLC45A3*	Del	50	66.7	2399
1q32.2	*PLXNA2*	Del	100	100	2374
2p22.1	*MORN2*	Dup	60 - 75	75	2455, 2292
2p12	*LRRTM4*	Del	60	100	2275
2q34	*ERBB4*	Del	37.5 - 50	42.8 - 66.7	2421, 2455, 2449
2q35	*ABCA12*	Del	50	66.7	2429
3p22.2	*CTDSPL,MIR26A1*	Del	40 - 50	66.7	2394, 2429, 2442
3p21.1	*SFMBT1*	Del	60 - 66.7	60 - 66.7	2241, 2396
3p11.1	*EPHA3*	Del	50	60 - 66.7	2421, 2427
3q26.1	*PPM1L*	Del	50	60	2427
3q28	*TP63*	dcDel	75	75	2279
4p15.31	*GPR125*	Del	100	100	2414
4p13	*GRXCR1*	Del	66.7 - 75	66.7 - 100	2145, 2414
4q27	*TNIP3*	Del	40 - 50	66.7	2145, 2442
4q31.3	*FAM160A1*	Del	40 - 60	66.7 - 100	2431, 2442
5p15.2	*SEMA5A,SNORD123*	Dup	66.7	80	2375
5p15.2	*LOC285692*	Dup	66.7	80	2375
5p13.3	*PDZD2*	Dup	50 - 100	66.7 - 100	2074, 2145, 2232
5q31.3	*PCDHA11*	Del	50 -100	60 - 100	2414, 2427
6p22.1	*HCG4,LOC554223*	Del	60	100	2442
6p21.33	*CCHCR1*	dcDel	75	100	2429
6q22.2	*SLC35F1*	Del	75	100	2401

Table 13 Family-wise analysis (continued).

7q11.23	*PMS2P11*	Dup	100	100	2241
7q22.1	*MUC17*	Del	75	100	2421
7q31.1	*IMMP2L,LRRN3*	Del	66.7	80	2427
7q31.1	*DOCK4*	Del	75	75	2279
7q36.1	*ABCF2*	Del	60	100	2394
7q36.2	*DPP6*	Dup	50	100	2386
8p23.2	*CSMD1*	Del	50	66.7	2401, 2435
8p23.1	*TNKS*	Dup	50	50	2292
8p22	*MSR1*	Del	50	50	2279
8p21.3	*PEBP4*	Dup	66.7	66.7	2414
8p21.2	*ADRA1A*	Dup	66.7	66.7	2414
8p12	*NRG1*	Dup	100	100	2414
8q21.3	*OTUD6B*	Del	75	100	2429
8q22.2	*OSR2*	Dup	50	50	2279
10p15.3	*ZMYND11*	Dup	75	75	2292
10q11.21	*ALOX5,MARCH8*	Dup	50	60	2427
10q21.3	*CTNNA3*	Del	50	50	2292
10q23.1	*NRG3*	Del	40 - 75	66.7 - 100	2435, 2431
10q23.1	*NRG3*	Dup	50	66.66667	2421
10q23.33	*CYP2C19*	Del	50	60 - 66.7	2399, 2427
12p13.33	*CACNA1C*	Del	75	100	2429
12q21.31	*ACSS3*	Dup	100	100	2232
12q21.31	*MIR548T*	Del	50	57.1	2449
12q23.1	*UHRF1BP1L*	Del	75	100	2399
12q23.2	*ARL1,SPIC,UTP20*	Dup	75	75	2279
12q24.33	*ANKLE2,PGAM5, POLE,PXMP2*	Del	66.7	66.7	2374
13q13.1	*KL*	Del	50 - 66.7	50 - 66.7	2241, 2279
13q21.33	*KLHL1*	Del	50 - 75	66.7 - 100	2421, 2427, 2442, 2435
14q11.2	*ANG,EDDM3A, EDDM3B,RNASE1, RNASE4,RNASE6*	Dup	50	50	2292
14q21.3	*MDGA2*	Del	75	75	2279
15q12	*GABRA5*	Dup	50	66.7	2435, 2421
15q14	*GOLGA8A*	Del	75	100	2399
16p12.3	*GPR139*	Del	50	50	2292
16p11.2	*EIF3C*	Dup	60 - 66.7	100	2248, 2442
17p11.2	*ALDH3A2,SLC47A2*	Del	50	66.7	2421, 2429
18p11.21	*FAM38B*	Del	100	100	2414
18q12.1	*FAM59A,MEP1B*	Dup	50	100	2386
19p13.3	*REXO1*	Del	50	60	2427
19p12	*ZNF626*	Del	50	66.7 - 100	2074, 2379
19q13.32	*EMP3,TMEM143*	Dup	50	50 - 60	2292, 2375
21q22.3	*TMPRSS2*	Dup	66.7	66.7	2414

Del – deletion; Dup – duplication; dcDel – double copy deletion; %TAF – percentage of total affected individuals in the family; %TCGF – percentage of the total cases (affected individuals) genotyped in the family

6 Discussion and Conclusion

Recently, several studies have reported CNVs as relevant contributors to human diversity and disease susceptibility including cancer (Sebat et al. 2004; Sharp et al. 2005; Feuk et al. 2006). With the recent advancements on SNP array technology, it is now possible to detect previously elusive genetic variations with high resolution.

In the present GWAS, four CNV identification algorithms were used in assessing copy number variation in the Finnish familial prostate and female breast cancers genotyped using Illumina HumanOmniExpress-12v1 with approximately 700,000 markers (SNPs) per sample and Illumina HumanCytoSNP-12 with approximately 300,000 markers per sample respectively. By default, the degree of overlapping of the SNPs in the HumanOmniExpress-12v1 platform is higher than in the other platform, however, both platforms cover the whole genome. Consequently, INDELs were identified in the prostate cancer dataset, though the study focuses on CNVs. More CNVs were observed in prostate cancer than in breast cancer with an average of 20 CNVs and 5 CNVs per sample respectively. Average size of CNVs in breast cancer is twice as large as the size obtained in the prostate cancer dataset. However, the ratio of deletions to duplication is 3:2 in the prostate cancer dataset as against approximately 1:1 in the breast cancer dataset (see Table 10). This discrepancy could be due to different genotyping platforms as well as experimental set up. However, it is not certain that the factors mentioned above are the primary factors responsible for the differences. To date, no work has been done comparing copy number variations in males and females, and until it is proven otherwise, the present study provides evidence that there are more structural variations (CNV) in male genomes than in the female genome. The study also provides evidence that the female genome is more stable than the male genome in term of deletions, though; the size of such events is larger in the female genome.

On the agreement of the algorithms, about 54% and 70% of the CNVs identified in the prostate cancer were detected by at least three and two of the individual-based algorithms respectively and about 30% (231/764) detected by only one of the three algorithms. 20% (47/231), 51% (118/231) and 28% (66/231) of the CNVs identified by only one of the individual-based algorithms were detected by PennCNV, QuantiSNP, and cnvPartition respectively. 11% (13/118) of the CNVs identified by QuantiSNP only were identified by CNstream (an algorithm for the detection and identification of copy number polymorphisms). In the case of cnvPartition, 10.6% (7/66) of the CNVs are CNPs. For PennCNV, however,

only 0.02% (1/47) of the CNVs identified by PennCNV only is a CNP (see Figure 7). A simple majority vote of 2/3 would have help to handle this in a more robust manner, but it might result in loss of biologically meaningful variants.

For the breast cancer however, 54.5% (196/359), 70% (251/359) and 30% (108/359) of the total CNVs identified in the breast cancer dataset were detected by three, two and only one of the three individual-based algorithms used in the study. The same statistics in term of perform and agreement of the algorithms was obtained in the two datasets (see Figure 10).

In term of performance, QuantiSNP performed best among the three algorithms sample-based algorithms. The three susceptibility loci identified in this study were identified by QuantiSNP. cnvPartition detected only two while PennCNV identified only one locus. Each of the loci was detected by at least two of the four algorithms used in the study. The susceptibility locus at 2p25.3 was detected by both QuantiSNP and CNstream. However, none of the other two sample-based algorithms detected this locus. The fact that the locus was detected by CNstream is a simple prove to justify that it is not a false positive. cnvPartition performs better in detecting CNVs of large sizes. The degree of agreement between the three sample-based algorithms is only about 55% and about 70% between any two algorithms (see Figures 9 and 10). This helps to justify the need for experimental (laboratory) validation of the CNVs detected by the algorithms.

Breast Cancer

The lower frequency of mutations at *BRCA1*/*BRCA2* loci in the Finnish and Southern Swedish populations as against the statistics obtained in other parts of the world probably suggests that there are susceptibility loci that could be peculiar to the Nordic population and are yet to be discovered. It was hypothesized that if *BRCA1*/*BRCA2* mutations could account for only about 30% of the breast cancer cases, then *BRCAX* should exist. In the present study with 84 *BRCA1/BRCA2*-mutation negative familial breast cancer cases, *BRCA1* variation was found in 0.012% (1/81) despite the fact that the families (cases) are negative for *BRCA1/BRCA2* mutations. This variation involves the deletion of exons 1 – 12 of *BRCA1* in the sample and was detected by all the four algorithms. The fact that it was detected by the four algorithms is a proof that the event is not a false positive.

CNstream's, a method for the detection of copy number polymorphisms (CNP), result clearly reveals that the samples are not closely related since they share only 29 common loci. Majority of the common loci have a frequency of 1.7% (2/116). The enrichment analysis (pathway commons) shows that both *BRCA1* and *BCL2L1* function in ATM-mediated response to DNA double-strand break. *BCL2L1* with some β-defensins was amplified at 20q11.21 locus in one sample. Ontology results show that *EPHA3* and the β-defensins have common function. Deletions at chromosome 8p locus harboring some β-defensins have been associated with prostate cancer (Klaus et al. 2008). The p-arm of chromosome 8 is frequently deleted and associated with disease progression in human cancers, including breast cancer (BrCa). The present study shows the deletion of *TUSC3* locus and amplification of *MSR1* locus in different individuals. Deletions at *MSR1* however, have been associated with hereditary prostate cancer. *EPHA3*, *MSR1*, *ERBB4* (close member of the *ERBB2* family) and *GPR142* together with other genes shown on the enrichment results, all have receptor activities, yet none of the loci is statistically significant in the study.

It could therefore be that multiple factors are responsible for the unexplained proportion of BrCa cases in the population in question. However, it could also be that the data does not support the hypothesis. Moreover, with larger sample sizes, it is possible to obtain significant statistics.

Prostate Cancer

The present GWAS, aimed at unraveling genetic predisposition to prostate cancer, identified three loci associated with prostate cancer: 2p25.3, 3p26.1 and 10q11.22; with each loci identified as being significant by at least two of the four algorithms used in the study. However, no gene was found in the loci.

2p25.3 deletion was found in 17.6% (18/102) of the cases and it involves six consecutive SNPs (rs1175867 – rs1175854). It spans 8,767bp and overlaps with known CNV in the Database of Genomic Variant (DGV) at "anyOverlap" criterion only, but with a higher frequency. The functional effect of this CNV is not clear because no known gene resides in the region of the deletion. However, a hypothetical protein LOC727982 is found located within 450kb of this CNV. At the moment, little is known about this protein.

Liu reported a deletion at 2p24.3 associated with aggressive prostate cancer (Liu et al. 2009). Like 2p25.3, no known gene was found in this locus and it also spans six SNPs. The significance of both loci is yet to be understood as both harbour no known gene. This calls for further research in the p-arm of chromosome 2.

3p26.1 deletion was found in 13.7% (14/102) of the cases and it involves four SNPs (rs1043364 – rs1704538). It does not overlap with any known CNV in the DGV at the two criteria used in this study. *GRM7* is located approximately 250kb away from the CNV. *GRM7* has been associated with age-related hearing impairment (Friedman et al. 2009). No clear association has been established between hearing and prostate cancer. Thus, the functional effect of this variation could not be established yet. Mention must however be made that 3p26 has earlier been associated with hereditary prostate cancer (HPC) (Rokman, Ikonen, et al. 2001).

600kb amplification at 10q11.22 was observed in 12.7% (13/102) of the cases and it involves several genes, yet none of them is known to play an important role in tumorigenesis. 10q11.22 together with 2p25.3 has not been reported previously.

Deletion at 1q21.2 (chr1:149039930-149060682) was found in 66-100% of the total cases genotyped and 42-75 % of the total cases in the family in five different families. This deletion involves *ARNT* and *CTSK*; *ARNT* plays a major role in cancer pathway. Variations in *ARNT* and *CTSK* have been associated with leukemia, acute myeloblastic and pycnodysostosis respectively. The deletion at 1q21 involves the deletion of *SETDB1* in families 2145 and 2279. *SETDB1* has been associated with breast cancer (Genetic Association Database - GAD) (http://geneticassociationdb.nih.gov/). *ARNT* and *CTSK* have also been implicated in different cancers (GAD). A 74kb 1q21.1 deletion was observed in 8.6% (7/81) of the breast cancer dataset (cases) with a p-value of 0.072. 1q21 therefore could be said to be a potential susceptibility locus to prostate cancer.

A family with five genotyped cases has her *SPSB1* locus (1p36.22 - chr1:9321241-9400868) amplified in 100% of the cases genotyped. Overexpression of *SPSB1* increased HGF-induced reporter gene expression and ERK phosphorylation in HEK293 cells. The ERK signaling pathway plays a role in several steps of tumor development (Kim et.al, 2010). A review carried out by Gonzalo Rodríguez-Berriguete et al. (2011) suggested that MAPK transduction pathways are involved in prostate cancer development. Mention must however be made that

the mother does not carry the variation though it was found in another mother from a different family but never in any of the other family members.

Amplification in *PDZD2* (Activated in Prostate cancer - 5p13.3) locus was observed in 66.7-100% and 50-100% of the total cases genotyped and total cases in the family respectively in three families. Chaib et al. (2001) suggested that accumulation of the *PDZD2* protein may be associated with the initiation or early promotion of prostate tumorigenesis. This amplification was observed also in the breast cancer dataset.

Deletion at *ERBB4* locus (2q34) was observed in 43-67% and 38-50% of the total cases genotyped and total cases in the family respectively in three families. *ERBB4* is a transmembrane receptor tyrosine kinase that regulates cell proliferation and differentiation and it has been associated with different cancer types including colorectal and lungs cancer (GAD).

Deletion at *CTDSPL* locus (3p22.2) was observed in 67% and 40-50% of the total cases genotyped and total cases in the family respectively in three families (2394, 2429 & 2442). Kashuba et al. (2004) found that the *CTDSPL* gene was homozygously deleted in about 15% of major epithelial cancers. Expression of the *CTDSPL* gene was reduced more than 20-fold in 11 of 12 carcinoma cell lines and in 3 of 8 tumor biopsies. Chang et al. (2008) found that intron 5 of the *CTDSPL* gene contains the microRNA MIRN26A1. This deletion at 3p22.2 also involves the MIRN26A1. MicroRNA-26a (miR-26a) is a tumor suppressor that is reduced in hepatocellular carcinoma (HCC). MicroRNAs (miRNA) are a diverse class of small, non–protein-coding RNAs that function as critical gene regulators. Several bioinformatics analyses indicate that each miRNA regulates hundreds of target genes, underscoring the potential influences of miRNAs on almost every biological pathway (Ambros 2004; Barte 2004). Recent evidence has shown that about half of the human miRNAs are located in cancer-associated genomic regions and can function as tumor suppressor genes or oncogenes depending on their targets (Calin et al. 2004; Calin and Croce 2006; Esquela-Kerscher and Slack 2006).

P63 was detected in a variety of human and mouse tissues, including proliferating basal cells of epithelial layers in the epidermis, cervix, urothelium, and prostate. Unlike p53, the p63 gene encodes multiple isotypes with remarkably divergent abilities to transactivate p53 reporter genes and induce apoptosis. Double copy deletion of *TP63* locus (3q28), involving

the mother, was observed in 75% of the total affected in a family. This variation was not found in any other sample or family.

Variations at *CSMD1* (8p23.2), *CYP2C19* (10q23.33), *KL* (13q13.1) and *EMP3* (19q13.32) loci were found in at least 50% of the total affected in two different families each. All of these genes have been associated with different types of cancer including breast cancer (*CYP2C19*, *KL*), prostate cancer (*EMP3*, *CYP2C19*) and lung cancer (*CYP2C19*) to mention but a few.

Furthermore, variations at *DOCK4* (7q31.1), *DPP6* (7q36.2), *TNKS* (8p23.1), *MSR1* (8p22), *ADRA1A* (8p21.2), *ALOX5* (10q11.21), *POLE* (12q24.33), *RNASE1* (14q11.2), and *TMPRSS2* (21q22.3) loci were found also in at least 50% of the total affected in one family each. All of these genes have also been implicated in different cancer types including: prostate cancer (*RNASE1*, *MSR1*, *TMPRSS2*, *ADRA1A* and *ALOX5* – prostatic neoplasm), breast cancer (*TNKS*, *POLE*), and colorectal cancer (*ALOX5*); amongst other.

It is worth mentioning that none of the above-mentioned loci (genes) is statistically significant based on the case-control association test, the future challenge will be to expand sample sizes and to follow co-segregation of given CNVs with cancer phenotype within families to identify which of the genes involved in the CNVs might contribute to familial breast and prostate cancer predisposition.

Conclusion

This genome-wide association study was carried out with three objectives:

1. To study CNVs' contribution to genetic diseases using familial Prostate and Breast cancers as case studies
2. To assess and compare the performance of different CNV calling algorithms with the aim of knowing the degree of agreement between the different algorithms used in the study
3. To study heritability of CNVs in a bid to unravelling the genetic predisposition to cancer types in question

The result of the study suggests that CNVs are important predisposing factors to the cancer types in question, though this still needs to be confirmed in a larger population. Although most studies on heritability of CNVs were done on trios comprising the father, mother and a child from the family; which probably suggests that closely related individuals have higher heritability than distant individuals. Several of CNVs identified in this study were found enriched in certain families. This result supports previous findings that CNVs are heritable. The degree of agreement between the three sample-based algorithms used in this study is only about 50%. This further justifies the need for experimental validation of CNVs identified by the algorithms.

In conclusion, the result of the current genome-wide scan reveals that there are several loci and genes that play important role in predisposing an individual to the cancer types in question. Some of these however, could be peculiar to certain populations.

Next Step

Having carried out a brute-force-genome-wide scan for susceptibility loci to prostate and breast cancers and with the various interesting results obtained in the study, it is expedient to follow up on the results by expanding the sample size and with the use of a more efficient approach. This could be in form of targeted sequencing or exome sequencing of certain genes. While loci such as 1p21, 2p25, 3p26 and 10q11 are potential loci for further investigation, *ARNT*, *CTSK*, *SETDB1*, *SPSB1*, *TP63*, *EPHA3*, *KL*, *CYP2C19* and *EMP3* are good candidate for targeted sequencing or exome sequencing.

7 References

Aceves, C, B Anguiano, and G Delgado. "Is iodine a gatekeeper of the integrity of the mammary gland?" *Journal of mammary gland biology and neoplasia* , 2005: 10 (2): 189–196.

Aitman, TJ, et al. "Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans." *Nature*, 2006: 439, 851–855.

Alonso, A, et al. "CNstream: A method for the identification and genotyping of copy number polymorphisms using Illumina microarrays." *BMC Bioinformatics*, 2010: 11:264.

Ambros, V. "The functions of animal microRNAs." *Nature*, 2004: 431:350–5.

AS, Bassett, Marshall CR, Lionel AC, Chow EW, and Scherer SW. "Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome." *Hum Mol Genet*, 2008: 17:4045—4053.

Bae, J S, et al. "Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurismal hemorrhage in Japanese population ." *Biochem Biophys Res Commun* , 2008: 373:593-596.

Bae, J S, et al. "Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population ." *Biochem. Biophys. Res. Commun.*, 2008: 373:593-596.

Bamshad, M, S Wooding, BA Salisbury, and JC Stephens. "Deconstructing the relationship between genetics and race." *Nat Rev Genet* , 2004: 5: 598–609.

Barte, l DP. " MicroRNAs: genomics biogenesis, mechanism and function." *Cell* , 2004: 116:281–97.

Bassett, A S, C R Marshall, A C Lionel, and E W Chow. "Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome." *Hum Mol Genet*, 2008: 17:4045—4053.

Berger, M F, M S Lawrence, F Demichelis F, y Drier, K Cibulskis, and A Y Sivachenko. "The genomic complexity of primary human prostate cancer." *Nature*, 2011: 470: 214–220.

Berry, R, et al. "Evidence for a prostate cancer-susceptibility locus on chromosome 20 ." *Am. J. Hum. Genet.* , 2000: 67: 82-91.

Berthon, P, et al. "Predisposing gene for early-onset prostate cancer, localized on chromosome 1q42.2-43 ." *Am. J. Hum. Genet.* , 1998: 62: 1416-1424.

Berx, G, et al. "E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers. ." *EMBO J.* , 1995: 14: 6107-6115.

Bobadilla, JL, M Jr Macek, JP Fine, and PM Farrell. "Cystic fibrosis: a worldwide analysis of CFTR mutations–correlation with incidence data and application to screening. ." *Hum Mutat* , 2002: 19: 575–606.

Borresen, A. L, et al. "Screening for germ line TP53 mutations in breast cancer patients. ." *Cancer Res.* , 1992: 52: 3234-32.

Bostwick DG, D G, and J Qian. "High-grade prostatic intraepithelial neoplasia." *Mod Pathol*, 2004: 17:360-379.

Botstein, D, and N Risch. " Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. ." *Nat Genet* , 2003: 33: Suppl228–237.

Broeks, A, et al. "ATM-heterozygous germline mutations contribute to breast cancer-susceptibility." *Am. J. Hum. Genet.* , 2000: 66: 494-500.

Bruder, C.E.G, et al. "Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles ." *American Journal of Human Genetics*, 2008: 82 (3) , pp. 763-771.

Cairns, P, et al. "requent inactivation of PTEN/MMAC1 in primary prostate cancer." *Cancer Res.*, 1997: 57: 4997-5000.

Calin, GA, and CM Croce. "MicroRNA signatures in human cancers. ." *Nat Rev Cancer*, 2006: 6:857–66.

Calin, GA, et al. "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers." *Proc Natl Acad Sci U S A* , 2004: 101:2999–3004.

Cantor, SB, et al. "BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function." *Cell* , 2001: 105: 149-160.

Cappuzzo, F, et al. "Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non-small-cell lung cancer." *J Natl Cancer Inst.* , 2005: 97:643-655.

Carpten, J. D, et al. "A transforming mutation in the pleckstrin homology domain of AKT1 in cancer." *Nature* , 2007: 448: 439-444.

Cerami, E.G, G.D Bader, B.E Gross, and C. cPath Sander. "open source software for collecting, storing, and querying biological pathways." *BMC Bioinformatics*, 2006: 7:497.

Chano, T, K Kontani, K Teramoto, H Okabe, and S Ikegawa. " Truncating mutations of RB1CC1 in human breast cancers." *Nature Genet.* , 2002: 31: 285-288.

Chokkalingam, A P, et al. "Prostate Carcinoma risk subsequent to diagnosis of benign prostatic hyperplasia: A population-based cohort study in Sweden ." *Cancer* , 2003: 98:1727-1734.

Colella, S, et al. "QuantiSNP: an objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data." *Nucleic Acids Res.* , 2007: 35, 2013–2025.

Collaborative Group on Hormonal Factors in Breast Cancer. "Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease." *Lancet* , 2002: 360 (9328): 187–95.

Consortium, The Gene Ontology. ""The Gene Ontology project in 2008"." *Nucleic Acids Res.* , 2008: 36 (Database issue): D440–4.

Coquelle, A, E Pipiras, F Toledo, G Buttin, and M Debatisse. "Expression of fragile sites triggers intrachromosomal mammalian gene amplification and sets boundaries to early amplicons." *Cell*, 1997: 89:215–225.

Coquelle, A, L Rozier, B Dutrillaux, and M Debatisse. "Induction of multiple double-strand breaks within an hsr by meganucleaseI-SceI expression or fragile site activation leads to formation of double minutes and other chromosomal rearrangements." *Oncogene*, 2002: 21:7671–7679.

Crawford, E D. "Epidemiology of prostate cancer." *Urology*, 2003: 62:3-12.

Daar, AS, and PA Singer. " Pharmacogenetics and geographical ancestry: implications for drug development and global health. ." *Nat Rev Genet* , 2005: 6: 241–246.

DeMarzo, A M, W G Nelson, W B Isaacs, and J I Epstein. "Pathological and molecular aspects of prostate cancer." *Lancet* , 2003: 361: 955–964.

Dong, J.-T, et al. "KAI1, a metastasis suppressor gene for prostate cancer on human chromosome 11p11.2 ." *Science* , 1995: 268: 884-886.

Dong, X, et al. "Mutations in CHEK2 associated with prostate cancer risk ." *Am. J. Hum. Genet.* , 2003: 72: 270-280.

Eagle, LR, X Yin, AR Brothman, BJ Williams, NB Atkin, and EV Prochownik. "Mutation of the MXI1 gene in prostate cancer ." *Nature Genet.* , 1995: 9: 249-255.

Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al. "Completing the map of human genetic variation." *Nature*, 2007: 447:161–165.

Eichler, E E, D A Nickerson, D Altshuler, and A M Bowcock. "Completing the map of human genetic variation." *Nature*, 2007: 447:161–165.

Erkko, H, et al. "A recurrent mutation in PALB2 in Finnish cancer families. ." *Nature* , 2007: 446: 316-319.

Esquela-Kerscher, A, and FJ Slack. "Oncomirs—microRNAs with a role in cancer." *Nat Rev Cancer* , 2006: 6:259–69.

Farrer, LA, LA Cupples, JL Haines, B Hyman, and WA Kukull. " Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium." *Jama*, 1997: 278: 1349–1356.

Feng, Z, K Mehrdad, MC Anne, F T Charles, Batish Sat Dev, and RL James. "The DNA replication FoSTeS/MMBIR mechanism can generate human genomic, genic, and exon shuffling rearrangements." *Nature Genet.* , 2009: 41, 849–853 .

Feuk, L, A R Carson, and S W Scherer. "Structural variation in the human genome." *Nat. Rev. Genet.*, 2006: 7:85-97.

Feuk, L, A R Carson, and S W Scherer. "Structural variation in the human genome." *Nat. Rev. Genet*, 2006: 7:85–97.

Fiegler, H., et al. "Accurate and reliable high-throughput detection of copy number variation in the human genome ." *Genome Res.*, 2006: 16:1566–1574.

Freeman, J L, et al. "Copy number variation: New insights in genome diversity." *Genome Res*, 2006: 16:949–961.

Freeman, J L, G H Perry, and Feuk L et al. "copy number variation: new insights in genome diversity." *Genome Research*, 2006: 16: 949–961.

Friedman, RA, et al. "GRM7 variants confer susceptibility to age-related hearing impairment." *Hum. Molec. Genet.* , 2009: 18: 785-796.

Friedrichsen, DM, et al. "Identification of a prostate cancer susceptibility locus on chromosome 7q11-21 in Jewish families ." *Proc. Nat. Acad. Sci.* , 2004: 101: 1939-1944.

Gaddipati, JP, et al. "Frequent detection of codon 877 mutation in the androgen receptor gene in advanced prostate cancers ." *Cancer Res.* , 1994: 54: 2861-2864.

Gallagher, E, et al. "Gain of imprinting of SLC22A18 sense and antisense transcripts in human breast cancer. ." *Genomics* , 2006: 88: 12-17.

Gasparini, P, R Rabionet, G Barbujani, S Melchionda, and M Petersen. " High carrier frequency of the 35delG deafness mutation in European populations. Genetic Analysis Consortium of GJB2 35delG. ." *Eur J Hum Genet* , 2000: 8: 19–23.

Giordano, SH, DS Cohen, AU Buzdar, G Perkins, and GN Hortobagyi. "Breast carcinoma in men: a population-based study ." *Cancer* , 2004: 101 (1): 51–7.

Gonzalez, E, et al. "The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility." *Science* , 2005: 307 (5714): 1434–1440.

Gronberg, H, A.-K Ahman, M Emanuelsson, A Bergh, J.-E Damber, and A Borg. "BRCA2 mutation in a family with hereditary prostate cancer ." *Genes Chromosomes Cancer* , 2001: 30: 299-301.

Guess, H A. "Benign prostatic hyperplasia and prostate cancer ." *Epidemiol Rev* , 2001: 23:152-158.

Hakansson, S, et al. "Moderate frequency of BRCA1 and BRCA2 germ-line mutations in Scandinavian familial breast cancer." *Am. J. Hum.*, 1997: Genet. 60: 1068-1078.

Hardy, G H. "Mendelian porportions in a mixed population." *Science*, 1908: 28: 49–50.

Hastings, PJ, JR Lupski, SM Rosenberg, and G Ira. "Mechanisms of change in gene copy number." *Nat Rev Genet*, 2009: 10: 551–564.

Haviv-Chesner, A, Y Kobayashi, A Gabriel, and M Kupiec. "Capture of linear fragments at a double-strand break in yeast." *Nucleic Acids Res.* , 2007: 35:5192–5202.

Healey, C. S, et al. " A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability." *Nature Genet.* , 2000: 26: 362-364.

Hugot, JP, M Chamaillard, H Zouali, S Lesage, and JP Cezard. " Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." *Nature* , 2001: 411: 599–603.

Hupe´, P, N Stransky, J.P Thiery, F Radvanyi, and E Barillot. "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions." *Bioinformatics*, 2004: 20, 3413–3422.

Iafrate, A J, et al. "Detection of large-scale variation in the human genome." *Nat. Genet.*, 2004: 36:949-951.

Ionita-Laza, I, AJ Rogers, C Lange, BA Raby, and C Lee. "Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis." *Genomics*, 2009: 93(1):22-26.

Ionita-Laza, I, et al. "On the analysis of copy-number variations in genome-wide association studies: a translation of family-based association test." *Gent Epidemiol* , 2008: 32:273-284.

Itsara, A, G M Cooper, C Baker, S Girirajan, and J Li. "Population analysis of large copy number variants and hotspots of human genetic disease." *Am J Hum Genet.*, 2009: 84:148–161.

Jemal, A, F Bray, MM Center, J Ferlay, E Ward, and D Forman. "Global cancer statistics." *CA: A cancer journal for clinicians*, 2011: 61 (2): 69-90.

Jonsson, B-A, et al. "-160C/A polymorphism in the E-cadherin gene promoter and risk of hereditary, familial and sporadic prostate cancer ." *Int. J. Cancer* , 2004: 109: 348-352.

Jupe, E. R, et al. "Single nucleotide polymorphism in prohibitin 3-prime untranslated region and breast-cancer susceptibility." *Lancet* , 2001: 357: 1588-1589.

Kim, J H, S M Dhanasekaran, R Mehra, S A Tomlins, W Gu, and J Yu. "Integrative analysis of genomic aberrations associated with prostate cancer progression." *Cancer Res* , 2007: 67: 8229–8239.

Knight, SJ, et al. "Subtle chromosomal rearrangements in children with unexplained mental retardation." *Lancet* , 1999: 354(9191):1676-81.

Korbel, J O, A E Urban, F Grubert, J Du, and T E Royce. "Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome." *Proc Natl Acad Sci U S A*, 2007: 104: 10110–10115.

Korbel, JO, et al. "Paired-end mapping reveals extensive structural variation in the human genome." *Science*, 2007: 318: 420–426.

Ku, C. S, N Naidoo, M Hartman, and Y Pawitan. " Genome-wide Association Studies of Cancers. eLS. ." *John Wiley*, 2010: eLS.

Kuo, MT, RC Vyas, LX Jiang, and WN Hittelman. "Chromosome breakage at a major fragile site associated with P-glycoprotein gene amplification in multidrug-resistant CHO cells." *Mol Cell Biol.* , 1994: 14:5202–5211.

Kuschel, B, et al. "Variants in DNA double-strand break repair genes and breast cancer susceptibility." *Hum. Molec. Genet.* , 2002: 11: 1399-1407.

Lapointe, J, C Li, J P Higgins, M van de Rijn, E Bair, and K Montgomery. "Gene expression profiling identifies clinically relevant subtypes of prostate cancer ." *Proc Natl Acad Sci USA* , 2004: 101: 811–816.

Le, Maréchal C, E Masson E, and J M Chen. "Hereditary pancreatitis caused by triplication of the trypsinogen locus." *Nature Genetics*, 2006: 38: 1372–1374.

Lee, J A, and J R Lupski. "Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders." *Neuron*, 2006: 52 (1): 103–121.

Lee, JA, CM Carvalho, and JR Lupski. "A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders." *Cell*, 2007: 131:1235–1247.

Lee, JA, CM Carvalho, and JR Lupski. "A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders"." *Cell*, 2007: 131 (7): 1235–47.

Lee, JW, et al. "PIK3CA gene is frequently mutated in breast carcinomas and hepatocellular carcinomas." *Oncogene* , 2005: 24: 1477-1480.

Lee, S, S Kasif, Z Weng, and C R Cantor. "Quantitative Analysis of Single Nucleotide Polymorphisms within Copy Number Variation." *PLoS ONE*, 2008: 3(12): e3906.

Li, J, et al. "Oncogenic properties of PPM1D located within a breast cancer amplification epicenter at 17q23." *Nature Genet.* , 2002: 31: 133-134.

Lieber, MR. "The mechanism of human nonhomologous DNA end joining. ." *J Biol Chem.*, 2008: 283:1–5.

Liskay, RM, A Letsou, and JL Stachelek. "Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells." *Genetics*, 1987: 115:161–167.

Lister, Sam. *Urine test could speed treatment of prostate cancer.* London: The Sunday Times, 2009, Retrieved 9 August 2010.

Liu, W, et al. "Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer." *Cancer Res.* , 2009: 69:2176–2179.

Lovett, ST. "Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences." *Mol Microbiol.* , 2004: 52:1243–1253.

Lovett, ST, RL Hurley, Jr Sutera VA, RH Aubuchon, and MA Lebedeva. "Crossing over between regions of limited homology in Escherichia coli. RecA-dependent and RecA-independent pathways." *Genetics*, 2002: 160:851–859.

MacPherson, G, et al. "Association of a common variant of the CASP8 gene with reduced risk of breast cancer." *J. Nat. Cancer Inst.* , 2004: 96: 1866-1869.

Matsuda, M, et al. "Mutations in the RAD54 recombination gene in primary cancers." *Oncogene*, 1999: 18: 3427-3430.

McCarroll, S.A, and D Altshuler. "Copy-number variation and association studies of human diseas." *Nature Genetic*, 2007: 39:S37-S42.

McCarroll, S.A., Altshuler, D. "Copy-number variation and association studies of human disease." *Nat. Genet.*, 2007: 39:S37-S42.

Mills, RE, K Walter, C Stewart, RE Handsaker, and K Chen. "Mapping copy number variation by population-scale genome sequencing." *Nature*, 2011: 470:59–65.

Narla, G, et al. "KLF6, a candidate tumor suppressor gene mutated in prostate cancer ." *Science* , 2001: 294: 2563-2566.

Nguyen, D Q, C Webber, and C P Ponting. "Bias of selection on human copy-number variants." *PLoS Genet*, 2006: 2: e20.

Olshen, AB, E.S. Venkatraman, R. Lucito, and M Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data." *Biostatistics*, 2004: 5, 557–572.

Olshen, AB, ES Venkatraman, R Lucito, and M Wigler. "Circular binary segmentation for the analysis of array-based DNA copy number data." *Biostatistics*, 2004: 5, 557–572.

Padiath, Q S, et al. "Lamin B1 duplications cause autosomal dominant leukodystrophy." *Nat Genet*, 2006: 38:1114–1123.

Parkin, DM, SL Whelan, J Ferlay, L Raymond, and J Young. "Cancer incidence in five continents." *Lyon, IARC*, 1997: Vol. VII (IARC Scientific publications No. 143).

Pico, A. R., T. Kelder, M. P. Van Iersel, K. Hanspers, B. R. Conklin, and C Evelo. "WikiPathways: Pathway Editing for the People." *PLoS Biology*, 2008: 6 (7): e184.

Piotrowski, A, et al. "Somatic mosaicism for copy number variation in differentiated human tissues." *Hum. Mutat.*, 2008: 29: 1118–1124. doi: 10.1002/humu.20815.

Pujana, MA, et al. "Network modeling links breast cancer susceptibility and centrosome dysfunction ." *Nature Genet.* , 2007: 39: 1338-1349.

Rao, DS, et al. "Huntingtin-interacting protein 1 is overexpressed in prostate and colon cancer and is critical for cellular survival ." *J. Clin. Invest.* , 2002: 110: 351-360.

Redon, R, S Ishikawa, K R Fitch, L Feuk, and G H Perry. "Global variation in copy number in the human genome." *Nature*, 2006: 444: 444–454.

Ridker, PM, JP Miletich, CH Hennekens, and JE Buring. "Ethnic distribution of factor V Leiden in 4047 men and women. Implications for venous thromboembolism screening." *Jama* , 1997: 277: 1305–1307.

Robbins, C M, W A Tembe, A Baker, S Sinari, T Y Moses, and S Beckstrom-Sternberg. "Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors." *Genome Res* , 2011: 21: 47–55.

Rokman, A, et al. "ELAC2/HPC2 involvement in hereditary and sporadic prostate cancer ." *Cancer Res.* , 2001: 61: 6038-6041.

Rokman, A, et al. "Hereditary prostate cancer in Finland: fine-mapping validates 3p26 as a major predisposition locus ." *Hum. Genet.* , 2005: 116: 43-50.

Rotger, M, et al. "Partial deletion of CYP2B6 owing to unequal crossover with CYP2B7." *Pharmacogenet.* , 2007: Genomics 17, 885–890.

Sachidanandam, R, D Weissman, S C Schmidt, J M Kakol, and L D Stein. "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." *Nature*, 2001: 409: 928–933.

Santoro, E, M DeSoto, and J Hong Lee. "Hormone Therapy and Menopause." *National Research Center for Women & Families*, 2009.

Sariego, J. "Breast cancer in the young patient." *The American surgeon* , 2010: 76 (12): 1397–1401.

Sebat, J, et al. "Large-scale copy number polymorphism in the human genome." *Science*, 2004: 305:525-528.

Sharp, AJ, et al. " Segmental duplications and copy-number variation in the human genome." *Am J Hum Genet* , 2005: 77:78-88.

Shen, M M, and C Abate-Shen. "Molecular genetics of prostate cancer: new prospects for old challenges. ." *Genes Dev* , 2010: 24: 1967–2000.

Siegel, R, E Ward, O Brawley, and A Jemal. "Siegel, R, et al (2011). "Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths."." *CA Cancer J Clin*, 2011: 61: 212–36.

Smith, J R, et al. "Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search ." *Science* , 1996: 274: 1371-1374.

Sommer, F, T Klotz, and B J Schmitz-Drager. "Lifestyle issues and genitourinary tumours." *World J Urol* , 2004: 21:402-413.

Steiner, P, D. M Barnes, W. H Harris, and R. A Weinberg. " Absence of rearrangements in the tumour susceptibility gene TSG101 in human breast cancer." *(Letter) Nature Genet.* , 1997: 16: 332-333.

Stephens, JC, DE Reich, DB Goldstein, HD Shin, and MW Smith. "Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. ." *Am J Hum Genet* , 1998: 62: 1507–1515.

Sun, X, et al. "Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer ." *Nature Genet.* , 2005: 37: 407-412.

Taylor, B S, N Schultz, H Hieronymus, A Gopalan, Y Xiao, and B S Carver. "Integrative genomic profiling of human prostate cancer." *Cancer Cell* , 2010: 18: 11–22.

Teo, YY, et al. "Singapore Genome Variation Project: A haplotype map of three South-East Asian populations." *Genome Res.*, 2009b: 19, 2154–2162.

Thai, TH, et al. "Mutations in the BRCA1-associated RING domain (BARD1) gene in primary breast, ovarian and uterine cancers ." *Hum. Molec. Genet.*, 1998: 7: 195-202.

Thompson, D, et al. "Evaluation of linkage of breast cancer to the putative BRCA3 locus on chromosome 13q21 in 128 multiple case families from the Breast Cancer Linkage Consortium." *Proc. Nat. Acad. Sci.*, 2002: 99: 827-831.

Tsukasaki, K, et al. "Mutations in the mitotic check point gene, MAD1L1, in human cancers ." *Oncogene* , 2001: 20: 3301-3305.

Vehmanen, P, et al. "Low proportion of BRCA1 and BRCA2 mutations in Finnish breast cancer families: evidence for additional susceptibility genes." *Hum. Molec. Genet.*, 1997: 6: 2309-2315.

Venturi, S. "Is there a role for iodine in breast diseases? ." *The Breast* , 2001: 10 (5): 379–382.

Walsh, T, et al. "Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer." *JAMA* , 2006: 295: 1379-1388.

Wang, K, et al. "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data." *Genome Res.*, 2007: 17, 1665–1674.

Wang, W, M. A Tucker, M. M Doody, R. E Tarone, and J. P Struewing. " A single nucleotide polymorphism in the 5-prime-UTR of RAD51 is associated with the risk of breast cancer among BRCA1/2 mutation carriers." *(Abstract) Am. J. Hum. Genet.*, 1999: 65: A22 only.

Witte, J S, et al. "Genomewide scan for prostate cancer-aggressiveness loci ." *Am. J. Hum. Genet.*, 2000: 67: 92-99.

Xu, J, et al. "A combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the International Consortium for Prostate Cancer Genetics ." *Am. J. Hum. Genet.* , 2005: 77: 219-229.

Xu, J, et al. "Linkage and association studies of prostate cancer susceptibility: evidence for linkage at 8p22-23 ." *Am. J. Hum. Genet.*, 2001: 69: 341-350.

Yager, JD, and NE Davidson. " Estrogen carcinogenesis in breast cancer ." *New Engl J Med* , 2006: 354 (3): 270–82.

Yanez, L, J Groffen, and DM Valenzuela. "c-K-ras mutations in human carcinomas occur preferentially in codon 12." *Oncogene* , 1987: 1: 315-318.

Yu, K.-D, et al. "Functional polymorphisms, altered gene expression and genetic association link NRH:quinone oxidoreductase 2 to breast cancer with wild-type p53." *Hum. Molec. Genet.* , 2009: 18: 2502-2517.

Zogopoulos, G, K C Ha, F Naqib, S Moore, and H Kim. "Germ-line DNA copy number variation frequencies in a large North American population." *Hum Genet*, 2007.

8 Appendices

Table S1 Genes Associated with Mendelian Disorders in Prostate Cancer Dataset

GENE	Locus	Gene MIM ID	Disorder / OMIM Phenotype	Phenotype MIM ID
ANG	14q11.2	105850	Amyotrophic lateral sclerosis 9	611895
ATP2A1	16p11.2	108730	Brody myopathy	601003
RHCE	1p36.11	111700	Rh-null disease, amorph type	
RHCE	1p36.11	111700	Blood group, Rhesus	
CACNA1C	12p13.33	114205	Brugada syndrome 3	611875
CACNA1C	12p13.33	114205	Timothy syndrome	601005
CYP2C19	10q23.33	124020	Clopidogrel, impaired responsiveness to	609535
CYP2C19	10q23.33	124020	Mephenytoin poor metabolizer	609535
CYP2C19	10q23.33	124020	Opremazole poor metabolizer,	609535
CYP2C19	10q23.33	124020	Proguanil poor metabolizer	609535
ARNT	1q21.3	126110	Leukemia, acute myeloblastic	
DPP6	7q36.2	126141	Ventricular fibrillation, paroxysmal familial	612956
PTPRC	1q31.3-q32.1	151460	Severe combined immunodeficiency, T cell-negative, B-cell/natural killer-cell positive	608971
PTPRC	1q31.3-q32.1	151460	Hepatitic C virus, susceptibility to	609532
ALOX5	10q11.21	152390	Asthma, diminished response to antileukotriene treatment in	600807
ALOX5	10q11.21	152390	Atherosclerosis, susceptibility to	
MSR1	8p22	153622	Barrett esophagus/esophageal adenocarcinoma	614266
MSR1	8p22	153622	Prostate cancer, hereditary	176807
PNP	14q11.2	164050	Immunodeficiency due to purine nucleoside phosphorylase deficiency	613179
PRKCA	17q24.2	176960	Pituitary tumor, invasive	
LHX3	9q34.3	600577	Pituitary hormone deficiency, combined, 3	221750
LPP	3q27-q28	600700	Leukemia, acute myeloid	601626
LPP	3q27-q28	600700	Lipoma	
CTSK	1q21.3	601105	Pycnodysostosis	265800
MYO1A	12q13.3	601478	Deafness, autosomal dominant 48	607841
NDN	15q11.2	602117	Prader-Willi syndrome	176270
TUFM	16p11.2	602389	Combined oxidative phosphorylation deficiency 4	610678
RAD51C	17q22	602774	Fanconi anemia, complementation group 0	613390
RAD51C	17q22	602774	Breast-ovarian cancer, familial, susceptibility to, 3	613399
TP63	3q28	603273	ADULT syndrome	103285
TP63	3q28	603273	Ectrodactyly, ectodermal dysplasia, and cleft lip/palate syndrome 3	604292
TP63	3q28	603273	Hay-Wells syndrome	106260
TP63	3q28	603273	Limb-mammary syndrome	603543

Table S1 Genes Associated with Mendelian Disorders in Prostate Cancer Dataset (continued)

TP63	3q28	603273	Orofacial cleft 8	129400
TP63	3q28	603273	Rapp-Hodgkin syndrome	129400
TP63	3q28	603273	Split-hand/foot malformation 4	605289
DYNC2H1	11q22.3	603297	Asphyxiating thoracic dystrophy 3	613091
DYNC2H1	11q22.3	603297	Short rib-polydactyly syndrome, type II, digenic	263520
DYNC2H1	11q22.3	603297	Short rib-polydactyly syndrome, type III	263510
KL	13q13.1	604824	Tumoral calcinosis, hyperphosphatemic	211900
KL	13q13.1	604824	Coronary artery disease, susceptibility to	
TUBA8	22q11.21	605742	Polymicrogyria with optic nerve hypoplasia	613180
PRODH	22q11.21	606810	Hyperprolinemia, type I	239500
PRODH	22q11.21	606810	Schizophrenia, susceptibility to, 4	600850
NIPA1	15q11.2	608145	Spastic paraplegia-6	600363
ATCAY	19p13.3	608179	Ataxia, cerebellar, Cayman type	601238
ASL	7q11.21	608310	Argininosuccinic aciduria	207900
PEX26	22q11.21	608666	Adrenoleukodystrophy, neonatal	202370
PEX26	22q11.21	608666	Refsum disease, infantile	266510
PEX26	22q11.21	608666	Zellweger syndrome	214100
PITPNM3	17p13.2	608921	Cone-rod dystrophy 5	600977
ALDH3A2	17p11.2	609523	Sjogren-Larsson syndrome	270200
ATXN10	22q13.31	611150	Spinocerebellar ataxia 10	603516
GUSB	7q11.21	611499	Mucopolysaccharidosis VII	253220
GRXCR1	4p13	613283	Deafness, autosomal recessive 25	613285

Table S2 Genes Associated with Mendelian Disorders in Breast Cancer Dataset

GENE	Locus	Gene MIM ID	Disorder / OMIM Phenotype	Phenotype MIM ID
ATP2A1	16p11.2	108730	Brody myopathy	601003
RHD	1p36.11	111680	Rh-negative blood type	
BRCA1	17q21.31	113705	Breast-ovarian cancer, familial, 1	604370
BRCA1	17q21.31	113705	Pancreatic cancer, susceptibility to, 4	614320
CACNA1C	12p13.33	114205	Brugada syndrome 3	611875
CACNA1C	12p13.33	114205	Timothy syndrome	601005
CHRNA7	15q13.3	118511	Schizophrenia, neurophysiologic defect in	
SLC1A1	9p24.2	133550	?Dicarboxylicaminoaciduria	222730
MSR1	8p22	153622	Barrett esophagus/esophageal adenocarcinoma	614266
MSR1	8p22	153622	Prostate cancer, hereditary	176807
PROS1	3q11.1	176880	Thrombophilia due to protein S deficiency	612336
EPCAM	2p21	185535	Colorectal cancer, hereditary nonpolyposis, type I	613244
EPCAM	2p21	185535	Diarrhea 5, with tufting enteropathy, congenital	613217
HNF1B	17q12	189907	Diabetes mellitus, noninsulin-dependent	125853
HNF1B	17q12	189907	Renal cysts and diabetes syndrome	137920
HNF1B	17q12	189907	Renal cell carcinoma	144700
ACACA	17q12	200350	Acetyl-CoA carboxylase deficiency	613933
TUSC3	8p22	601385	Mental retardation, autosomal recessive 7	611093
FGF10	5p12	602115	Aplasia of lacrimal and salivary glands	180920
FGF10	5p12	602115	LADD syndrome	149730
TUFM	16p11.2	602389	Combined oxidative phosphorylation deficiency 4	610678
PARK2	6q26	602544	Adenocarcinoma of lung, somatic	211980
PARK2	6q26	602544	Adenocarcinoma, ovarian, somatic	167000
PARK2	6q26	602544	Parkinson disease, juvenile, type 2	600116
PARK2	6q26	602544	Leprosy, susceptibility to	607572
DNAI2	17q25.1	605483	Ciliary dyskinesia, primary, 9, with or without situs inversus	612444
EHMT1	9q34.3	607001	Kleefstra syndrome	610253
KANK1	9p24.3	607704	Cerebral palsy, spastic quadriplegic, 2	612900
COX4I2	20q11.21	607976	Exocrine pancreatic insufficiency, dyserythropoietic anemia, and calvarial hyperostosis	612714
GLIS3	9p24.2	610192	Diabetes mellitus, neonatal, with congenital hypothyroidism	610199
PFKM	12q13.11	610681	Glycogen storage disease VII	232800
DOCK8	9p24.3	611432	Hyper-IgE recurrent infection syndrome, autosomal recessive	243700
DOCK8	9p24.3	611432	Mental retardation, autosomal dominant 2	614113
MBD5	2q23.1	611472	Mental retardation, autosomal dominant 1	156200
LIPA	10q23.31	613497	Cholesteryl ester storage disease	278000
LIPA	10q23.31	613497	Wolman disease	278000

Table S3 Enrichment analysis result for gene list with odds ratio >1 in the PrCa dataset.

GO ANALYSIS_ODDSRATIO>1			
GO	TERM	GENES	P-VALUE
Biological process			
GO:0007156	homophilic cell adhesion	*CDH4,PCDHA4,PCDHA11,PCDHA1,PCDHA5,PCDHA9,PCDHA8,PCDHA6,PCDHA7,PCDHA10,PCDHA12,PCDH9,PCDHA3,PCDHA2*	6.00E-09
GO:0016337	cell-cell adhesion	*CDH4,CTNNA3,PTPRC,PCDHA4,PCDHA1,PCDHA8,PCDHA6,PCDHA7,PCDHA11,PCDHA9,PCDHA5,PCDHA10,PCDH9,PCDHA12,PCDHA3,PCDHA2*	9.20E-07
GO:0007155	cell adhesion	*DGCR6,CDH4,CTNNA3,LPP,NRXN3,SEMA5A,PTPRC,PCDHA4,PCDHA1,PCDHA8,PCDHA6,PCDHA7,PCDHA11,PCDHA9,PCDHA5,PCDHA10,PCDHA12,PCDH9,PCDHA3,PCDHA2*	1.30E-03
GO:0022610	biological adhesion	*CDH4,SEMA5A,PCDHA4,PCDHA1,CTNNA3,PCDHA8,PCDHA6,PCDHA7,DGCR6,PTPRC,NRXN3,PCDHA11,PCDHA9,PCDHA5,PCDHA10,PCDHA12,PCDH9,PCDHA3,LPP,PCDHA2*	1.30E-03
GO:0007399	nervous system development	*TP63,KLHL1,PRKCA,ALDH1A2,MDGA2,NAIP,NDN,CDH4,PCDHA4,PCDHA1,PCDHA8,PCDHA6,PCDHA7,SEMA5A,NRXN3,PCDHA11,PCDHA5,ATXN10,PCDHA10,GABRA5,PCDHA3,ALDH3A2,PCDHA2*	1.40E-03
Molecular function			
GO:0004522	pancreatic ribonuclease activity	*ANG,RNASE1,RNASE11,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	1.56E-11
GO:0016894	endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 3'-phosphomonoesters	*RAD51C,ANG,RNASE1,RNASE11,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	1.56E-11
GO:0016892	endoribonuclease activity, producing 3'-phosphomonoesters	*RNASE1,ANG,RNASE11,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	4.53E-11
GO:0004521	endoribonuclease activity	*APEX1,RNASE1,ANG,RNASE11,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	4.54E-09
GO:0004540	ribonuclease activity	*ANG,RNASE1,APEX1,RNASE11,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	9.89E-08
GO:0004519	endonuclease activity	*RAD51C,RNASE1,APEX1,ANG,RNASE11,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	2.13E-07
GO:0004518	nuclease activity	*RAD51C,RNASE1,APEX1,ANG,RNASE11,REXO1,RNASE12,RNASE6,RNASE9,RNASE4,RNASE10*	1.81E-06

Table S3 Enrichment analysis result for gene list with odds ratio >1 in the PrCa dataset (continued)

GO:0005509	calcium ion binding	*ATP2A1,CDH4,PCDHA4,PITPNM3,PCDHA1,PCDHA8, PCDHA6,PCDHA7,PRKCA,ALOX5,PCDHA11,NRXN3, CACNA1C,PCDHA5,PCDHA9,PLCG2,PCDHA10,PCD HA12,PCDH9,PCDHA3,PCDHA2*	5.00E-04
GO:0016787	hydrolase activity	*TMPRSS2,IMMP2L,AGBL4,TMEM55B,AADACL4,NAIP ,RNASE10,PAPPA2,ANG,RNASE11,ARL1,PLCG2,DYN C2H1,C3AR1,RNASE1,DPP6,APEX1,GUSB,USP18,RN ASE6,TUFM,OSGEP,RNASE9,ATP2A1,CTSK,KL,PTPR C,MYO1A,RAD51C,ATAD3B,DUSP16,REXO1,RNASE1 2,TUBA8,RNASE4,NDST4,AADACL3*	7.00E-04
GO:0016788	hydrolase activity, acting on ester bonds	*C3AR1,PLCG2,RNASE1,APEX1,USP18,RNASE6,AADA CL4,RNASE9,RNASE10,PTPRC,RAD51C,RNASE11,AN G,DUSP16,REXO1,RNASE12,RNASE4*	1.20E-03
Cellular component			
GO:0031226	intrinsic to plasma membrane	*TMPRSS2,MSR1,NRG3,CDH4,PCDHA4,PCDHA1,RHC E,PCDHA8,PCDHA6,HCN2,PCDHA7,KL,PTPRC,PCD HA11,NRXN3,PCDHA5,GABRA5,PCDHA10,PCDHA3, PRKD1,PCDHA2,C3AR1*	3.70E-02
GO:0005887	integral to plasma membrane	*TMPRSS2,MSR1,NRG3,CDH4,PCDHA4,PCDHA1,RHC E,PCDHA8,PCDHA6,HCN2,PCDHA7,KL,PTPRC,PCD HA11,NRXN3,PCDHA5,GABRA5,PCDHA10,PCDHA3, PRKD1,PCDHA2,C3AR1*	3.70E-02
KEGG PATHWAY ANALYSIS ODDS RATIO > 1			
KEGG ID	KEGG pathway	GENES	P-VALUE
830	Retinol metabolism	*CYP2C19,DHRS3,ALDH1A2*	4.20E-02
330	Arginine and proline metabolism	*PRODH,ALDH3A2,ASL*	4.20E-02
4666	Fc gamma R-mediated phagocytosis	*PTPRC,PRKCA,PLCG2*	4.50E-02
1100	Metabolic pathways	*ALOX5,PRODH,GUSB,ALDH1A2,CYP2C19,PLCG2,SG MS1,ASL,NDST4,DHRS3,ALDH3A2*	4.50E-02
4012	ErbB signaling pathway	*NRG3,PRKCA,PLCG2*	4.50E-02
4020	Calcium signaling pathway	*CACNA1C,PRKCA,ATP2A1,PLCG2*	4.50E-02
PATHWAY COMMONS ANALYSIS ODDS RATIO > 1			
PATHWAY ID	NAME	GENES	P-VALUE
DB_ID:769	Proline catabolism	*PRODH,ASL*	1.60E-02
DB_ID:206	Platelet activation triggers	*PRKCA,PLCG2*	3.70E-02

Table S4 Enrichment analysis result for gene list with odds ratio > 2 in the PrCa dataset

GO ANALYSIS ODDS RATIO > 2			
GO	TERM	GENES	P-VALUE
Biological process			
GO:0008361	regulation of cell size	*EMP3,CDH4,NRG3,TP63,NDN,SGMS1,PAPPA2*	9.00E-04
GO:0016049	cell growth	*TP63,NRG3,EMP3,CDH4,NDN,SGMS1,PAPPA2*	9.00E-04
GO:0032535	regulation of cellular component size	*SGMS1,PAPPA2,TP63,NRG3,EMP3,CDH4,NDN*	2.20E-03
GO:0090066	regulation of anatomical structure size	*CDH4,SEMA5A,PCDHA4,PCDHA1,CTNNA3,PCDHA8, PCDHA6,PCDHA7,DGCR6,PTPRC,NRXN3,PCDHA11, PCDHA9,PCDHA5,PCDHA10,PCDHA12,PCDH9,PCD HA3,LPP,PCDHA2*	5.20E-03
GO:0065008	regulation of biological quality	*ARNT,PTPRC,CDH4,NRG3,TP63,NDN,SGMS1,EMP3, HCN2,GRXCR1,CTSK,PAPPA2,C3AR1*	5.40E-03
GO:0040007	growth	*SGMS1,PAPPA2,TP63,NRG3,EMP3,CDH4,NDN*	1.35E-02
KEGG PATHWAY ANALYSIS ODDS RATIO > 2			
KEGG ID	KEGG pathway	GENES	P-VALUE
4514	Cell adhesion molecules (CAMs)	*PTPRC,CDH4*	6.90E-02
4010	MAPK signaling pathway	*DUSP16,FGF22*	1.10E-01
5200	Pathways in cancer	*ARNT,FGF22*	1.10E-01
PATHWAY COMMONS ANALYSIS ODDS RATIO > 2			
PATHWAY ID	NAME	GENES	P-VALUE
DB_ID:1045	Glypican pathway	*PTPRC,DUSP16*	1.70E-01
DB_ID:916	Transcription	*EIF3C,POLRMT*	1.70E-01
DB_ID:1031	Glypican 1 network	*PTPRC,DUSP16*	1.70E-01

Table S5 Enrichment analysis result of gene list with odds ratio > 1 in the BrCa dataset

	GO ANALYSIS OF GENE LIST WITH OR > 1		
GO	TERM	GENES	P-VALUE
	Biological process		
GO:0042742	defense response to bacterium	*DEFB115, DEFB123, DEFB121, DEFB116, DEFB124, DEFB119, DEFB118*	1.95E-02
GO:0009617	response to bacterium	*DEFB115, DEFB123, DEFB121, DEFB116, DEFB124, DEFB119, DEFB118, EPHA3*	4.52E-02
	Molecular function		
GO:0004872	receptor activity	*CHRNA7, EPHA3, LOC619207, GPR142, ROBO1, MSR1, OR4M1, CD300E, CD300C, OR4Q3, ERBB4, GRM8, SEMA5A, OR4K5, GRM5, OR4N4, CD300LD, OR4K2, OR4N2, OR4K1, CD300A, TAS2R1, CD300LB, GPCRLTM7, SEMA4B, OR4M2*	3.21E-02
GO:0004035	alkaline phosphatase activity	*ALPPL2, ALPP*	3.21E-02
GO:0004888	transmembrane receptor activity	*CHRNA7, EPHA3, LOC619207, GPR142, ROBO1, MSR1, OR4M1, CD300C, OR4Q3, ERBB4, GRM8, SEMA5A, OR4K5, GRM5, OR4N4, OR4K2, OR4N2, OR4K1, TAS2R1, GPCRLTM7, OR4M2*	3.21E-02
GO:0008046	axon guidance receptor activity	*SEMA5A, ROBO1*	3.21E-02
GO:0060089	molecular transducer activity	*CHRNA7, EPHA3, LAT, LOC619207, GPR142, ROBO1, MSR1, OR4M1, CD300E, CD300C, OR4Q3, ERBB4, GRM8, SEMA5A, OR4K5, GRM5, OR4N4, CD300LD, CHN2, OR4K2, CLNK, SH2B1, R4N2, OR4K1, CD300A, TAS2R1, CD300LB, GPCRLTM7, SEMA4B, OR4M2*	3.21E-02
GO:0004871	signal transducer activity	*CHRNA7, EPHA3, LAT, LOC619207, GPR142, ROBO1, MSR1, OR4M1, CD300E, CD300C, OR4Q3, ERBB4, GRM8, SEMA5A, OR4K5, GRM5, OR4N4, CD300LD, CHN2, OR4K2, CLNK, SH2B1, R4N2, OR4K1, CD300A, TAS2R1, CD300LB, GPCRLTM7, SEMA4B, OR4M2*	3.21E-02
	KEGG PATHWAY ANALYSIS		
KEGG ID	KEGG pathway	GENES	P-VALUE
4020	Calcium signaling pathway	*CHRNA7, CAMK2D, CACNA1C, ATP2A1, ERBB4, GRM5*	1.30E-03

Table S5 Enrichment analysis result of gene list with odds ratio > 1 in the BrCa dataset (continued)

4740	Olfactory transduction	*CAMK2D, OR4K2, OR4N2, OR4K1, OR4M1, OR4Q3, OR4M2, OR4K5, OR4N4*	1.30E-03
790	Folate biosynthesis	*ALPPL2, ALPP*	6.30E-03
310	Lysine degradation	*WHSC1, BBOX1, EHMT1*	6.30E-03
4360	Axon guidance	*EPHA3, SEMA4B, SEMA5A, ROBO1*	1.20E-02
4720	Long-term potentiation	*CAMK2D, CACNA1C, GRM5*	1.42E-02
4520	Adherens junction	*MLLT4, CTNNA3, IQGAP1*	1.61E-02
970	Aminoacyl-tRNA biosynthesis	*FARS2, TARSL2*	4.19E-02
4722	Neurotrophin signaling pathway	*CAMK2D, YWHAE, SH2B1*	4.75E-02
	WIKIPATHWAYS ANALYSIS		
GENE SET ID	GENE SET NAME	GENES	P-VALUE
WP501	GPCRs, Class C Metabotropic glutamate, pheromone	*GRM8, GRM5*	1.26E-02
WP716	Retinol metabolism (BiGCaT, NuGO)	*DHRS3, CYP2E1*	3.85E-02
	PATHWAY COMMONS ANALYSIS		
PATHWAY ID	NAME	GENES	P-VALUE
DB_ID:433	Homologous recombination repair of replication-independent double-strand breaks	*BCL2L1, BRCA1*	2.75E-02
DB_ID:1035	Lissencephaly gene (LIS1) in neuronal migration and development	*YWHAE, IQGAP1*	2.75E-02
DB_ID:435	ATM mediated response to DNA double-strand break	*BCL2L1, BRCA1*	2.75E-02
DB_ID:432	Homologous Recombination Repair	*BCL2L1, BRCA1*	2.75E-02
DB_ID:767	Double-Strand Break Repair	*BCL2L1, BRCA1*	3.12E-02
DB_ID:707	Post-translational protein modification	*PROS1, PIGW*	4.07E-02

Table S6 Enrichment analysis result of gene list with odds ratio > 2 in the BrCa dataset

	GO ANALYSIS		
GO	TERM	GENES	P_VALUE
	Biological process		
GO:0007565	female pregnancy	*PSG11, PSG4, PSG2, PSG5, PSG9, PSG1, PSG7, PSG6*	1.86E-09
GO:0051704	multi-organism process	*EPHA3, PSG11, PSG4, PSG5, PSG1, PSG7, PSG2, PSG9, PSG6*	2.00E-04
GO:0022414	reproductive process	*PSG11, PSG4, PSG2, PSG5, PSG9, PSG1, PSG7, PSG6*	1.60E-03
GO:0000003	reproduction	*PSG11, PSG4, PSG2, PSG5, PSG9, PSG1, PSG7, PSG6*	1.60E-03
	Molecular function		
GO:0016814	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amidines	*APOBEC3B, APOBEC3A*	4.18E-02