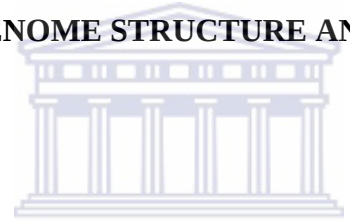
UNIVERSITY *of the*
WESTERN CAPE

A COMPUTATIONAL CHARACTERISATION OF THE RELATIONSHIP BETWEEN GENOME STRUCTURE AND DISEASE GENES



A thesis submitted in fulfilment of the requirements for the degree of *Magister Scientiae in Bioinformatics* at the South African National Bioinformatics Institute (SANBI), Faculty of Natural Sciences, The University of the Western Cape (UWC).

by

Tracey Deborah Kibler

May 2012

Supervisor: Doctor Nicki Tiffin

Co-Supervisor: Professor Alan Christoffels

Keywords

Disease genes

Genome structure

Base composition

Genetic variation

Gene length

Position effect

Translational research

Homologous recombination events

Recombination hotspots

Frequency of recombination

Abbreviations

DNA	Deoxyribonucleic acid
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
cM	CentiMorgan
SNP	Single Nucleotide Polymorphism
bp	base pair
kb	kilo base pair (10^3 base pairs)
Mb	mega base pair (10^6 base pairs)
OMIM	Online Mendelian Inheritance in Man
IQR	Interquartile range
GWAS	Genome-wide Association Studies
MRCA	Most Recent Common Ancestor

Abstract

This is a pilot study to investigate the relationship between disease gene status and the structure of the human genome with specific reference to regions of recombination. It compares certain characteristics of a control set of genes, with no reported association or function in any known disease, with a second set of well-curated genes with a known association to a disease.

One of the benefits of recombination is the introduction of new combinations of genetic variation in the genome. Recombination hotspots are regions on the chromosome where higher than normal frequencies of breaking and rejoining between homologous chromosomes occur during meiosis. The hotspot regions exhibit both a non-random distribution across the human genome and varying frequencies of breaking and rejoining.

The study analyzed a set of features that represent general properties of human genes; namely base composition (percentage GC content), genetic variation (single nucleotide polymorphisms - SNPs), gene length, and positional effect (distance from chromosome end), in both the disease-associated gene set and the control set. These features were linked to recombination hotspots in the human genome and the frequency of recombination at these hotspots. Descriptive statistics was used to determine differences between the occurrences of these features in disease-associated genes compared to the control set, as well as differences in the occurrence of these same features in subset of genes containing

an internal recombination hotspot compared to the genes with no internal recombination hotspot.

The study found that disease-associated genes are generally longer than those in the control set, which is consistent with previous studies. It also found that disease-associated genes are much more likely to contain a recombination hotspot than those genes with no disease association. The study did not, however, find any association between disease gene status and the other set of features; namely GC content, SNP numbers or the position of a gene on the chromosome. Further analysis of the data suggested that the increased probability of disease-associated genes containing a recombination hotspot is most likely an effect of longer gene length and that the presence of a recombination hotspot is not sufficient in its own right to cause disease gene status.

May 2012

Declaration

I declare that *A computational characterisation of the relationship between genome structure and disease genes* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Tracey D. Kibler May 2012

Signed: _____

Acknowledgements

I would like to express my gratitude to all those who made it possible to complete this thesis. I want to thank the South African National Bioinformatics Institute (SANBI) for allowing me the opportunity to complete my masters' degree with them.

I am deeply indebted to my supervisor Dr Nicki Tiffin whose help, suggestions and encouragement helped me in all areas of my research and in writing, not only the thesis, but also with the Python scripts. Thank you, also, to my co-supervisor Prof. Alan Christoffels.

I would like to thank Galen Wright for his careful review of this thesis and Sumir Panji and Darlington Mapiye for their advice and assistance with the technical sections. To all my other colleagues and friends at SANBI, thank you for your support and encouragement.

Mostly, I would like to give my special thank you to my husband Rolf van Zyl whose patience, support and love enabled me to complete this work. He presented me with the opportunity to do this and without his continued support and encouragement I would have been lost.

Index

Table of Contents

Table of Contents (cont..)

Table of Contents (cont..)

CHAPTER THREE

Table of Contents (cont..)

List of Tables

Table of Figures

Table of Figures (cont..)

Table of Figures (cont..)

CHAPTER ONE

Background and Introduction

In recent years there have been hundreds, if not thousands, of candidate disease genes predicted by genome-wide studies that have been performed on large groups of patients, as well as animal experimental models of disease, and cell culture models of disease. It is for this reason that the objective of present-day computational approaches to disease-associated gene identification is to try to isolate the 'most likely' disease gene candidates for further empirical analysis by translational researchers so that this massive amount of information can be put into practical applications (Tiffin *et al*., 2006; Tiffin *et al*., 2009). Computational disease gene predictions attempt to efficiently identify genes of diagnostic, prognostic and therapeutic value for further experimental validation. The data used for computational analysis include gene structure and sequence data, functional annotation of candidate genes, the characteristics of known disease genes, gene regulatory networks and protein-protein interactions, and data from animal models and disease phenotype. There is however, a negative aspect of these methods. They are typically developed using training sets and training data that are mainly Eurocentric (Tiffin *et al*., 2006; Tiffin *et al*., 2009). This Caucasian-centric bias is regularly shown in genome-wide association studies (GWAS). The participation ratio for a GWAS is usually ~10:1 individuals with European ancestry compared to all other ethnic groups combined (http://www.genome.gov/gwastudies). Also, the majority of the GWAS that do involve non-European ethnic groups generally include

smaller sample sizes than those with participants of European ancestry. There are a number of reasons for this bias; (1) better funding for Caucasian-centric studies, the majority of researchers with sufficient funding for GWAS come from Europe and USA where the population base is predominantly of European ancestry, and (2) there is an increased population complexity in African populations (Need and Goldstein, 2009), the African continent has the second largest population (after Asia) and this implies greater diversity between subgroups in Africans resulting in an increased likelihood of population stratification. This creates a need for larger sample sizes and more complex analysis.

In this thesis, I investigate the possibility of a fresh approach to enable disease-associated gene prediction, using the position of genes within the genome structure. This approach is novel because it is the first of its kind to investigate whether the distance of a gene from a recombination hotspot and the frequency of the recombination at that hotspot may be related to the likelihood that a specific gene is the underlying cause of a specific disease. This could therefore, be seen as a pilot study to investigate other possible approaches to predicting "most likely" disease-associated gene candidates, by analyzing the relationship between disease-associated genes and genome structure. Since epidemiological evidence has clearly and consistently shown that disease occurrence and genetics underlying disease can vary substantially between populations and ethnic groups (Via *et al*., 2009) if such a relationship between gene and genome structure exists, it will provide a novel way of prioritising disease genes in a population/ethnic-specific way based on the unique haploblock structure of populations.

The remainder of this chapter will provide the background to the study, as well as reviews current knowledge and the state of the art in this field.

Chapter 2 outlines the materials and methodology used in the study. This includes, a description of the process used for the assembly of training gene sets, the scoring system used to describe the relationship between the gene and its flanking recombination hotspots, and the analysis used to compare the score distribution for disease-associated genes and genes with no reported association or function in any known disease, hereafter referred to as the 'control set'.

Chapter 3 describes the results of the study and particularly the differences and similarities in the scores of the genes in the control set. It also describes the results of analysis done on the collective list of the same genes (disease-associated and control set) that have been separated into genes that contain an internal hotspots compared to genes that have no internal hotspots.

In chapter 4, the likelihood of whether a gene underlying a disease might be due to the effect of a genes proximity to the recombination hotspots and frequency of recombination at these hotspots is discussed. This chapter also outlines the strengths and limitations of the study as well as presenting directions for future research.

1.1. The Process and Mechanisms of Homologous Recombination

Genetic recombination is a crucial element of evolution and forms the basis of our genetic history. Genetic recombination plays two essential biological roles; it guarantees the consistent transfer of genetic information from one generation to the next and it generates new combinations of genetic variants (Zhang *et al*., 2009). There are two types of genetic recombination; homologous DNA recombination where genetic material is exchanged between different regions of two sister chromatids and, the far rarer, non-homologous DNA recombination where genetic material is exchanged between different chromosomes. Due to the fact that it is less likely to occur naturally, it is the frequency of non-homologous recombination that has key implications for genetic reliability, genetic progression and more importantly, human disease.

Non-homologous recombination, involves the exchange of genetic material between different chromosomes. Non-homologous recombination repairs double-strand breaks in DNA without the need for a homologous template. The repair is guided by short homologous DNA sequences called micro-homologies that are present in single-stranded overhangs on the ends of double-strand breaks. Non-homologous recombination has been implicated in many diseases (reviewed in Chen *et al*., 2010) and can also lead to insertions and deletions as well as to translocations and telomere fusion (reviewed in Chen *et al.*, 2010 and Baird. 2008).

Homologous DNA recombination, involves the exchange of genetic material between two highly similar or even identical molecules of DNA. In humans, this is an essential process occurring during meiosis where genetic materials are exchanged

between two newly duplicated chromosomes. These chromosomes can be divided into two types; autosomes and sex chromosomes. Human somatic cells contain 23 pairs of chromosomes (22 autosomes pairs and one pair of sex chromosome (XX in female and XY in male)).

As this thesis focuses on hotspot recombination within genes on the chromosome it is important to note that 95% of the human Y chromosome is unable to recombine with the X chromosome, except for small pieces of pseudo-autosomal regions at the telomeres (remaining 5%) and therefore, for our purpose, the X chromosome and Y chromosome will be considered as one.

1.1.1. Meiosis and Crossing over

In sexually reproducing organisms, such as humans, meiosis takes place in specialized diploid (46) cells called zygotes and results in haploid (23) daughter cells with only one set of homologous chromosomes (Figure 1). During meiosis, a homologous chromosome pair, consisting of a chromosome from the mother and a chromosome from the father, aligns along the centre of the nucleus and exchanges sections or fragments of chromosome in a process of the recombination referred to as "crossing over". Crossing over can cause alleles, previously on the same chromosome, to be separated onto two different chromosomes. The further apart the alleles are on the original chromosome, the greater the chance that a cross-over event will occur and the greater the chance that the alleles will be separated after crossover. Crossing over produces a new and unique chromosome containing genetic information different from the parents, and hence, is referred to as DNA recombination.

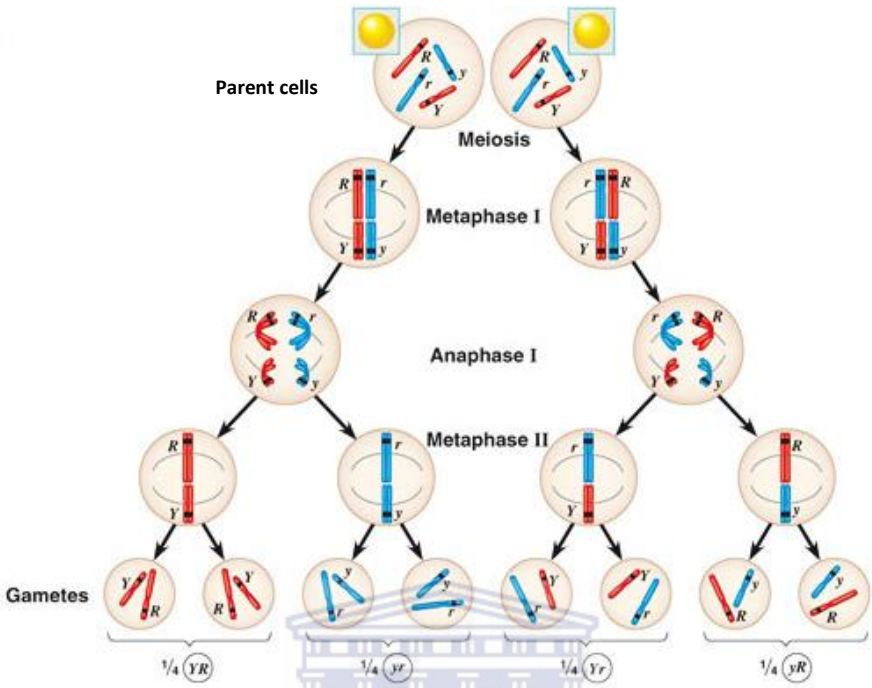


Figure 1 The Stages of Meiosis – An Overview

Meiosis begins with one diploid cell (46) containing two copies of each chromosome, one from the mother and one from the father and results in four haploid cells (23) containing one copy of each chromosome. There are two major phases of meiosis, meiosis I and meiosis II. During meiosis I, the parent cell has two sets of chromosomes (diploid) and each single cell divides into two. During meiosis II, those two cells each divide into four daughter cells containing a single set of chromosomes each (haploid).

(Adapted from URL: http://t1.gstatic.com/images)

1.1.2. The Holliday Junction

The most widely accepted model for DNA recombination was first proposed by Robin Holliday in 1964 (Holliday, 1964) and is based on the formation of a 'Holliday junction' as shown in Figure 2.

A 'Holliday junction' is formed when two double-stranded DNA molecules (one from each parent) separate into four strands in order to exchange segments of genetic information. This occurs during meiosis when a double strand break (DSB) occurs on one of the chromosomes. This break allows for sections of the DNA surrounding it to be degraded by an enzyme thereby creating a single strand DNA that has the potential to bind to homologous DNA on another chromosome. The Holliday junction can resolve in two possible ways; gene conversion or crossing over. In gene conversion, small amounts of DNA sequence information are transferred between chromosomes having no effect on the subsequent gametes. In crossing over, much larger sections of DNA sequence information are exchanged and this has a much greater effect on the subsequent gametes, as all genetic material beyond the recombination points is exchanged.

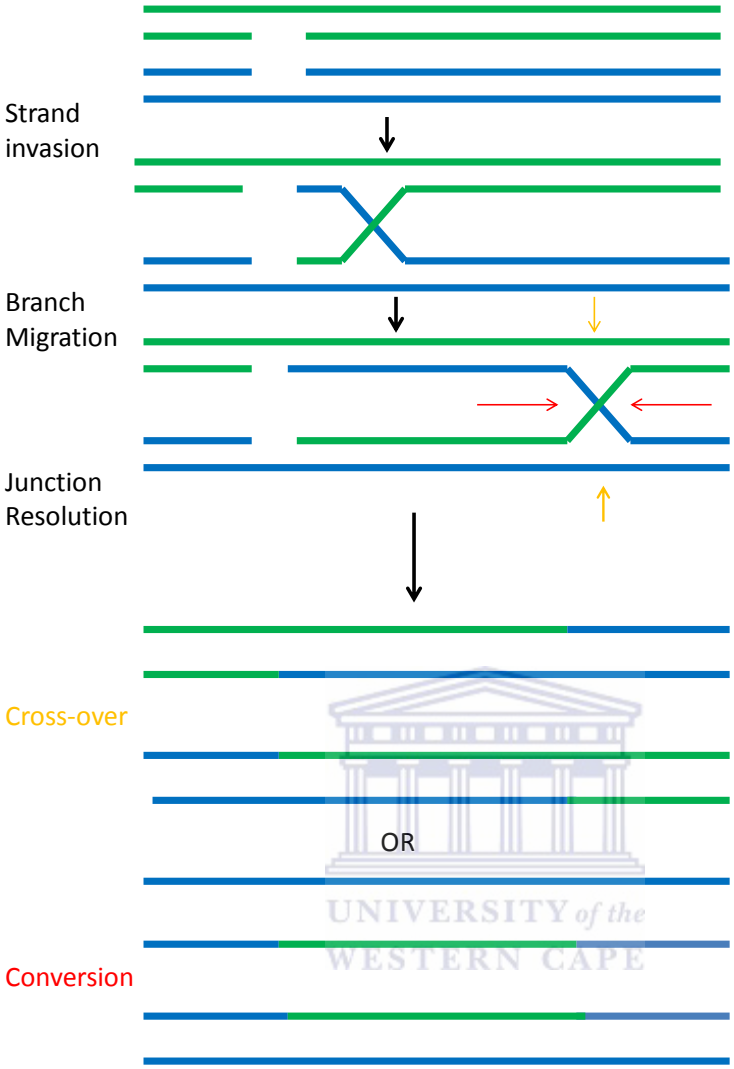


Figure 2 Holliday model of recombination

The Holliday model for homologous recombination shows single-strand breaks occurring at the same point on one strand of each parental DNA. The free ends of each broken strand then migrate across to the other DNA helix, where both strands are joined together. The resulting crossover junction is called a Holliday junction. Depending on how this structure is resolved either cross-over or gene conversion products result.

1.1.3. The CentiMorgan

The unit for measuring the rate at which recombination occurs on a DNA molecule is called the centiMorgan (cM). It is usually calculated in terms of the expected number of recombinations that occur between two loci per generation. One centiMorgan is equal to a 1% chance that in a single generation a marker at one genetic locus on a chromosome will be separated from a marker at a second locus due to crossing over (Figure 3). In other words, if two loci are 1cM apart then during meiosis they will recombine on average 1 out of 100 generations. Since the unit is incremental, if two loci are separated by 40cM then one would expect to observe forty recombination events on average per generation between the two loci. Reference, however, is seldom made to distances greater than 50cM as this would imply a recombination rate of greater than 50% and would be equivalent to random assortment. This would then suggest that the loci are either on separate chromosomes or that the distance between the two was too great to be of any significance.

The genetic distance between two genes can be estimated by calculating the number of offspring that exhibit two linked genetic traits and then estimating the percentage of offspring where the traits are not linked. The higher the percentages of offspring that do not show both traits, the farther apart on the chromosome the two genes are. Genes with a percentage value lower than 50% are referred to as linked.

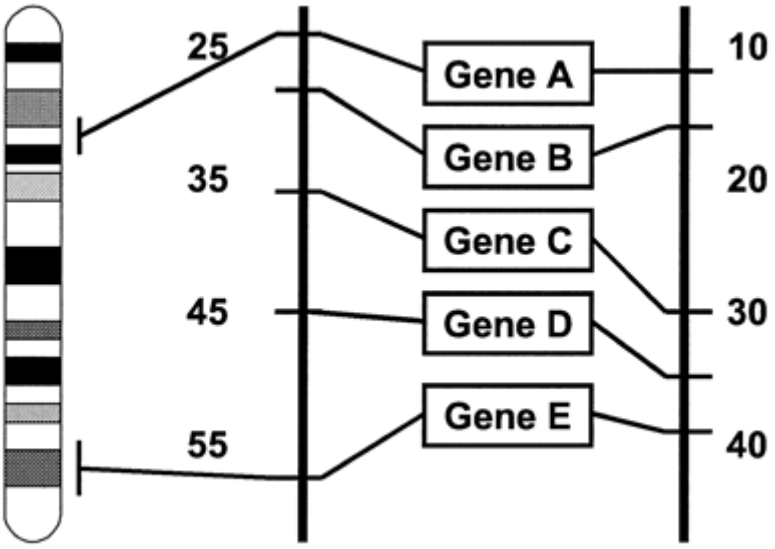


Figure 3 Illustration of the CentiMorgan

A centiMorgan is the unit of genetic distance that represents a percentage probability of recombination during meiosis. If two genes are 5 cM apart, there is a 5% chance they will break apart during meiosis.

1.1.4. Gene Mapping

There are important applications for gene mapping in research; (1) it is useful in locating the position of genes on a chromosome. For example, if two genes are linked and the position of the first gene is known the position of the second gene can be assumed as it must be within close proximity, (2) it is valuable for estimating genetic risk. For example, if a gene cannot be investigated directly, then one could use a variation found at a closely linked locus to identify the presence or absence of an unfavorable allele.

Gene mapping can be done by either building a physical map or a genetic map. A physical map illustrates the actual physical location of genes on a chromosome while a genetic map represents the distribution of a set of genes on a chromosome with the distance between loci expressed as percent recombination, or centiMorgan. In humans, the rate of recombination on most chromosomes is lower in males than in females and therefore female genetic maps are longer than male genetic maps. On average, the genetic maps of females are 90% longer than the same maps in males however, the base pair number per chromosome remains the same and therefore their physical maps are identical.

The standard method for studying the rate of recombination in the human genome begins with building a genetic map (Figure 4). Genetic maps can be constructed using family-based linkage analysis. Linkage analysis is done by using the genotypes of a notable number of individuals within a family to identify sufficient numbers of genetic markers that can be used to determine chromosomes in current generations that are recombinants of those in earlier generations. This data can then be used to

calculate the number of recombination events between markers and to show the position of its known genes, or genetic markers, relative to each other in terms of recombination frequency (in centiMorgans). Due to the difference in genetic maps between males and female, the frequency of recombination rate between males and females also differs. The distance in cM can be given as either female or male based or alternatively a sex-averaged map can be constructed using the "sex-average" distance in cM between the two loci.

Comparisons between genetic maps and physical maps have shown that some regions of chromosomes are more likely to be involved in crossing over than others. These regions are referred to as recombination hotspots and are approximately one or two thousand base pairs (bp) in length. These hotspot regions are, in turn, flanked by "cold spot" regions. These are sections on the chromosome where the frequency of recombination is lower than average (Lichten *et al*., 1995). Recombination hotspot regions are investigated in this study due to their potential effect on DNA sequence variation in human chromosomes and because these regions could possibly be used to identify alleles that cause disease.

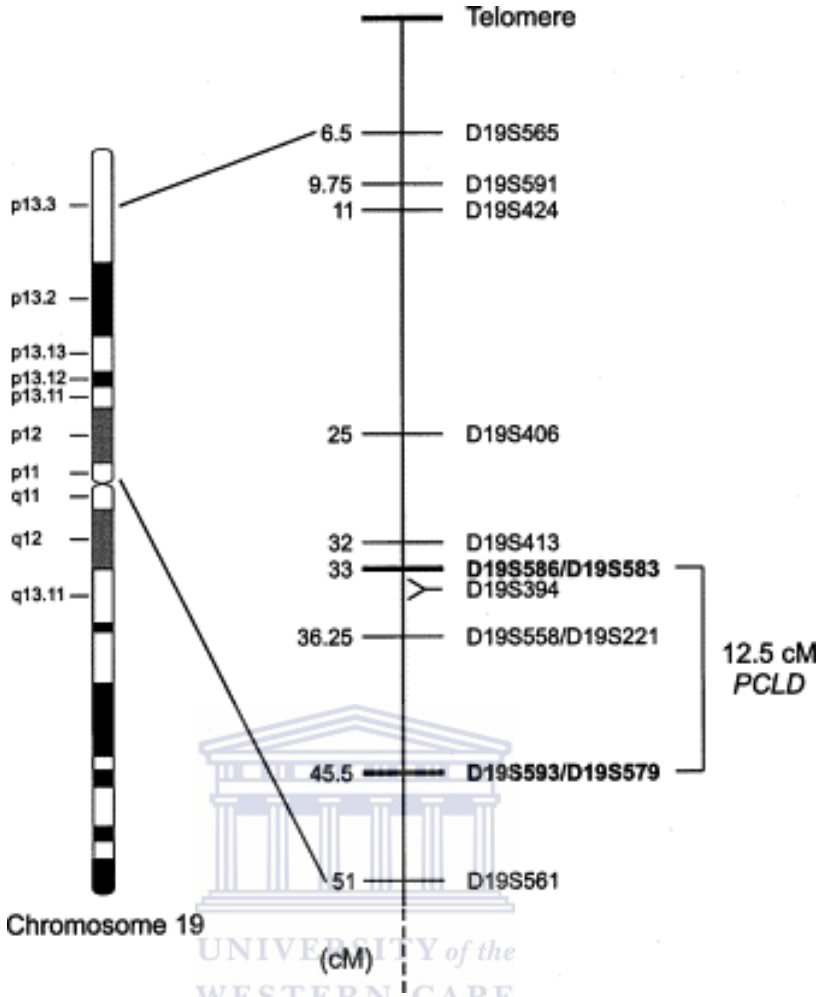


Figure 4 Comparison of a Genetic map and Physical map

A genetic map, as illustrated on the right, shows the genes on a chromosome with the distances between loci given in centiMorgans, i.e. as a percent of recombination. A physical map, as illustrated on the left, describes the physical location of genes on chromosomes.

(Received from URL: http://t1.gstatic.com/images?q=tbn:ANd9GcSurlxYn3ZxkuyWeSPnCaRH7vNuSSXiyOGP9f24NidNJAluAW5hsg)

1.1.5. Haploblock Structure

In humans, 99.5% of our DNA sequence is identical. It is however, the variations in our DNA that affect an individual's susceptibility to disease risk. A variation at a single base pair, called a SNP, can have a significant influence. A haplotype is a combination of alleles at multiple linked loci that an individual will inherit as a unit from the parent and are common to related individuals. A haploblock is a set of SNPs on the same chromosomes that are inherited as a block and are common in unrelated individuals. These 'haploblocks' indicate regions on the chromosome that have not been altered by recombination.

1.1.6. Linkage Disequilibrium

As discussed earlier in the chapter, the frequency of recombination is a measure of the genetic distance between two sites on a chromosome. At a recombination frequency of 1%, two loci will separate once in every 100 recombinations during meiosis. Loci separated by >50cM are said to be unlinked as recombination has reached a maximum of 50%. In this scenario there is an equal chance that the loci could stay together or separate during meiosis. As the distance between two loci decreases and eventually reaches 0cM for loci that are very close to one another, the recombination frequency will reach 0% and the level of linkage will increase. In other words, at 0cM distance, there is only one possible conclusion, the loci are linked. This indicates that equilibrium is impossible, and hence crossing over will never occur and the two loci are said to be in linkage disequilibrium (Figure 5). In this scenario the loci, as well as the entire region between them, will be inherited as a single unit (Micklos and Freyer. DNA Science: A First Course, 2nd Edition).

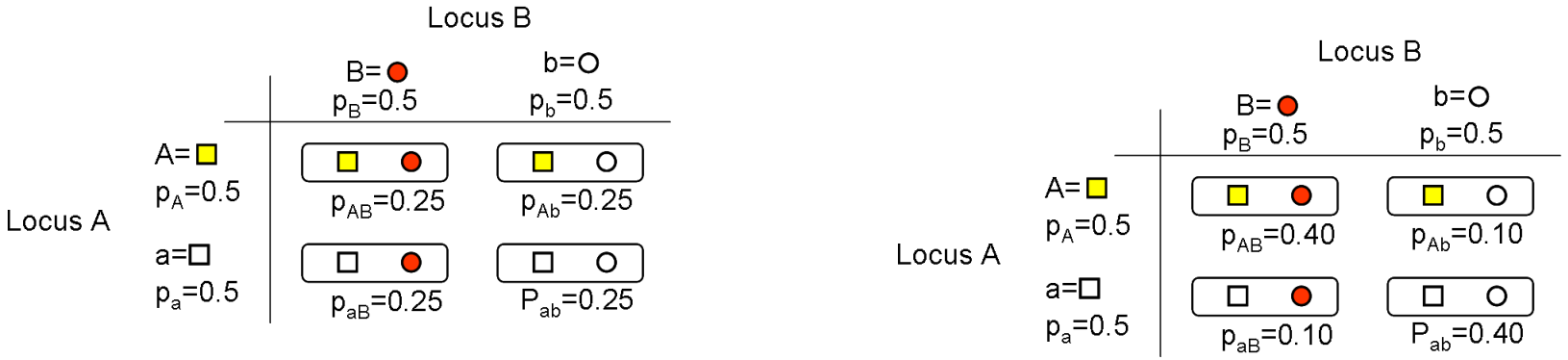


Figure 5 Linkage Disequilibrium

Two loci are in linkage disequilibrium (LD) when pairs of alleles from the two loci are inherited together in the same gametes more often than expected at random from their corresponding alleles frequencies.

On the left, two loci, locus A and locus B, each one with two alleles (A,a and B,b) and allele frequencies as $p_A=p_a=0.5$ and $p_B=p_b=0.5$. Locus A and B are in linkage equilibrium (LE) when the frequencies of the four possible gametes AB, Ab, aB, and ab are:

$p_{AB} = p_A p_B = 0.25$ $p_{Ab} = p_A p_b = 0.25$ $p_{aB} = p_a p_B = 0.25$ $p_{ab} = p_a p_b = 0.25$

And the sum of the gamete frequencies equals 1.

On the right, the actual gamete frequencies are different from what it is expected under LE and shows that locus A and B are in linkage disequilibrium: $p_{AB} = p_A p_B = 0.40$ $p_{Ab} = p_A p_b = 0.10$ $p_{aB} = p_a p_B = 0.10$ $p_{ab} = p_a p_b = 0.40$

(URL http://www.biometris.wur.nl/UK/Tutorial+on+LD+mapping/Concepts+and+definitions/What+is+LD/)

1.2. Characterising Recombination Hotspots

As previously mentioned, recombination hotspots are regions within the genome where increased rates of recombination occur (Lichten & Goldman 1995). Research has shown that most hotspots share morphological traits such as structure and appearance. These homologous traits infer a more recent common ancestor, and can be used to reconstruct evolutionary histories. On estimate, a hotspot is in the region of 1.5 to 2.0 kilobases in length.

1.2.1. The International HapMap Project

Recombination rates, and therefore identification of recombination hotspots, have been calculated at a genome-wide level (Myers *et al*., 2005), using the data presented by the HapMap consortium (The International HapMap Consortium 2005, URL: www.hapmap.org). Although understanding of the nature of chromosomal recombination hotspots across the genome is limited (Li and Stephens, 2003; Stumpf and Goldstein, 2003; McVean *et al*., 2004; Frazer *et al*., 2007) some analyses have didentify short motifs that are associated with recombination hotspots (Myers *et al*., 2008). Data regarding recombination rates and hotspots can be accessed and downloaded by population from the HapMap website, and resources are available at this, and other, public databases for viewing the data (Barnes 2006; Frazer *et al*., 2007).

The HapMap project was designed to create a public-access database identifying patterns of common sequence variations. The aim of the project is to create a central database of human genetic variation that can be used for genetic studies of human health and disease (Manolio *et al*., 2008). The data was collected from several

populations from different ancestral geographic locations to ensure that the HapMap included most of the common variation and some of the less common variation in different populations. First initiated in 2002, the aim was to identify genetic variations distributed amongst the different population groups, by identifying SNPs spaced at approximately 5-kilobase intervals with a minor allele frequency of at least 5%.

The HapMap project studied the linkage disequilibrium (LD) relationships across the human genome in four different ethnic groups (The International HapMap Project, 2003). These included a panel of 30 trios from the Yoruba, Nigeria (YRI); a panel of 30 CEPH trios from US Utah residents with European ancestry (CEU); and a panel of 45 unrelated Japanese individuals from Tokyo (JPT) and 45 unrelated Han Chinese individuals from Beijing (CHB). The two statistical measures of linkage disequilibrium, used by the project are; (1) D′ - if two SNPs have not been separated by recombination during the history of the sample, then D′ = 1, and (2) r^2 - when two alleles are always observed together then $r^2 = 1$.

As previously mentioned, there are 22 autosome pairs, and one pair of sex chromosomes, XX in females and XY in males. Since only two portions of the Y chromosome are homologous with the X chromosome, and hence do not recombine, these are not genotyped in the HapMap database and therefore do not feature in the analysis.

1.3. Current Protocols for Estimating Recombination Rates

In review, recombination frequency refers to the rate with which a single chromosomal crossover will take place between two genes during meiosis and is measured by counting the number of recombinant offspring and dividing it by the total number of offspring (Hudson, 1987). In humans, it has been possible to construct whole-genome genetic maps by scoring markers across the entire genome based on three methods for estimating recombination rates; (1) sperm typing, (2) pedigree data analysis and (3) by studying linkage disequilibrium (Zheng *et al*., 2010).

1.4. Defining a Gene and Gene Features

The genes included in both the disease-associated gene list as well as list of genes that have not been documented with a disease association (the control set) can be separated into three types; protein-coding genes, non protein-coding genes (pseudogenes), and RNA. Below is a brief description of each as well as the justification for including these genes in this study.

Protein-coding genes are heredity DNA units passed from parent to offspring that code for a protein. These are the genes responsible for the physical and inheritable characteristics or phenotype of the offspring. Protein-coding genes include exons, which code for the amino acid sequence of the protein; introns, which contain regulatory regions and splicing sequences elements. The UTR regions are found at both the 5' and 3' ends of the gene and contain regulatory elements. UTR regions are included in the analysis as mutations in these regions could affect the expression of the gene and hence could influence disease phenotype (Chen *et al*., 2006, Pickering and Willis. 2005).

MicroRNA (miRNA) is a short ribonucleic acid (RNA) molecule found in eukaryotic cells and is usually only a few nucleotides in length. miRNAs have been included in the study as several of these molecules have been found to have links with human disease (de Pontual *et al*, 2011, Thum *et al*., 2007) as well as in cancer (Farazi *et al*., 2011, Thomson *et al*., 2005). For example, microRNA-21 was one of the first microRNAs to be identified as an oncogenic-associated microRNA.

Pseudogenes are a subset of non protein-coding genes and are dysfunctional paralogues of their functional counterparts but have lost their protein-coding ability.

Pseudogenes were included in the study for several reasons. Firstly, there have been pseudogenes previously implicated in disease (McEente *et al*., 2010). These, therefore, fit our criteria for the disease gene set and should consequently also be included in the control set. Secondly, the definition of a pseudogene is inconsistent and many pseudogenes have been subsequently predicted to have biological functions (reviewed in Svensson *et al*., 2006). The removal of pseudogenes from the control set is unlikely to be accurate based on the poor understanding of how to determine if a gene is non-functional. The fact that the estimated range of pseudogenes in humans is between 3600 and 20000 (Svensson *et al*., 2006, Ohshima *et al*., 2003) gives an indication of how inconsistent not only the figures are but also the definition of what a pseudogene actually is. The expression products have been found for 4-6% of pseudogenes and this too is likely to be an underestimation (Harrison *et al*., 2005). Lastly, pseudogenes have very similar sequences to their functional counterparts with sequence conservation of an estimated 67% and inclusion of these genes is therefore unlikely to affect or skew analysis of gene characteristics in the control set (Svensson *et al*., 2006). There is also evidence of conservation and functionality of pseudogenes in mouse and drosophila (Balakirev and Ayala, 2003).

Figure 6 illustrates the gene including the coding and non-coding regions. The coding region of a gene's DNA or RNA codes for proteins and is composed of exons. The transcription start codon binds the region nearer the 5' end while the stop codon is nearer the 3' end. The 5’ untranslated region (UTR) begins at the transcription start site and ends just before the start codon (usually AUG) of the coding region. The 3’ UTR region starts with the nucleotide immediately following the stop codon of the

coding region. The 5'- and 3' UTR regions are contained within exons but do not code for amino acids.

For the analysis, I have reviewed certain gene features of known disease-associated genes and compared them to a control set (see Table 1). These include gene length, base composition (GC content), genetic variation (SNP count and SNP density), positional effect (distance of a gene from the chromosome end depending on its location with respect to the centromere), and recombination hotspots (presence/absence of a recombination hotspot within the gene or distance from nearest hotspot and frequency of recombination of internal/nearest hotspot).

For the purpose of this thesis a gene includes 5'UTR, exons, introns and 3'UTR. The GC content and genetic variations (SNP count) were both pre-calculated values obtain from the Ensembl database version 65. A chromosome has two arms, commonly referred to as the p arm and q arm, which are separated by the centromere. The position effect (distance to chromosome end) was calculated separately for each arm by comparing gene position to centromere position individually for each chromosome. For the p arm distance, distance from chromosome end was calculated by subtracting the chromosome start (0) from the position of the start of the gene. For the q arm distance, distance from chromosome end was calculated by subtracting the position of the end of the gene from the position of the chromosome end. It is important to note that the centromeres are based on Ensembl assemble GRCh37 and are averaged at a length of 3 million base pairs each. This is due to the fact that the centromeric region of a chromosome is hard to sequence accurately because it is heterochromatic. Frequently, sequencing centers simply report centromeres as gaps.

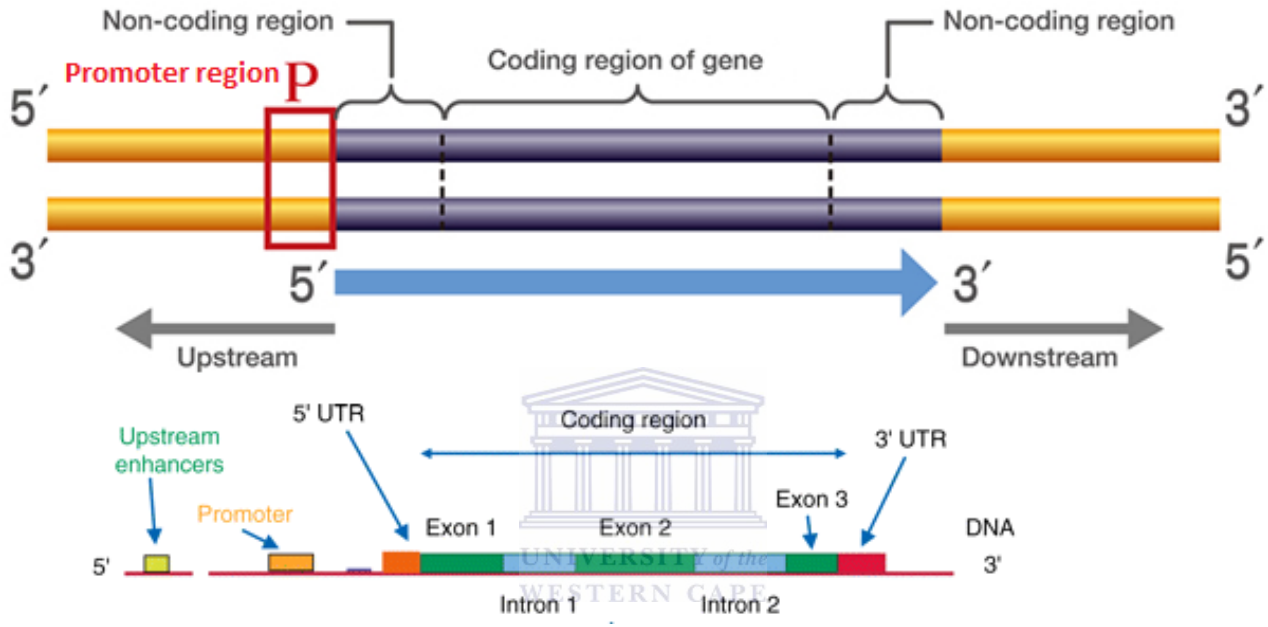


Figure 6 Illustration of the Human gene

The coding region of a gene's DNA or RNA codes for proteins and is composed of exons. The transcription start codon binds the region nearer the 5' end while the stop codon is nearer the 3' end. The 5’ untranslated region (UTR) begins at the transcription start site and ends just before the start codon (usually AUG) of the coding region. The 3’ UTR region starts with the nucleotide immediately following the stop codon of the coding region.. The 5’- and 3’ UTR regions are contained within exons but do not code for amino acids.

Table 1 **Gene Feature Directory**

The list of features reviewed in known disease-associated genes compared to the control set.

	Database	Description	Resource
GC content			
	Martview	GC content as a % of gene	http://www.ensembl.org/biomart/martview/
SNP density			
	SNPMart 65	Number of SNPs per 10kbp	ftp://ftp.ensembl.org/pub/release-65/mysql/snp_mart_65/
Gene Length			
	EnsemblMart 65	Length of gene in bp	ftp://ftp.ensembl.org/pub/release-65/mysql/ensembl_mart_65/
Position			
	EnsemblMart 65	Shortest distance of gene to chromosome end	ftp://ftp.ensembl.org/pub/release-65/mysql/ensembl_mart_65/
Recombination			
	HapMap	Rate of recombination per gene	http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/

1.4.1. Base Composition: percentage GC content

The GC-content of a gene refers to the percentage of guanine or cytosine nitrogenous bases present in a DNA molecule. As the GC bases pair by three hydrogen bonds as opposed to the AT (adenine/thymine) bases that pair only by two hydrogen bonds, DNA with a high GC-content is more stable than DNA with a low GC-content. The GC content of human DNA varies extensively across the entire human genome from between 30% to 60%. Nonetheless Spencer *et al*., 2006 and Fullerton *et al*., 2001 found that there was an increase in GC content that is highly localized to recombination hotspots. One explanation for this is that GC-rich regions promote the occurrence of hotspots, and this assumption is supported by the association between recombination hotspots and the occurrence of double-strand breaks in GC-rich regions (Mucha *et al*., 2000). Freudenberg *et al*., 2009 used partial correlation analysis to determine a casual relationship between GC content, exon density and recombination rate in the human genome. The Zhao *et* al., 2003 study found that the distribution of SNPs among the gene structure categories, both intergenic and genic (intron and exonic) was dependant on the GC content of chromosomes.

1.4.2. Genetic Variation: Single Nucleotide Polymorphisms (SNPs)

Genetic variation can be described on two tiers; firstly, at population level, where genetic variation is depicted as a percentage of allele frequency in a population; and secondly at individual level, where genetic diversity occurs either as homozygous or heterozygous at a specific locus. With the increase in the number of sequenced genomes the available data on genetic variation has increased substantially and this

has the added benefit of the inclusion of many more rare variations (for example, SNPs with low minor allele frequencies).

The average number of differences between genome of two individuals remains relatively constant at a rate of approximately 0.1% of nucleotide sites, on average, one variant per 1000 base pairs (Wang, 1998).

The most common type of genetic variation in the human population is called a single nucleotide polymorphism, or SNP. There are an estimated 10 million common SNP sites constituting 90% of the variation in the population (Reich *et al*., 2003). Figure 7 illustrates that SNPs are single-nucleotide substitutions of one base for another. In the human genome, each SNP location can have only four different versions: one for each nucleotide, adenine (A), cytosine (C), guanine (G) and thymine (T). In order to be classified as a SNP, a minimum of two versions of a sequence must occur in at least 1% of the population. There are approximately 10 million SNPs within the 3-billion-nucleotide human genome which translates to about one in every 300 nucleotide base pairs.

SNPs occur throughout the genome and can act as biological markers to track an associated disease, or the inheritance of disease genes within families and/or to predict the risk of certain individuals to develop a particular disease (Genetic Home Reference handbook, US National Library of Medicine URL: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp). They can be used as markers to compare regions of the genome between cohorts, i.e. individuals presenting the disease versus those who do not. SNPs are also helpful for locating a gene that produces phenotypically different individuals (Zollner and von Haeseler, 2000).

Researchers have found that ordinarily SNPs are not responsible for a disease state but instead serve as biological markers for locating a disease on the human genome map. This is due to the fact that SNPs associated with a disease are usually located in close proximity to the gene associated with a certain disease (Folkersen *et al*., 2010).

For this reason SNPs can be used to search for and isolate the disease-causing gene by using association studies to detect differences between the SNP patterns of the two groups and, thereby indicating which pattern is most likely associated with the disease-causing gene (Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources URL:http://www.ncbi.nlm.nih.gov/).

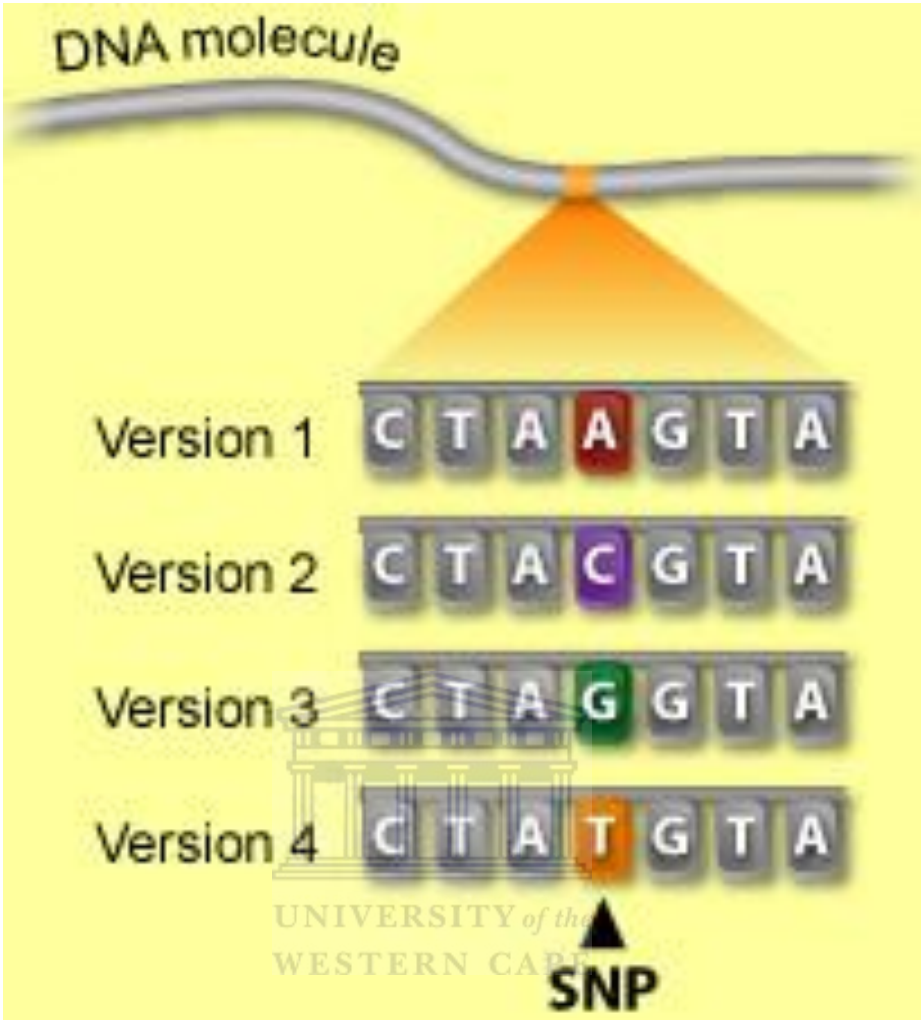


Figure 7 Single Nucleotide Polymorphisms (SNPs)

SNPs are single-nucleotide substitutions of one base for another. In the human genome, each SNP location can have only four different versions: one for each nucleotide, adenine (A), cytosine (C), guanine (G) and thymine (T). In order to be classified as a SNP, a minimum of two versions of a sequence must occur in at least 1% of the population. There are approximately 10 million SNPs within the 3-billion-nucleotide human genome which translates to about one in every 300 nucleotide base pairs (URL http://learn.genetics.utah.edu/content/health/pharma/snips/).

If SNPs are randomly distributed across the genome, one can assume that the difference in SNP count per gene could be proportional to the length of the gene. This was confirmed by Zhao *et al*., 2003 with a finding that the number of SNPs per chromosome was correlated with the length of the chromosome. They estimated that there was on average 8.33 SNPs per 10kb across the genome.

As discussed earlier, as more populations' genomes are sequenced, the number of known SNPs increases, and so these values may reflect the limits of our knowledge, in 2003 and today, rather than the absolute number of SNPs. This is especially true given the fact that African genomes are the most diverse and yet only a very limited number have been sequenced (The 1000 Genomes Project Consortium, 2010). It is therefore probable that the SNP density data may change as our knowledge base of human genomes increases.

Also important to note is that the genome structure and genetic variation differ considerably between different populations, but African populations, in particular, are more genetically diverse than European and Asian populations. This is generally believed to be the result of a bottleneck that occurred when population groups migrated out of Africa, from where it is thought that humans originated. This is known as the “Out of Africa” hypothesis (Stinger *et al*., 1988). Nucleotide diversity and haplotype diversity decrease in populations according to their geographic distance from Africa (Teo *et al*., 2005; Conrad *et al*., 2006; Jakobsson *et al*., 2008; Tishkoff *et al*., 2009).

Table 2 classifies SNPs found in both the coding and non-coding region of a gene and describes their functional effects in humans.

Table 2 Functional effects of SNPs.

SNPs are found in both the coding and non-coding region of a gene and have different functional effects in humans.

Type	Properties
Coding region	
Synonymous	Both alleles produce the same polypeptide sequence
Nonsynonymous	Both alleles produce a different polypeptide sequence
Nonsense	Results in a premature stop codon
Missense	Results in a different amino acid
Non-coding region	
Promoter/regulatory region	Does not change the amino acid, but can affect the level, location or timing of gene expression
Upstream	Change the sequence without any known transcription factor binding site
Enhancer	Alter the binding site for a transcription factor
Gene Splicing Site	Break the consensus splicing site sequence

Figure 8 shows that there is a very strong correlation between the number of SNPs per chromosome and the length of the chromosome. Therefore, I have also reviewed SNP count and SNP density per 10kb to determine if these are related to gene status and whether there is a difference in the degree of variation between disease genes and the control set.

1.4.3. Gene Length

Research has shown that proteins known to be associated with human disease show a clear trend for being longer than the rest of the proteins in the human genome. For clarity, a protein is involved in a disease when its analogous gene has impaired function of expression due to a mutation that is so severe that it produces a certain phenotype that is classified as disease (Lopez-Bigas *et al*., 2004, Adie *et al*., 2005). The Lopez-Bigas study concluded that the average length of a disease protein is 699 amino acids, while the average lengths of non-disease proteins range from 460 to 508. The Adie *et al* study found that the median length of disease genes, those found in OMIM, was 27kb while those genes not found in OMIM have an average length of 19kb. These results are endorsed by the fact that during DNA replication, the mutation rate could be as high as 1 mistake per 100 (10-2) to 1,000 (10-3) nucleotides and that at this rate there is a substantial escalation in the risk of disease-causing gene mutations in longer genes compared to shorter genes (Johnson *et al*., 2000).

Table 3 illustrates the difference in protein length and gene length that has been published in previous literature.

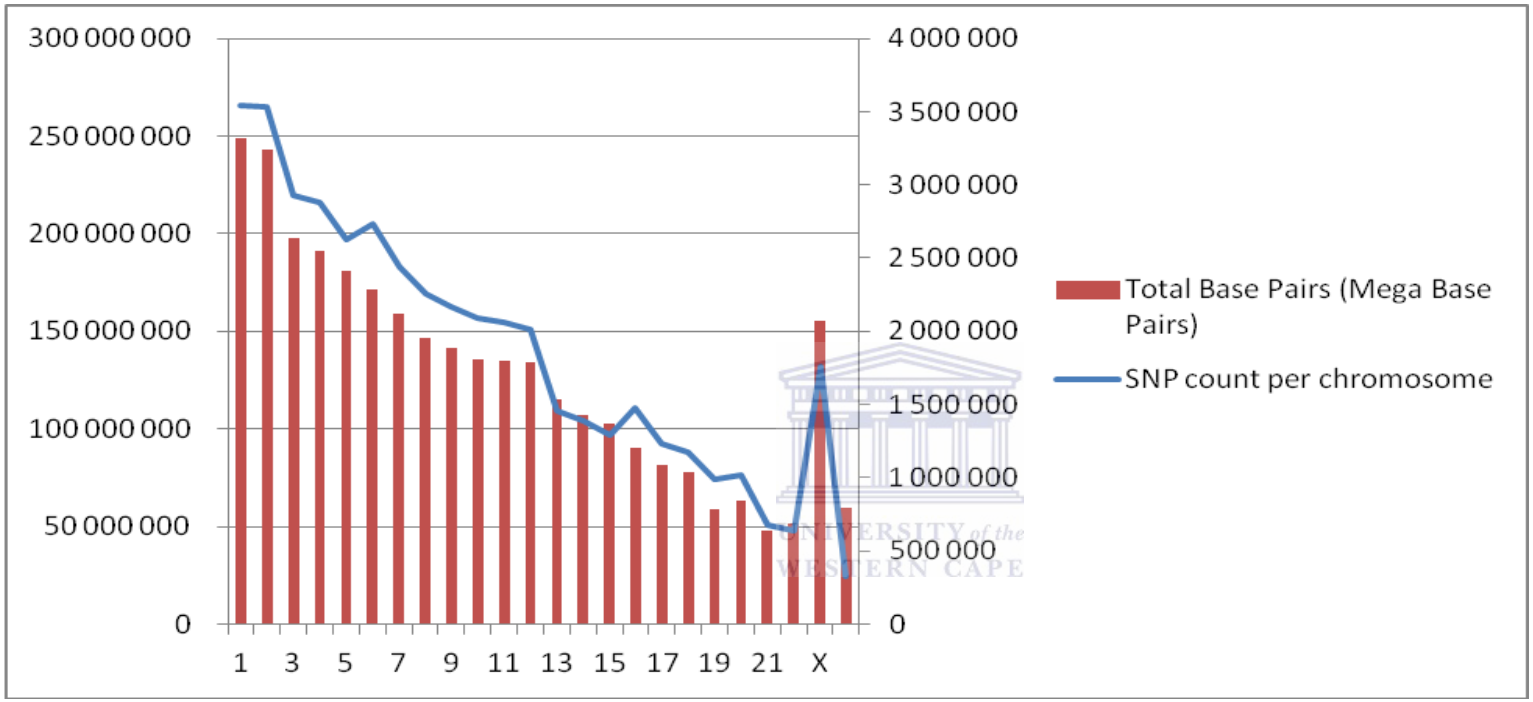


Figure 8 SNP number per chromosome (based on Ensembl build GRCh37.p7, Feb 2009)

An illustration of the total number of SNPs per chromosome versus the chromosome length shows that there is a correlation between the two suggesting that the SNP count is proportional to the chromosome length. (Data obtained from Ensembl).

Table 3 Previously reported data of gene and protein length

Illustration of the comparative results of protein length, calculated in amino acids, and gene length, calculated in base pairs, reported in previous literature.

	Results	References	Data	Source
DISEASE				
Protein length	699 amino acids	Lopez-Bigas et al., 2004	1567 genes associated with hereditary diseases in humans and their protein sequences	OMIM
Gene length	27kb	Adie et al., 2005	1,084 genes	OMIM
N**ON-DISEASE**				
Protein length	460-508 amino acids	Lopez-Bigas et al., 2004	10 000 similarly sized protein sets selected from all the human genome proteins	Ensembl v15.33.1
Gene length	19kb	Adie et al., 2005	~ 18,000 known genes not known to be involved in human disease	Ensembl (2005)

1.4.4. Position Effect: Distance from chromosome end

The theoretical basis for investigating the effect of the physical distance that a gene lies from the chromosome ends, based on its position in relation to the telomere/centromere and gene-start or gene-end, is the possibility that more recombination may occur in regions closer to chromosome ends, as they may have more flexibility to enable recombination.

Once the human genome sequence was compiled and analyzed by the International Human Genome Sequencing Consortium in 2001 it was confirmed that the rate of recombination near the centromere, or region between the p and q arm, is generally repressed while the rate generally increases near the telomeres, or chromosome ends (Ames *et al*., 2008, International Human Genome Sequencing Consortium, 2001).

A chromosomal break can cause a "loss-of-function" phenotype by disrupting the coding sequence of a gene, or by separating the gene from its adjacent regulatory region. It could, alternatively, initiate a "gain-of-function" phenotype by splicing the regulatory sequences from one gene to the coding sequences from another gene and causing alternate expression. In either instance, the breakpoint provides an invaluable clue to the exact physical location of the disease gene. A pitfall to this is that sometimes, due to the tertiary structure and folding dynamics of the DNA molecule, breakpoints can alter expression of a gene located hundreds of kilo base pairs away.

Adie *et al*., 2005 found that there was a noteworthy difference ($p<0.01$) in the distance of a disease gene to the nearest neighbouring gene with a median value of 52kb, while genes not known to be involved in disease had a median distance of

46kb (Adie *et al*., 2005). To my knowledge, however, there are no published figures on the distance of a disease gene or non-disease gene to the chromosome end.

Figure 9 gives a very broad illustration of the position of all known protein-coding genes as well as other genes (novel protein-coding genes, Pseudogene, miRNA genes, rRNA, snRNA genes, snoRNA genes, and other miscellaneous RNA genes) on the chromosome in relation to the chromosome ends showing that, in general, genes are fairly evenly distributed along the chromosomes.

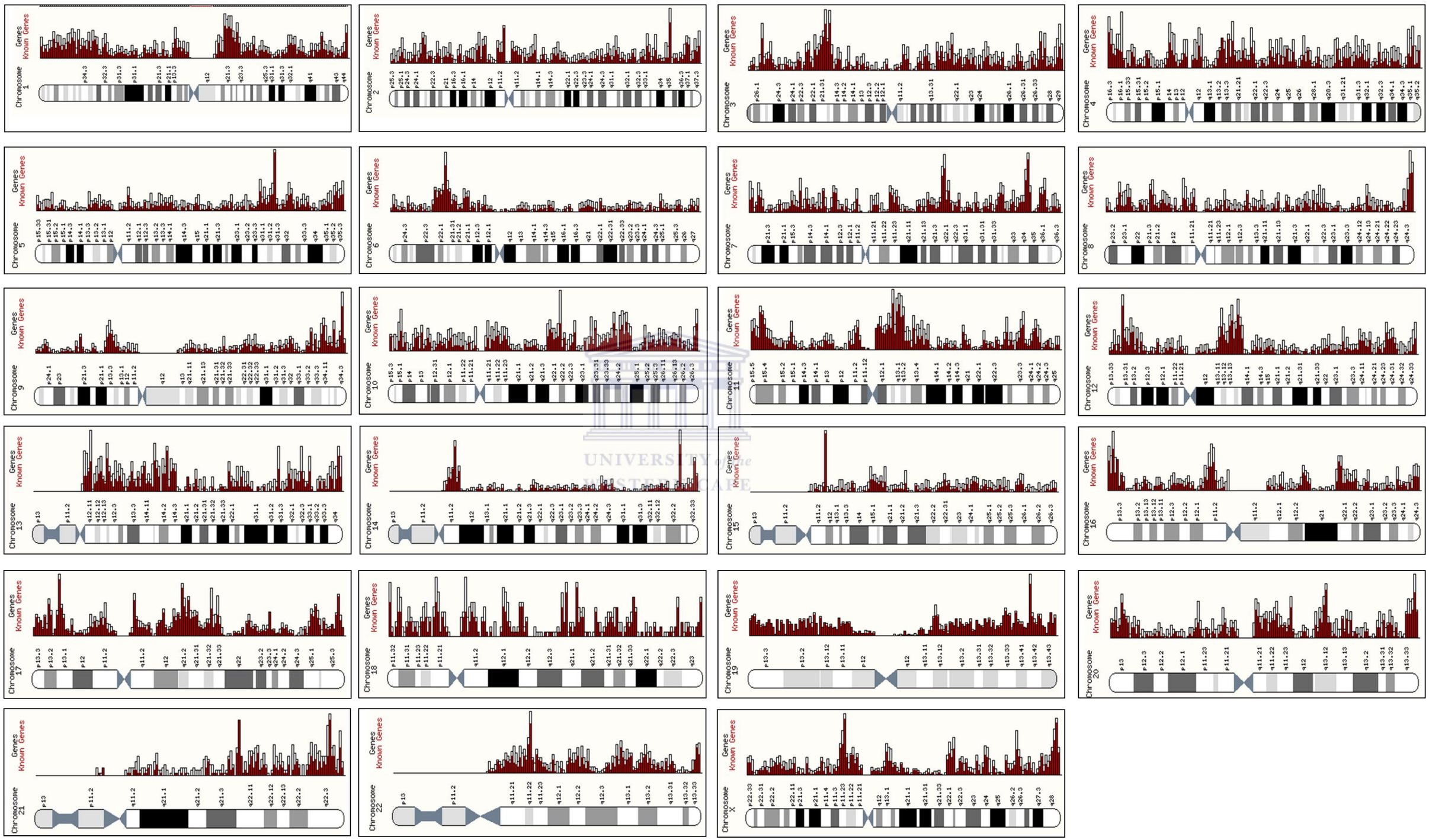


Figure 9 Gene Distribution per chromosome

A graphical illustration of the position of all known protein-coding genes as well as other genes, such as pseudogenes, on the chromosome in relation to the chromosome ends. This shows that, in general, genes are fairly evenly distributed along the chromosomes.

1.5. Data Resources

1.5.1 Online Mendelian Inheritance in Man (OMIM)

The resource used for disease-associated gene predictions was the Online Mendelian Inheritance in Man (OMIM) database (URL: http://omim.org, McKusick-Nathans Institute of Genetic Medicine, John Hopkins University (Baltimore, MD) 2012), human genome reference assembly GRCh37.3 from October 2011). This is a comprehensive and curated database of human genes and genetic phenotypes detailing all disease-associated genes (~21 000 genes) that have been implicated in disease. The OMIM data is based on published, well-referenced and current literature. At present, OMIM reports ~2800 defective genes or loci that have been described conclusively as that cause of "*a single disorder with a sole means of transmission*" (Mendelian diseases), while the remaining entries are described as "*being of uncertain inheritance*", either because the same condition can result from more than one genetic defect (complex diseases), or because the phenotype may overlap with another (Parton. 2003).

The data is freely available and updated daily. This database was initiated by Dr. Victor A. McKusick in the 1960's in order to assemble all known Mendelian traits and disorders at one source. In a joint collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins, OMIM went online in the 1980's.

From here on, the following description of what constitutes a gene will hold true for all references made to OMIM throughout this study unless explicitly stated otherwise; gene includes 5'UTR, exons, introns and 3'UTR.

When reference is made to disease-associated genes this includes protein-coding genes, non protein-coding genes (pseudogenes) and RNA. The diseases included in the OMIM database are no longer exclusively single gene (Mendelian) diseases but also include multifactorial diseases, also called complex or polygenic diseases, caused by both environmental factors as well as mutations in multiple genes such as cancer, chromosome abnormalities, resulting from abnormalities in chromosome number and structure e.g. down syndrome and lastly mitochondrial diseases, a genetic disorder caused by mutations in the mitochondrial DNA, e.g the eye disease called Leber's hereditary optic atrophy. The genes included in OMIM are genes with a direct relationship to these disease and those of known function (Amberger *et al*., 2009).

1.5.2. The Ensembl Project

In order to distinguish between disease-associated genes and genes not previously reported as being implicated in any known disease and to circumvent the problem of genes that have not, as yet, been implicated in disease, or false positives, I constructed a list of genes from the whole genome, using the Ensembl project database (Flicek *et al*., 2011 URL: http://www.ensembl.org/), and removed all genes that were mentioned in OMIM.

Ensembl is a joint project between European Bioinformatics Institute (EBI), part of the European Molecular Biology Laboratory (EMBL), and the Wellcome Trust Sanger Institute (WTSI) in Cambridge, United Kingdom started in 1999 prior to the completion of the first draft human genome. The main goal of the project was to automatically annotate the genome. The idea was to integrate this annotation with

other available biological data and make it publicly available via the web in the timeliest and most cost effective manner (Flicek *et al*., 2011).

Ensembl is an open-source database that offers specific tools for comprehensive data analysis and mining with the DNA sequences and assemblies provided by various projects around the world. The Ensembl BioMart data-mining tools provide an easy to use, generic system, capable of handling large amounts of input data and BioMart is able to perform advanced queries of numerous biological data through a single web interface (Smedley *et al*., 2009, URL: http://www.ensembl.org/biomart/martview/). The Ensembl data-mining tools used in this study are; (1) EnsemblMart version 65, the .ftp file contains the automated annotation of eukaryotic genomes and (2) Ensembl SNPMart version 65, the .ftp file contains Ensembl variation data from dbSNP and other sources.

From here on, the following description of what constitutes a gene will hold true for all references made to Ensembl throughout this study unless explicitly stated otherwise; all genes include the 5'UTR and 3'UTR regions as well as the exons and introns. Although one gene may have multiple RNA transcripts, the genomic DNA sequence was used to define the gene in this study.

The Ensembl human gene set, on which the control set is based, includes all protein-coding genes that have been automatically annotated using Ensembl genebuild pipeline as well as all transcripts based on mRNA and proteins in public scientific databases. All automatically-annotated pseudogenes and non-coding RNAs are also included.

1.6 Rationale for current study

The main motivation for this study stems from the reality that significant changes in gene structure as a consequence of recombination are known to underlie diseases. Non-homologous recombination has been implicated in many diseases (reviewed in Chen *et al*., 2010), and in particular to translocations. When a translocation occurs a segment of a chromosome shifts from one position to another, either within the same chromosome or to another chromosome. It is the latter translocation event that can result in gross changes to DNA structure. These chromosomal aberrations have been shown to be pivotal in tumour development and that ultimately leads to the formation of cancer genes, commonly referred to as oncogenes. Even though homologous recombination, and genetic variation, is a necessary process it is these extreme scenarios where non-homologous recombination introduces a disease phenotype through a gross change in variation that lead us to question whether more subtle changes in variation, introduced through homologous recombination, could also predispose a gene to causing disease through increased variation. Identifying the effects of homologous recombination on likelihood of genes to underlie disease could offer a novel, population-specific, approach to prioritising good candidate genes for further investigation.

It may soon be possible to extrapolate the occurrence of translocations to the wider genome for the purpose of identifying aetiological genes using a similar technique to Hamkin *et al.*, 2012. This is because during recombination events, genes that are in closer proximity to recombination hotspots are more likely to undergo recombinant

events that place them next to variant regulatory sequences, and may therefore be more prone to altered or dysregulated expression leading to pathogenic effects.

The aim was to develop a dataset of disease-associated genes and a control set of genes that have not previously been reported as having an association with any known disease and to investigate various gene characteristics of these datasets and how they relate to the likelihood of a gene to underlie disease. The main issue was to determine whether a gene's proximity to a hotspot is likely to be of universal importance to disease gene status or whether it should be considered for a group of diseases characterized by chromosomal instability such as the case in cancers.

The main goal was to use this as a pilot study to investigate whether distance from recombination hotspots and frequency of recombination at the hotspot may be related to the likelihood of a gene to cause disease. The secondary goal was to attempt to identify additional gene features, such as gene length or SNP count, as a supporting characteristic to classify disease gene status. These goals could thereby potentially identify other criterion that can be used in concert with existing tools in the prediction of most likely disease gene candidates.

This goal was accomplished by answering the following questions:

(A) A general characterization of the properties of disease-associated genes compared to the control set

When comparing disease-associated gene characteristics to the same characteristics of genes in the control set:

i. Is there a difference in %GC content in disease-associated genes compared to the control set?

ii. Is there a difference in SNP count and SNP density in disease-associated genes and the control set?

iii. Is there a difference in length between disease-associated genes and the control set?

iv. How does the distance that disease-associated genes and genes in the control set sit from the end of the chromosome impact on the gene status?

(B) A specific characterization of recombination hotspots in disease-associated genes compared to the control set:

When comparing all genes that contain an internal hotspot and genes with no internal hotspot:

v. Is there a difference in %GC content between genes with hotspots and genes with no hotspots?

vi. Is there a difference in SNP count and SNP density in genes with hotspots compared to genes with no hotspots?

vii. Is there a difference in length between genes with hotspots and genes with no hotspots?

viii. How does the distance that all genes with a hotspot and those with no hotspot sit from the end of the chromosome impact on gene status?

When comparing the recombination hotspot either lying within the gene or nearest to the gene:

ix. Are disease-associated genes more likely to contain a recombination hotspot than genes in the control set?

x. Do the hotspots lying in disease-associated genes exhibit a higher frequency of recombination than the hotspots in genes in the control set?

xi. Do the hotspots lying nearest to disease-associated genes with no internal hotspot exhibit a higher frequency of recombination than the hotspots lying nearest to genes in the control set?

xii. Is there a difference in frequency, distance and overall scoring of hotspots nearest to disease-associated genes compared to hotspots nearest genes in the control set?

CHAPTER TWO

Material and Methods

This research focuses on the characterization of disease-associated genes and the control set of genes with no documented disease association with regards to the position of the gene within the genome structure.

An overview of the methods used to achieve this, as illustrated in Figure10, can be divided into four key sections;

(1) Developing the scoring system

(2) Compiling of the test sets and assembling of recombination hotspot data

(3) Collecting gene characteristics

(4) The data analysis

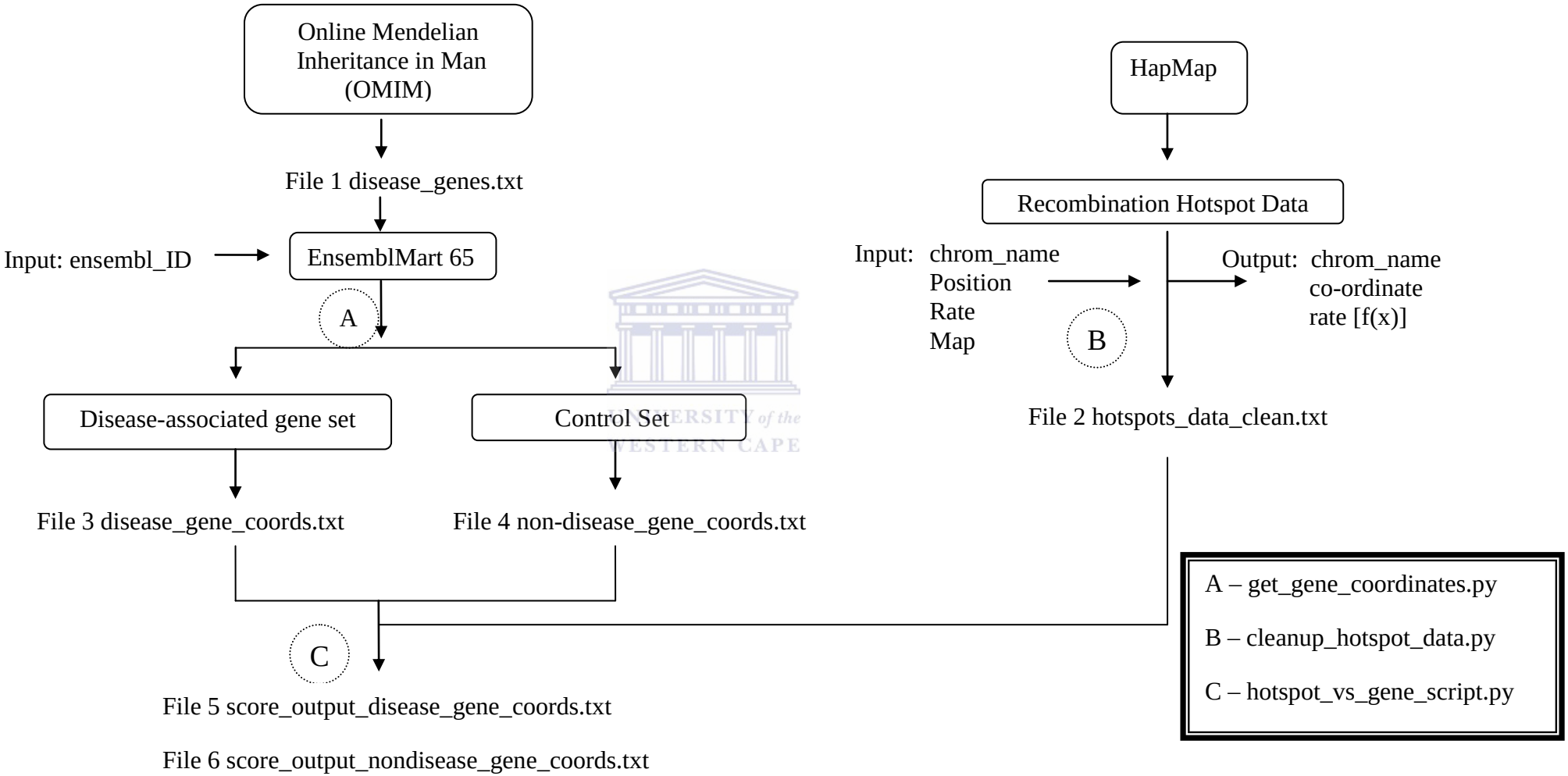


Figure 10 Overview of methods applied to compilation of test sets and assembly of recombination hotspot data

2.1. Developing the Scoring System

I developed a pragmatic, empirically determined scoring system to represent the distance from recombination hotspots and frequency of recombination at those hotspots, for all genes in the genome (Illustrated in Figure 11).

$$\text{Score} = \frac{\text{Frequency of recombination (in centiMorgans per mega base pair)}}{\text{Distance (in mega base pairs)}}$$

This scoring system took into consideration all recombination hotspots on the same chromosome as the gene, and the frequency of recombination events at flanking hotspots as well as recombination hotspots that lie further away. A pipeline was then established to calculate the scores for all genes in a given population. The design enables us to use the formula broadly on any populations given the positions of recombination hotspots in that population, however since this is a pilot study I initially wanted to establish if there was a relationship between hotspots and disease-associated genes before extending the study to specific populations.

The scoring system also incorporates the possibility that a gene lies very close to a recombination hotspot with a low frequency of recombination events, but that further away there may be a recombination hotspot of high frequency of recombination. This is because such a high frequency recombination hotspot could potentially have a greater effect on co-segregation of the gene with its regulatory sequences than the more proximal hotspot. This required a matrix of scores to be established for each gene, relating to multiple recombination hotspots in the region surrounding the gene and the frequency of recombination at each hotspot. For each disease-associated genes all the hotspots on the same chromosome were considered and a score was

generated, then the highest scoring hotspot for each disease-associated gene was selected for further analysis. In this way the dominant effect of local hotspots on the gene was determined. Consideration was also given to the possibility that a recombination hotspot occurs within the span of the gene and these instances were flagged with an artificially assigned distance value of 0.1. (This value was chosen as 0.1 cannot be a naturally occurring measure of distance in base pairs). The scoring system allowed us to assign a score to each gene according to its relationship with nearby recombination hotspots. The score reflects the likelihood that the coding section of the gene could recombine causing different combinations of alleles to come together; and consequently whether recombination events could cause altered functioning of that gene.

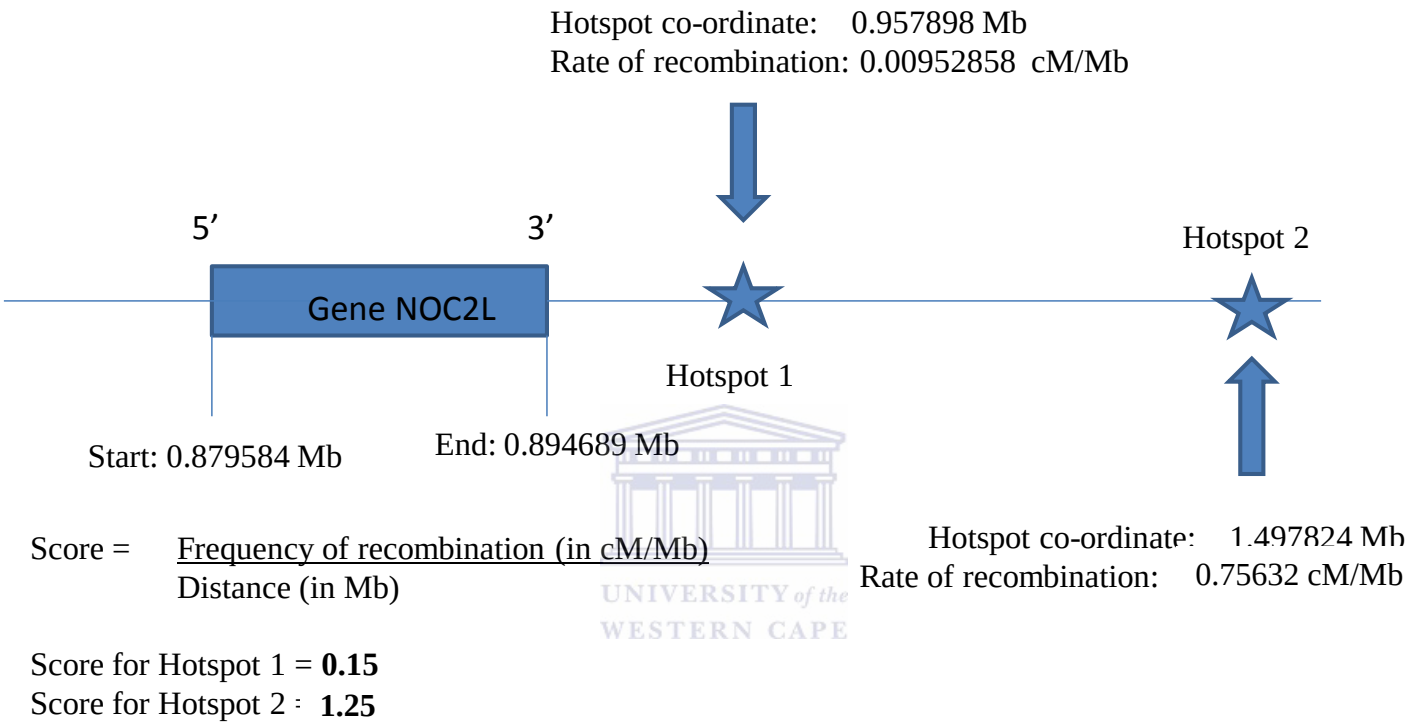


Figure 11 **Graphical illustration of scoring system metric**

A graphical illustration showing the concept behind the design of the scoring system used for this study. The hotspot with the higher score was selected for further analysis. This figure demonstrates that a hotspot with a lower frequency of recombination (f(x)) lying nearer to the gene has less impact on the gene than a hotspot with a higher frequency of recombination lying farther away from the gene.

2.2. Compilation of Test Sets and Recombination Hotspot Assembly

The first step of this section involved the assembly of a list of genes as control sets. There were two sets of genes required to test whether distance/frequency metrics of recombination hotspots correlate with the likelihood of a gene to have an aetiological role; (1) disease-associated genes; and (2) genes that have no documented association with a disease.

The OMIM database was selected as the source of disease-associated genes for several reasons: (1) it represents a very stringent, well curated database that is updated daily and therefore contains validated, up-to-date disease phenotype/genotype information; (2) only validated gene-disease information is included so the proximity to tagging SNPs, as in GWAS studies is not considered sufficient evidence to associate a gene with a disease – further evidence is required to implicate the gene with the disease. OMIM contains not only information about Mendelian diseases but also information of genes implicated in complex diseases, in which multiple genes contribute to the disease phenotype but each gene is insufficient on its own to cause the disease phenotype. I used the gene list from OMIM for both Mendelian diseases and complex diseases because it is known that with Mendelian diseases there is a straight forward pattern of inheritance, a single gene mutation is sufficient to cause a disease phenotype. However, in complex diseases the cause is more difficult to trace. In complex diseases it is a combination of factors, such as allele variation or environment, which lead to disease phenotype. I hypothesis that if multiple genes work in combination to cause a certain disease phenotype, if each of the genes has increased variation due to the proximity to a

hotspot then the overall increased variation of all causative genes may still influence the likelihood of the combined genes to cause the phenotype.

I also reviewed other databases such Genetic Association Database (GAD), (http://geneticassociationdb.nih.gov/), and Human Gene Mutation Database (HGMD), (http://www.hgmd.cf.ac.uk/ac/index.php), as appropriate sources. However, I chose to select the OMIM database as the only source for disease gene data for the following reasons: (1) OMIM is freely available, (2) there was a high level of common genes; (3) less stringent data curation (OMIM data manually curated, no automation); and (4) these other databases include negative results of association studies which makes disease gene selection complicated.

The OMIM .ftp file generated a file that included MIM numbers, Entrez gene IDs (the phenotype gene IDs were excluded) plus approved gene symbols.

FTP file: (ftp://anonymous:tracey%40sanbi.ac.za@grcf.jhmi.edu/OMIM/)

I input the Entrez gene IDs into Ensembl BioMart software, EnsemblMart 65 database which generated the control set list into a text file (file 1 – disease_gene.txt). All duplicate genes were removed.

The next step was to download the recombination hotspot data. The recombination rates and hotspot data were compiled from the genotyping data and these genotypes correspond to the four HapMap populations: 90 Caucasian individuals from 30 CEPH family trios, 90 individuals from 30 family trios from the Ibadan people (Nigeria, West Africa), and 90 unrelated individuals from Southeast Asia (45 Han Chinese from Beijin, and 45 Japanese from Tokyo). This data was downloaded for

the combined populations for the reason that this is a pilot study and I initially wanted to establish if there was a relationship between hotspots and disease-associated genes before extending the study to specific populations.

Recombination rates and hotspot data were extracted from the bulk data download on the HapMap website. Using the HapMap II release 28 data published in August 2010 based on NCBI B37 assembly. A Python script (*cleanup_hotspot_data.py*) was used to extract the necessary data from the HapMap download, removed all unnecessary information and generated a text file (file 2 – hotspots_data_clean.txt). Using the Ensembl gene ID as input, I used a Python script (*get_gene_coordinates.py*) to extract the gene co-ordinates as well as the chromosome name from the Ensembl database, human genome assembly (GRCh37) version 65, released in December 2011, contains 23 171 gene models and the manually curated annotation from Havana of 45 484 genes to create a final set of 56 478 genes, including coding genes, non protein-coding genes (pseudogenes) and RNA genes. This was the data used to create the control set of genes from the whole genome. Using file 1 as input, a second Python script (*hotspot_vs_gene_script.py*) divided all the genes in Ensembl into two categories. The Python script selected all genes that were previously identified by OMIM and separated the data into two text files (file 3 – disease_gene_coords.txt, those found in OMIM and file 4 – nondisease_genes_coords.txt, all other genes).

The output files included the Ensembl gene ID, the gene name, chromosome name, and the gene start and gene end co-ordinates for both lists of disease-associated genes and the control set. Using these files the script then selected the top scoring hotspot

data for each gene using the scoring system I developed. The data generated two output files (file 5 – score_output_disease_gene_coords.txt and file 6 – score_output_nondisease_gene_coords.txt). These files contained information for the gene (Ensembl ID, maximum score output, chromosome name, gene name and gene start and end co-ordinates) as well as hotspot information for that particular gene (hotspot position, hotspot frequency and distance of the gene to the nearest hotspot).

2.3. Collecting Gene Characteristics Data

For the purpose of this thesis I have identified certain gene characteristics that I have reviewed between disease-associated genes and the control set. These include gene length, GC content, SNP density and distance of a gene from the chromosome ends, presence/absence of a recombination hotspot, distance to nearest hotspot and frequency of recombination of internal/nearest hotspot. The rationale for this is explained in chapter 1.

2.3.1. Determining Base Composition: GC Content

The percentage GC content was extracted from the Ensembl BioMart database, EnsemblMart_65, by uploading the Ensembl gene ID. Using the 'Homo Sapiens genes (GRCh37.6) dataset, I refined the search by selecting the ID list limit in the 'gene' subheading under the FILTERS table and entered the Ensembl gene ID. Using the %GC content under the 'features' subheading in the ATTRIBUTES table to further refine the search.

2.3.2. Calculating Genetic Variation: SNP count

Our SNP density data were obtained using a Python script (*get_all_snp_info.py*) to download all pre-calculated data for SNPs in batches per chromosome from the Ensembl BioMart database, SNPMart_65 (December 2011), using Entrez gene IDs. A second script (*split_genes_by_chrom.py*) was used to categorize the SNPs by chromosome into disease-associated genes and the control set in order to make SNP counting computationally manageable.

A third script (*get_gene_snp_count.py*) was used to count the SNPs for each gene, chromosome by chromosome to generate separate output files (SNP_count_disease_genes_chrom_(x).txt). The output files included the following data; Ensembl gene ID, gene name, gene chromosome name, gene start and gene end, gene length as well as SNP count and SNP density.

2.3.3. Establishing Gene Length

A Python script (*get_gene_coordinates.py*) was used to extract the co-ordinates of the gene per chromosome from the Ensembl BioMart database, EnsemblMart_65. The data, which included gene start (5'UTR) and gene end (3'UTR) positions on the chromosome, was input into an Excel spreadsheet and the SUM function was used to determine the length of all the genes in both output files (i.e. disease_gene_coords.txt and non_disease_gene_coords.txt) by using gene end – gene start.

2.3.4. Mapping Position Effect: Distance from chromosome end

The distance of the gene from the chromosome end was calculated separately for each arm. The p arm was calculated using the position of the gene in relation to the centromere start position and the q arm was calculated using the position of the gene in relation to centromere end position.

Using the gene co-ordinates (gene start (5'UTR) and gene end(3'UTR)) and the centromere start position (p arm) as well as end position (q arm), I calculated the distance of each individual gene from the chromosome start, at position 0 base pair, by subtracting gene start from chromosome start. I then calculated the distance of the gene from the chromosome end, using the chromosome length data from Ensembl,

by subtracting chromosome end (or chromosome length) from gene end. The results were generated into an excel file, per chromosome (file 7 – distance_from_chromo_ends.xls). I then used the MIN function in to determine the shortest distance of each gene from a chromosome end on each arm.

2.4. **Data analysis**

In statistics there are two main fields of study: descriptive and inferential. Descriptive statistics gives a graphical representation of a data set while inferential statistics uses mathematical probabilities by attempting to make the best possible conclusion from a small sample of data by extrapolating the data to make inferences about the larger population (*Introductory Statistics*, 2nd Edition (1995) Wiley). This thesis focuses on descriptive statistics as the analysis performed is on all the known genes in the entire human genome and therefore there is no need to make use of inferential statistics to make any assumptions about the larger population. The dataset is complete and not a sample from a larger population.

For this study I use boxplots (Figure 12), a graphical tool for the easy visualization of data. The advantages of a boxplot include the ability to graphically display the layout and spread of a dataset at a glance, they provide an indication of the data's symmetry and skewness and one can quickly and easily compare more than one dataset side-by-side. The main attribute of a boxplot is that, unlike other methods for displaying data, it shows outliers, points that lie beyond one and a half times the length of the box. Mild outliers lie outside the lower quartile region while those outside the upper quartile are considered extreme outliers. Outliers usually only represent a small portion of the data and there are many reasons for the appearance of outliers; human error or simply through natural deviations in populations. An outlier could also be the result of a flaw in the research hypothesis which may warrant further investigation.

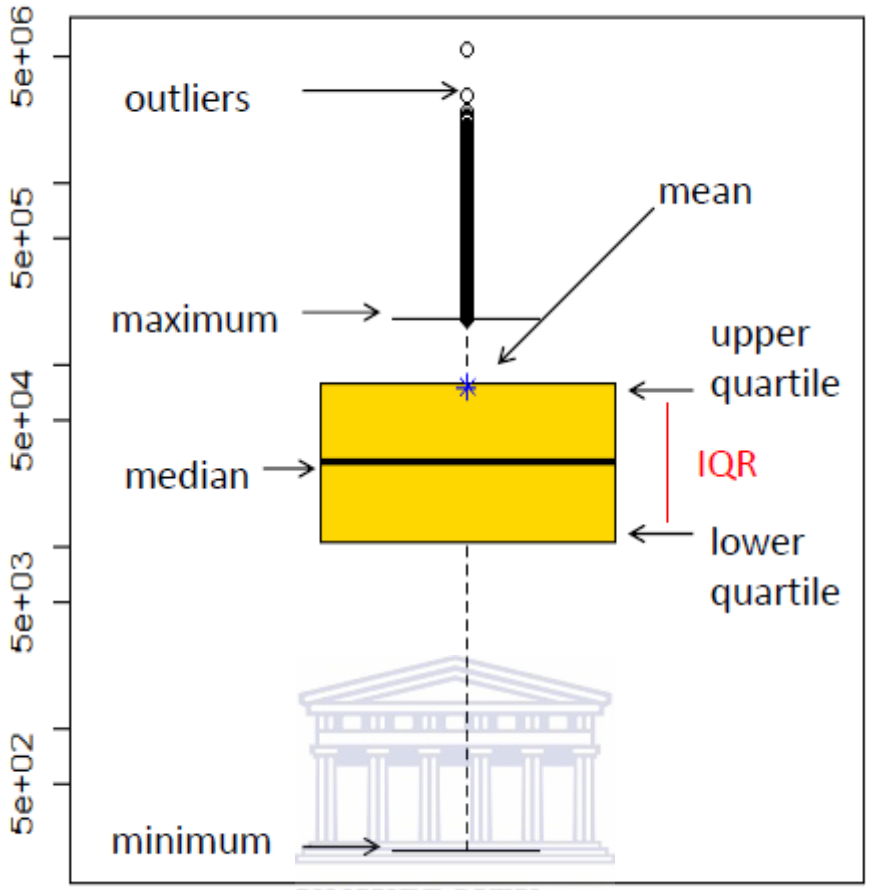


Figure 12 Boxplot - Display of Distribution

The black line dividing the box represents the median of the data. The lower and upper edges of the box are represented by the lower and upper quartiles of the data. This means that 50% of the observations fall within the range of the box, while 25% fall below and above the lower (Q1) and upper (Q3) quartiles, respectively. The whiskers mark those values which are 1.5 x IQR from the upper and lower quartiles. The IQR is the inter quartile range: the distance between Q1 and Q3. The outliers are defined as any data values that are above or below the threshold; Q3 + 1.5 x IQR and Q1 – 1.5 x IQR

Boxplots are used to display differences between datasets and the distribution of the dataset elements without postulating about the underlying statistical distribution, in other words they are non-parametric. The layout of the data helps show asymmetry in the distribution of the data, as well as to identify outliers. The programme language R was used to generate boxplots and density plots by inputting the data into scripts. Analysis of the figures was used to answer the following questions:

(1) Are there any major differences in the characteristics of disease-associated genes compared to the control set?

(2) Are there any considerable differences in the characteristics of genes containing internal hotspots compared to genes with no internal hotspot?

(3) Is there a variation in the occurrence of recombination hotspots and frequency of recombination in disease-associated genes when compared to genes in the control set?

2.5. Software

Python was used extensively to generate and analyze data for this study (URL http://www.python.org, Python version 2.7.2). It is an open source, general-purpose, programming language used by Scientists for most kinds of software development. It was first released in 1991 and comes with extensive standard libraries and is very user friendly. It may be used in conjunction with other programming languages but is mostly used as a scripting language (Shaw, 2012).

The other programming language and software environment used extensively for this study, R, is also an open source language used for both developing statistical software and data analysis (URL http://cran.r-project.org/mirrors.html, R version 2.13.1). I use R for the purpose of generating all the boxplot and density plots.

CHAPTER THREE

Results

The disease-associated gene dataset used for this section of the study was extracted from the OMIM database and the control dataset from the Ensembl database. The disease-associated gene set included a total of 13095 genes while the control set totaled 38256 genes. The first sets of figures (Figures 13 – 16), generated using an R script, were designed to compare disease-associated genes and the control set. The second sets of figures (Figure 17 – 20) were designed to compare genes that contain an internal hotspot to those genes with no internal hotspots, also using an R script. The intention of the final sets of figures (Figures 21 – 24) was to determine whether the position and frequency of recombination of a hotspot could impact on the likelihood of a gene to underlie disease. Each set of figures consists of both a boxplot as well as a density plot. The two plots are two different ways to represent the same data. In any data analysis a visual representation of the data is the best way to illustrate the results and helps with the final interpretation of these results.

A total of 12 boxplot figures were analysed to answer the following questions:

3.1. Are there any major differences in the characteristics of disease-associated genes compared to the genes in the control set?

3.1.1. Analysis of the Base Composition: GC content

In the dataset of 13095 disease-associated genes, the average GC content per gene is 46.83%, while in the control set of 38256, the average GC content per gene is 45.60%.

The two plots, shown in Figure 13, represent the percentage distribution of GC content in disease-associated genes compared to genes in the control set. A review of the plots seems to confirm only a very minor difference in percentage distribution of GC content and in fact when comparing the mean and median values of both datasets the difference is negligible. The density plot shows an estimate of the actual densities and represents a different view of the same data shown by the boxplot. In this case, it also shows that there clearly is no substantial difference in the percentage distribution of GC content in disease-associated genes compared to genes in the control set. It may, however, be important to note the number of outliers that are visible in both the disease-associated gene and non-disease gene data. These outliers outside the upper quartile and are therefore considered extreme outliers. Since they differ substantially from the rest of the data, they may prove to be worthy of further investigation.

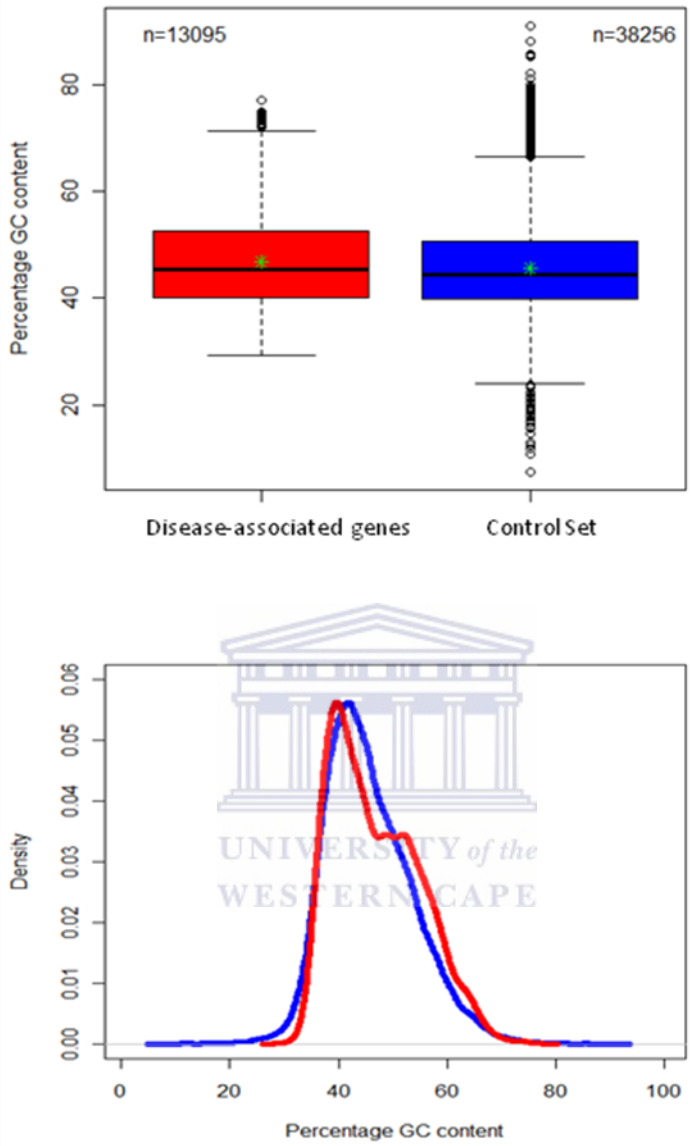


Figure 13 Percentage GC content in disease-associated genes compared to genes in the control set

This boxplot figure, and its associated density plot, illustrates the percentage GC content in disease-associated genes compared to the control set. It shows that there is a negligible difference between the two sets.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.1.2. Analysis of the Genetic Variation: SNP count

In the dataset of 13095 disease-associated genes, the average SNP count is 1089.52 per gene (including both exonic and intronic regions), while the average SNP density is 1.59 SNPs per 10kb. Conversely, in the control set of 38256 genes, the average SNP count is 251.67 per gene (including both exonic and intronic regions), while the average SNP density is 1.62 SNPs per 10kb.

The two plots, shown in Figure 14, represent the SNP density per 10kb in disease-associated genes compared to genes in the control set. A review of the plots validates that there is no considerable difference in SNP density per 10kb between disease-associated genes and genes in the control set with little difference in the mean and median values of both datasets. The density plot reiterates this finding. It may, however, again be important to note the substantial number of outliers visible in both the disease-associated gene and non-disease gene data. These outliers outside the upper quartile and are therefore considered extreme outliers. Since they differ substantially from the rest of the data, they may prove to be worthy of further investigation.

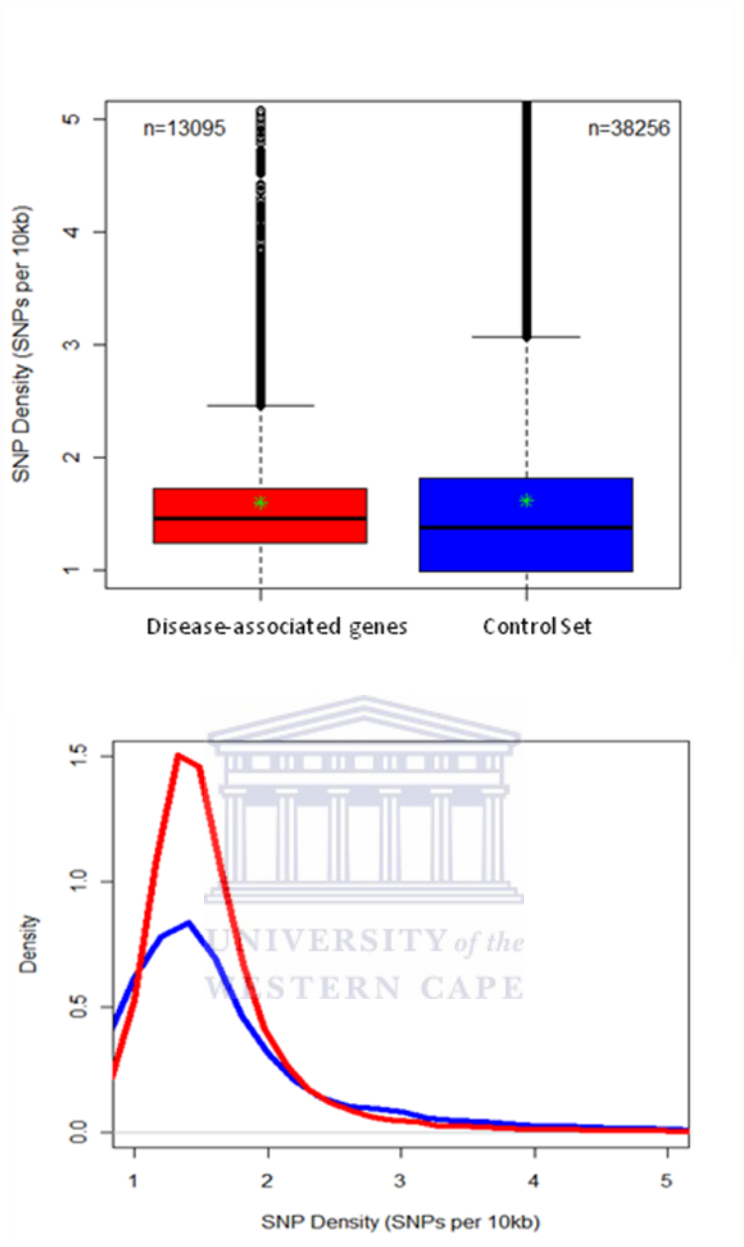


Figure 14 SNP density in disease-associated genes compared to genes in the control set

This boxplot figure, and its associated density plot, illustrates the difference in SNP density in disease-associated genes compared to genes in the control set. It shows a very little difference in SNPs per 10kb between the two sets.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.1.3. Analysis of Gene Length

In the dataset of 13095 disease-associated genes, the average length of a gene is 75518bp. The largest gene, transmembrane protease, serine 3 (TMPRSS3), is 5379013bp and the shortest gene, Phosphatidylinositol N-acetylglucosaminyltransferase subunit Y (RP11-466G12.4.1), is 216bp. Conversely, in the control set of 38256, genes, the average length is 17402bp. The largest gene, ATP/GTP binding protein-like 4 (AGBL4), is 1491058bp and the shortest gene, T cell receptor delta diversity 1 (TRDD1), is 7bp.

The two plots, illustrated in Figure 15, show the distribution of gene length in disease-associated genes compared to genes in the control set. A review of these plots indicates that there is a difference in distribution, with disease-associated genes showing a greater tendency to be longer than genes in the control set. The density plot shows a different view of the same data and it is clear that disease-associated genes are generally longer than genes in the control set.

The outliers, for this data, are concentrated towards the higher end of the scale indicating that it is the much longer genes in both groups that may, in reality, be the cause of this result as they are considerably different from the remaining data. These data points lie outside the upper quartile and are therefore considered extreme outliers. Since they differ substantially from the rest of the data, they may prove to be worthy of further investigation.

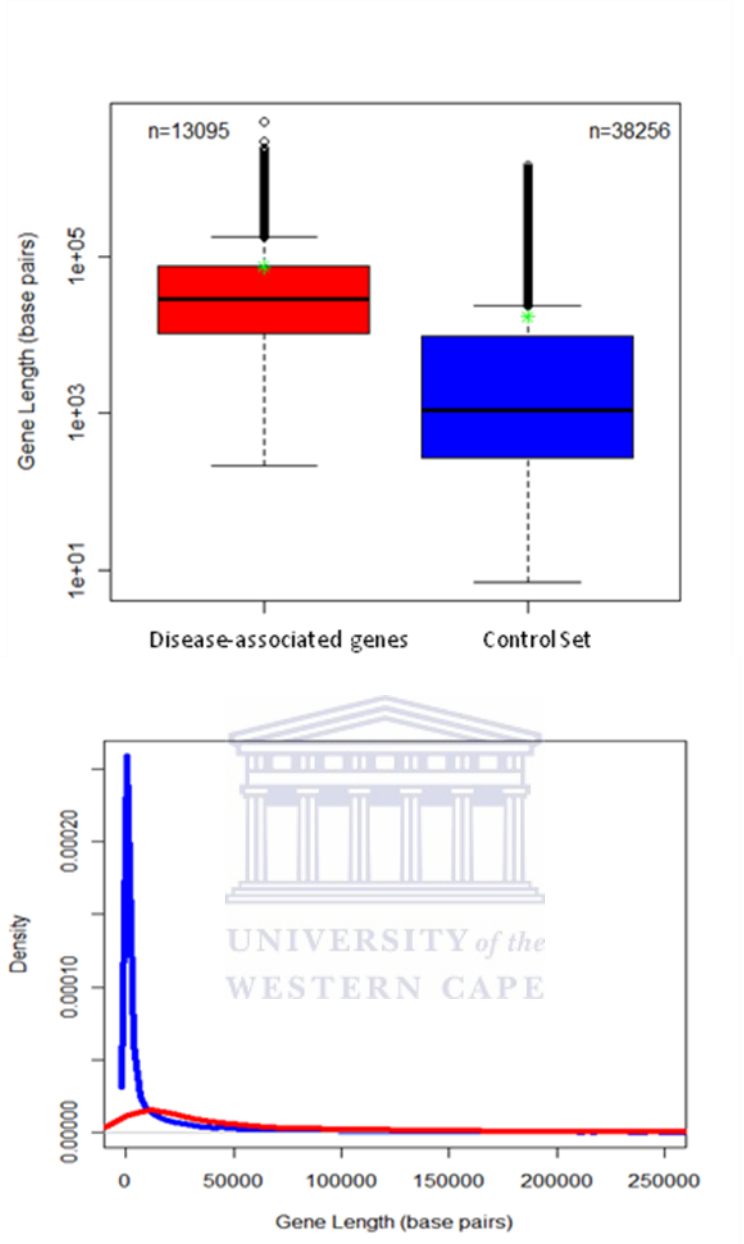


Figure 15 Length of disease-associated genes compared to genes in the control set

This boxplot figure, and its associated density plot, illustrates the difference in gene length of disease-associated compared the genes in the control set. It shows that disease-associated genes are longer in length than the genes in the control set.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.1.4. Analysis of the Position Effect: Distance from chromosome end

In the dataset of 13095 disease-associated genes, 4443 genes were closer to the chromosome start while 8652 genes were closer to the chromosome end. In the control set of 38256 genes, 12423 genes were closer to the chromosome start, while 25833 genes were closer to the chromosome end.

The two plots, illustrated in Figure 16, show the distribution of both disease-associated genes and genes in the control set within the chromosome length. A review of these plots confirms that there seems to be only a slight difference in the distribution, with neither disease-associated genes nor the genes in the control set showing any trend for occurring nearer to the chromosome ends. The density plot validates these findings with a clear indication that neither disease-associated genes nor genes in the control set tend to lie closer to the chromosome ends.

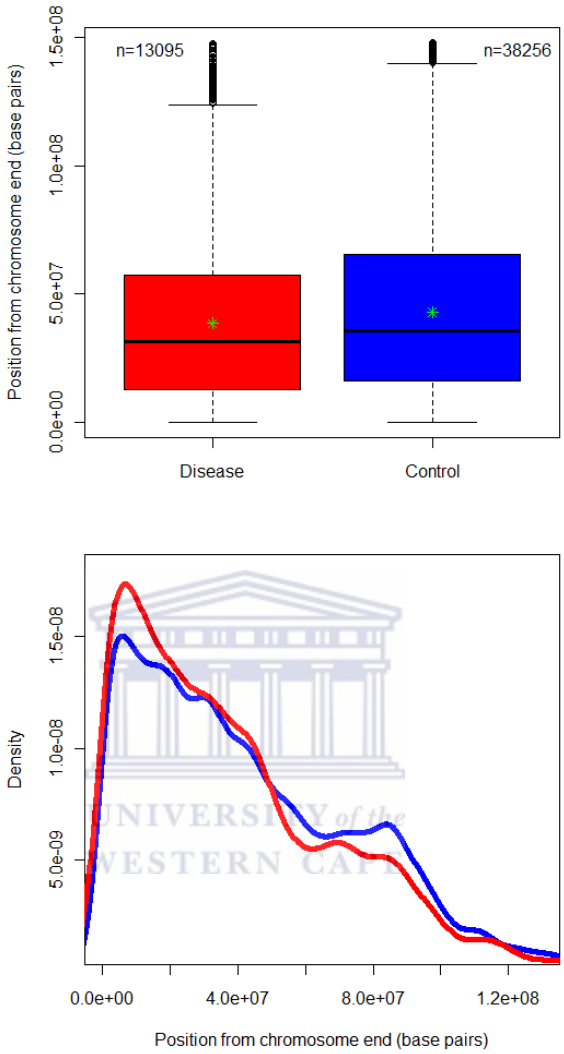


Figure 16 Position of genes on chromosome

This boxplot figure, and its associated density plot, illustrates the position of disease-associated genes from the chromosome ends compared the position of the genes in the control set from the chromosome ends. It shows that position of the gene on the chromosome has little effect on disease-gene status.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.2 Are there considerable differences in the characteristics of genes containing internal hotspots compared to genes with no internal hotspots?

By further categorising the disease-associated gene and list of genes in the control set into a subset of genes that; (1) contain an internal hotspot and (2) do not contain an internal hotspot, additional analysis determined whether the presence of the hotspot within the gene differed between disease-associated genes and genes in the control set.

3.2.1. Analysis of the Base Composition: GC content

From the datasets of 13095 disease-associated genes and 38256 genes in the control set (a total of 51351 genes), 34850 genes contained internal hotspots while 16501 contained no internal hotspots. The genes with an internal hotspot had an average GC content of 45.62% per gene while the genes with no internal hotspot had an average GC content of 46.52%.

The two plots, shown in Figure 17, illustrate the percentage distribution of GC content of genes that contain an internal hotspot compared to those that have no hotspot. A review of the plots confirms that there is no substantial difference in percentage distribution of GC content here either and in fact when comparing the mean and median values of both dataset the difference is again very small. The density plot illustrates that there is no substantial difference in the percentage distribution of GC content in genes that contain an internal hotspot compared to genes that do not contain a hotspot.

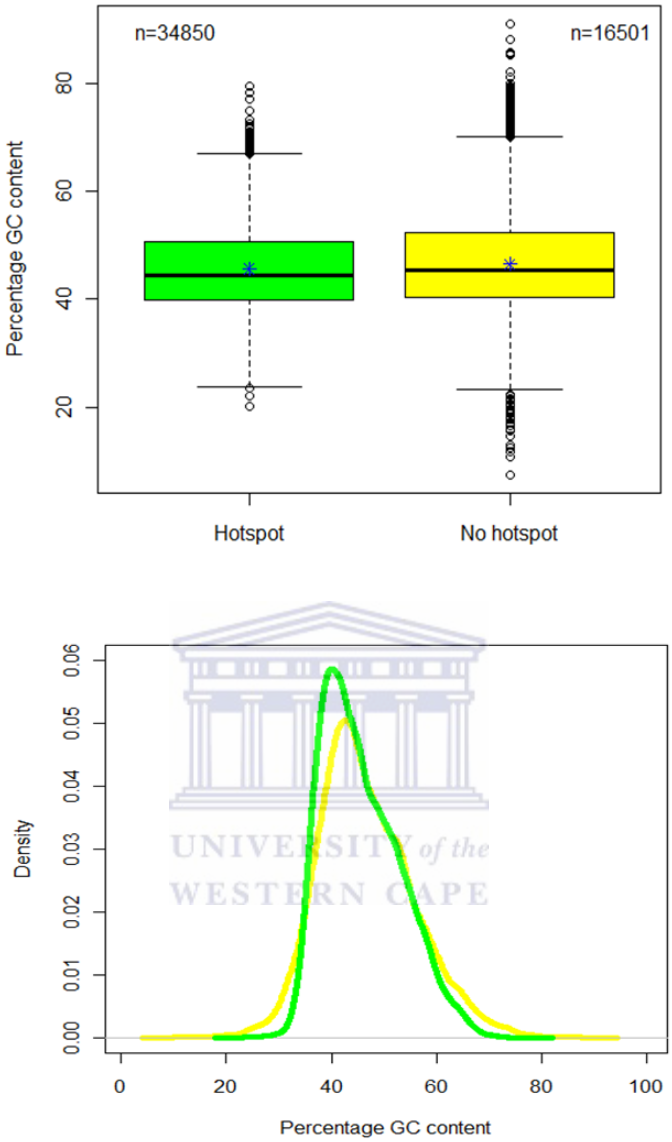


Figure 17 Percentage GC content in genes with an internal hotspot compared to genes with no internal hotspot

This boxplot figure, and its associated density plot, illustrates the percentage GC content in genes that contain an internal hotspot compared to genes that do not contain an internal hotspot. It shows a negligible difference between the two sets.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.2.2. Analysis of Genetic Variation: SNP count

From the datasets of 13095 disease-associated genes and 38256 genes in the control set (a total of 51351 genes), 34850 genes contained internal hotspots while 16501 contained no internal hotspots. The genes with an internal hotspot had an average SNP density of 1.633 per 10kb while the genes with no internal hotspot had an average SNP density of 1.5738 per 10kb.

The two plots, shown in Figure 18, illustrate the SNP density of genes that contain an internal hotspot compared to those that have no hotspot. A review of the plots seems to confirm that there is no substantial difference in the SNP density here either and in fact when comparing the mean and median values of both dataset the difference is again very small and the density plot reaffirms this finding.

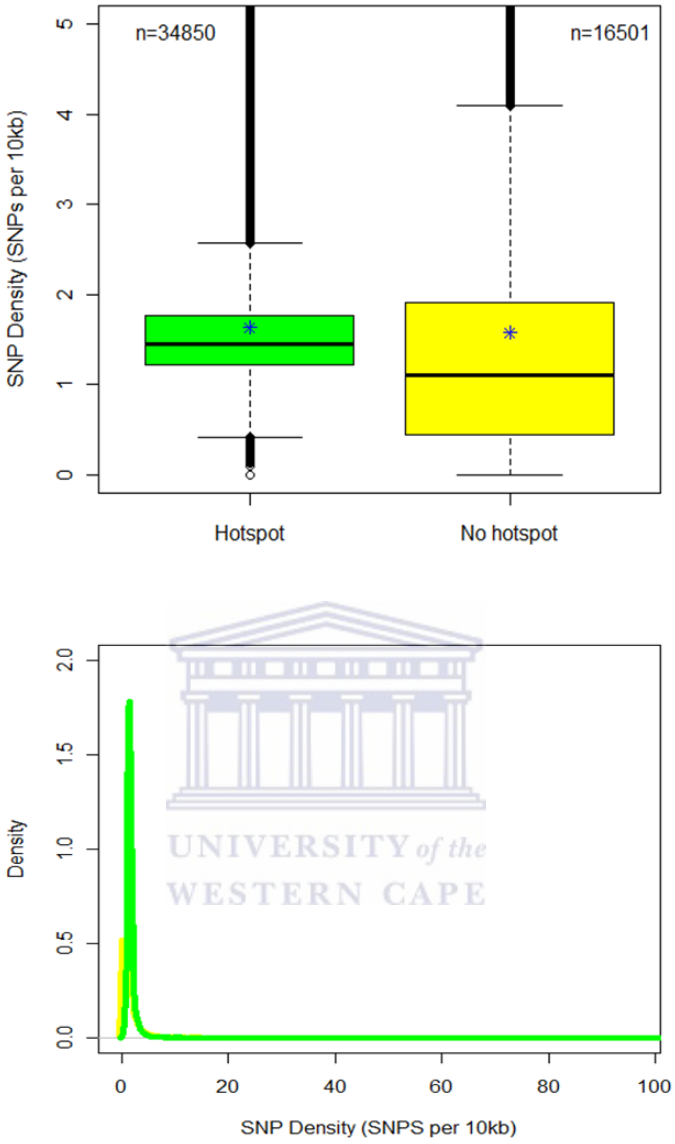


Figure 18 SNP density in genes that contain an internal hotspot compared to genes that do not contain an internal hotspot

This boxplot figure, and its associated density plot, illustrates the SNP density of genes that contain an internal hotspot compared to genes that do not contain an internal hotspot. It shows that there is very little difference between the two sets.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.2.3. Analysis of Gene Length

From the datasets of 13095 disease-associated genes and 38256 genes in the control set (a total of 51351 genes), 34850 genes contained internal hotspots while 16501 did not contain an internal hotspot. The average length of a gene with an internal hotspot is 46663bp compared to the average length of a gene with no internal hotspot, 1722bp.

The two plots, illustrated by Figure 19, show the distribution of gene lengths in genes with an internal hotspot compared to genes with no internal hotspot. A review of these plots indicates that there is indeed a difference in distribution, with genes containing a hotspot showing a greater tendency to be longer than genes with no internal hotspot. The density plot shows an estimate of the actual densities and also clearly shows that genes containing an internal hotspot are longer than genes with no internal hotspot.

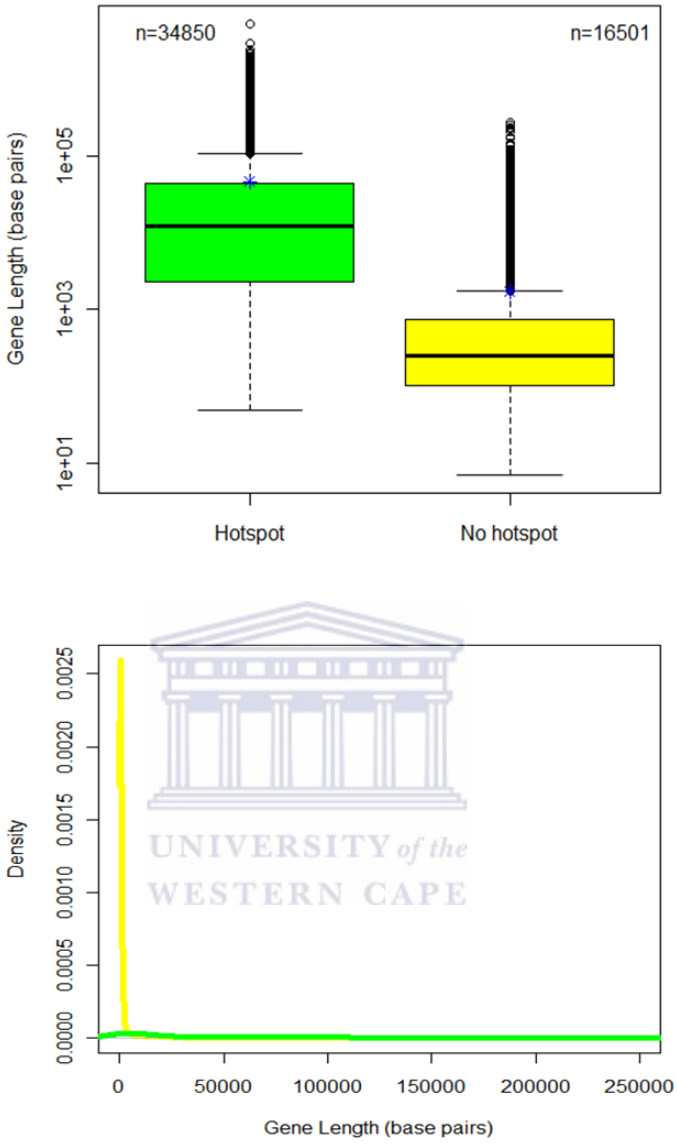


Figure 19 Length of genes that contain an internal hotspot compared to genes that do not contain an internal hotspot

This boxplot figure, and its associated density plot, illustrates the difference in length of genes that contain an internal hotspot compared to genes that do not contain an internal hotspot. It shows

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.2.4. Analysis of Position Effect: Distance from chromosome end

From the datasets of 13095 disease-associated genes and 38256 genes in the control set (a total of 51351 genes), 34850 genes contained internal hotspots while 16501 contained no internal hotspots. As mentioned in the previous section, in the dataset of 13095 disease-associated genes 4443 genes were closer to the chromosome start while 8652 genes where closer to the chromosome end. In the control set of 38256 genes, 12423 genes were closer to the chromosome start, while 25833 genes were closer to the chromosome end.

The two plots, illustrated in Figure 20, show the distribution within the chromosome length of genes with an internal hotspot and genes with no internal hotspot. A review of these plots, as well as the Figure 14 in Chapter 1, confirms that there seems to be only a slight difference in the distribution, with neither genes that contain a hotspot nor genes that do not showing any trend for occurring near the chromosome ends. The density plot also gives an indication that neither genes with a hotspot nor genes with no hotspot tend to lie close to the chromosome ends.

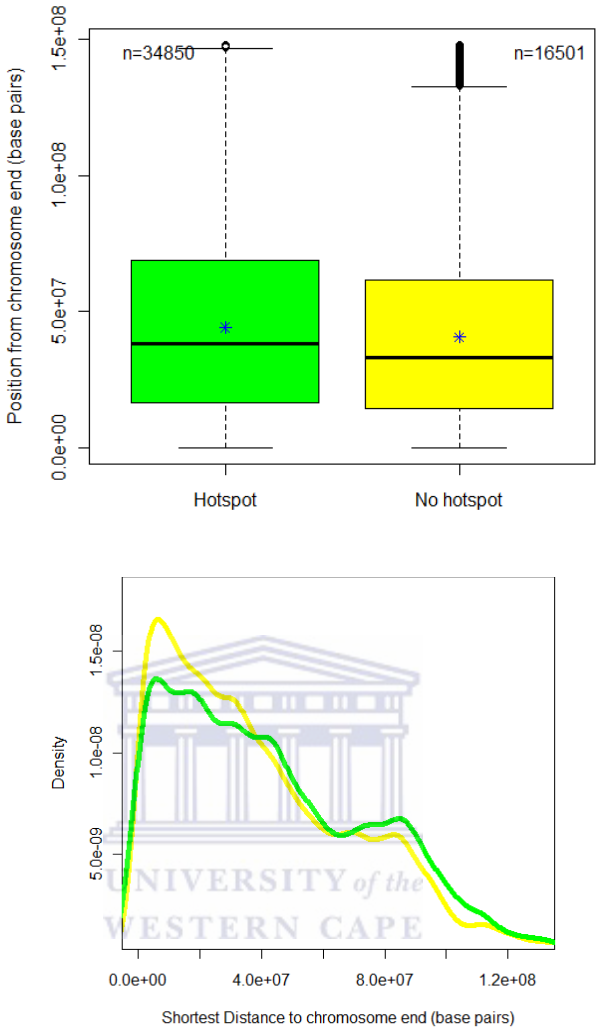


Figure 20 Position of genes on chromosome

This boxplot figure, and its associated density plot, illustrates the position from the chromosome ends of genes that contain an internal hotspot compared the position from chromosome ends of the genes with no internal hotspot. It shows that position of the gene on the chromosome has little effect on whether a gene contains an internal hotspot or not.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value while the asterisks (*) shows the mean value. The outliers (1.5 x IQR) are shown as circles.

3.3. Is there variation in the frequency of recombination in the hotspots of disease-associated genes compared to the frequency of recombination in the hotspots of the genes in the control set?

3.3.1. Analysis of the occurrence of recombination hotspots in disease-associated genes compared to recombination hotspots of the genes in the control set

There were 13095 disease-associated genes of which 12488 (95.37%) contained hotspots and there were 38256 genes in the control set of which 22362 (58.45%) contained hotspots.

These results, shown in Figure 21, indicate that not only are disease-associated genes longer, as seen in the previous two sections, but that disease-associated genes are also highly likely to contain a hotspot.

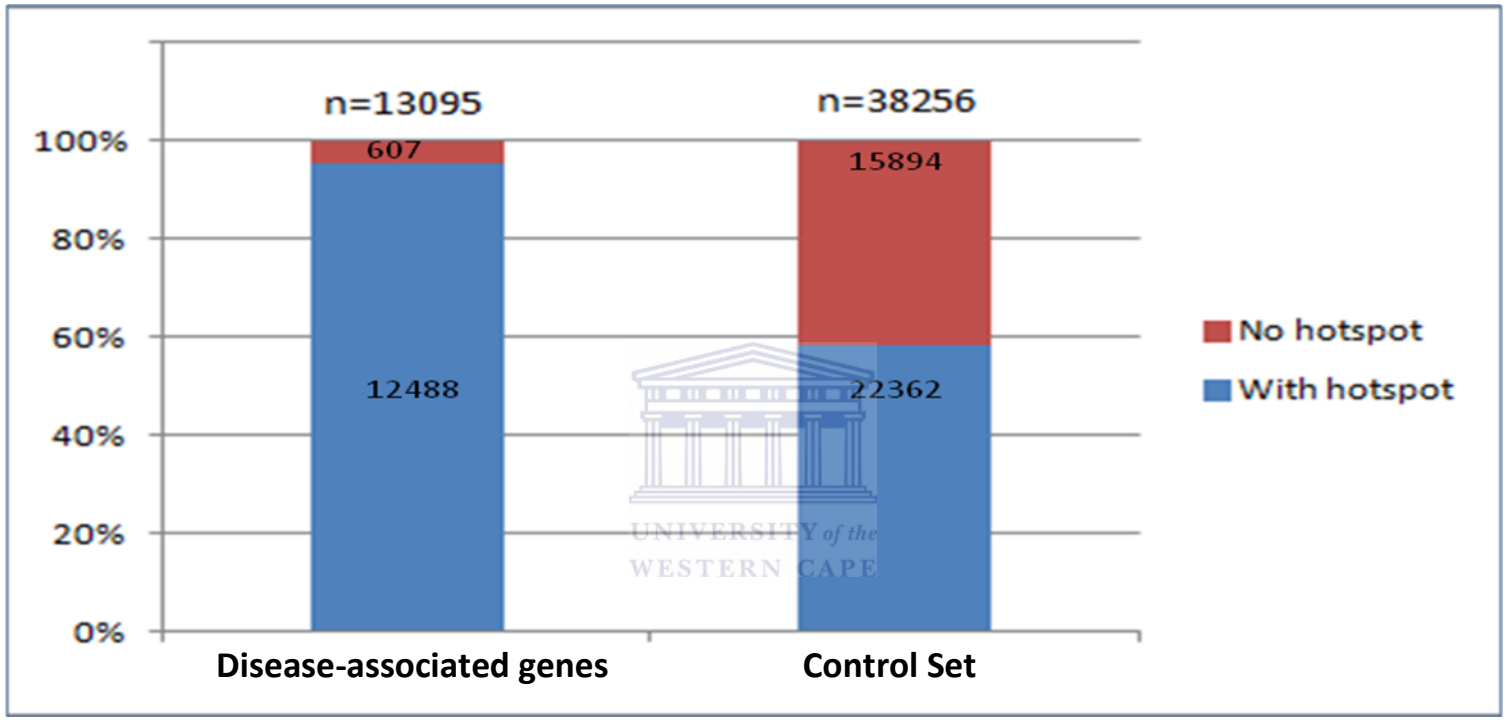


Figure 21 Comparison of hotspot position in disease-associated genes versus hotspot position of the genes in the control set

A bar chart representing the distribution of recombination hotspots within disease-associated genes and the genes in the control set. It shows that disease-associated genes are longer than the genes in the control set and also contain more internal hotspots.

3.3.2. Analysis of the frequency of recombination in the hotspots of disease-associated genes compared to the frequency of recombination in the hotspots of the genes in the control set

A visual inspection of the boxplot in Figure 22 reveals that the distribution of the frequency of recombination in disease-associated genes that contain an internal hotspot compared with the genes in the control set that contain an internal hotspot is slightly higher in disease-associated genes (mean value of 9.937) than genes in the control set (mean value of 5.882). The distribution for both datasets is skewed to the left with the bottom whisker much longer than the top whisker. From this boxplot we may determine that there is slightly higher frequencies of recombination in disease-associated genes that contain an internal hotspot than in genes in the control set that contain an internal hotspot.

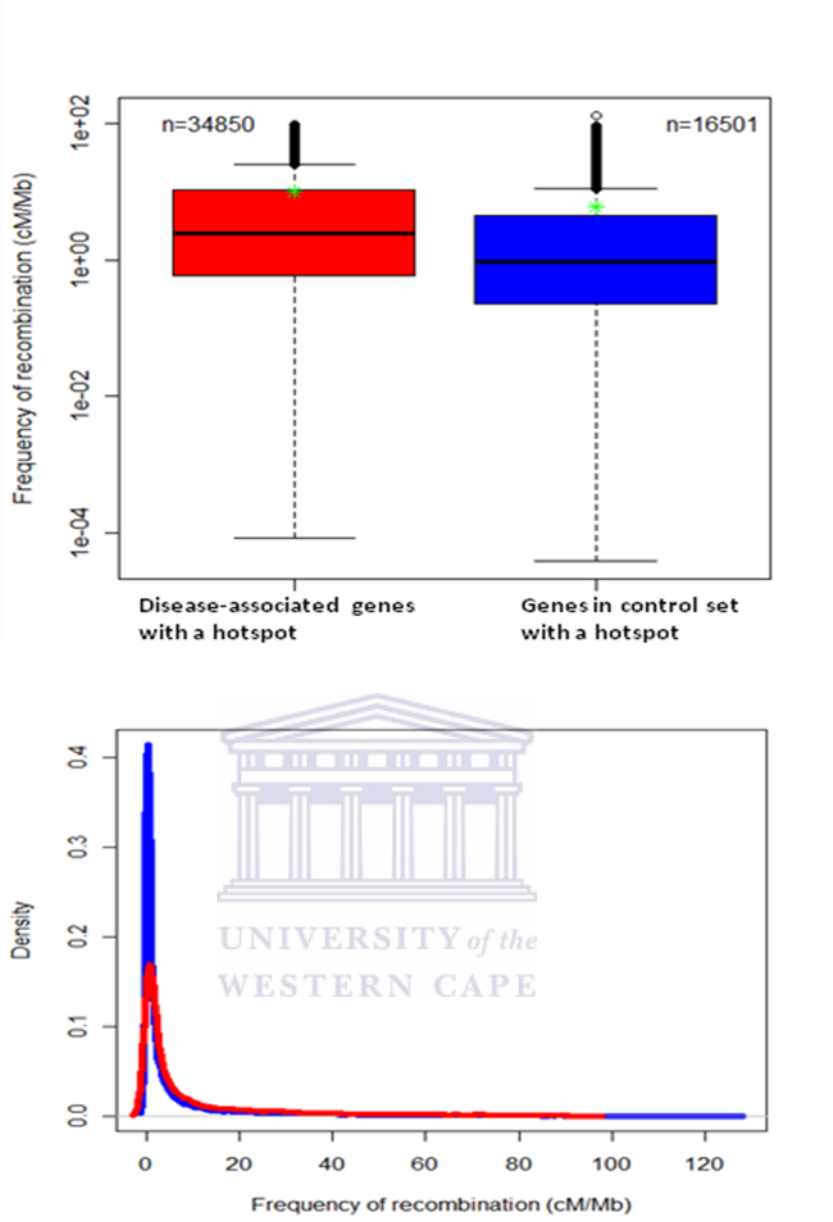


Figure 22 Frequency of recombination of internal hotspots of disease genes compared to the frequency of recombination of internal hotspots of genes in the control set

This boxplot figure, and its associated density plot, illustrates the frequency of recombination of internal hotspots of disease genes compared to the internal hotspots of genes in the control set. It shows that there is somewhat more recombination in the hotspots of disease-associated genes.

3.3.3. Analysis of the frequency of recombination of the highest scoring hotspot for each disease-associated gene compared to the frequency of recombination of the highest scoring hotspot of each gene in the control set.

A visual assessment of the boxplot in Figure 23 reveals that the distribution of the frequency of recombination of the highest scoring hotspot for each disease-associated gene compared to the frequency of recombination of the highest scoring hotspot for each gene in the control set is, again, fairly similar with a mean value of 20.021 and 18.009 respectively. The distribution for both datasets is skewed to the left with the bottom whisker much longer than the top whisker. This difference is more prevalent in the frequency of recombination in the highest scoring hotspot that lies closest to genes in the control set. From this boxplot we may determine that there is an inconsequential difference between the frequencies of recombination in the highest scoring hotspot closest to genes in the control set than in the hotspots closest to disease-associated genes. There is however, more variance in the frequency of recombination in the highest scoring hotspot closest to genes in the control set but this may only be due to the difference in population size.

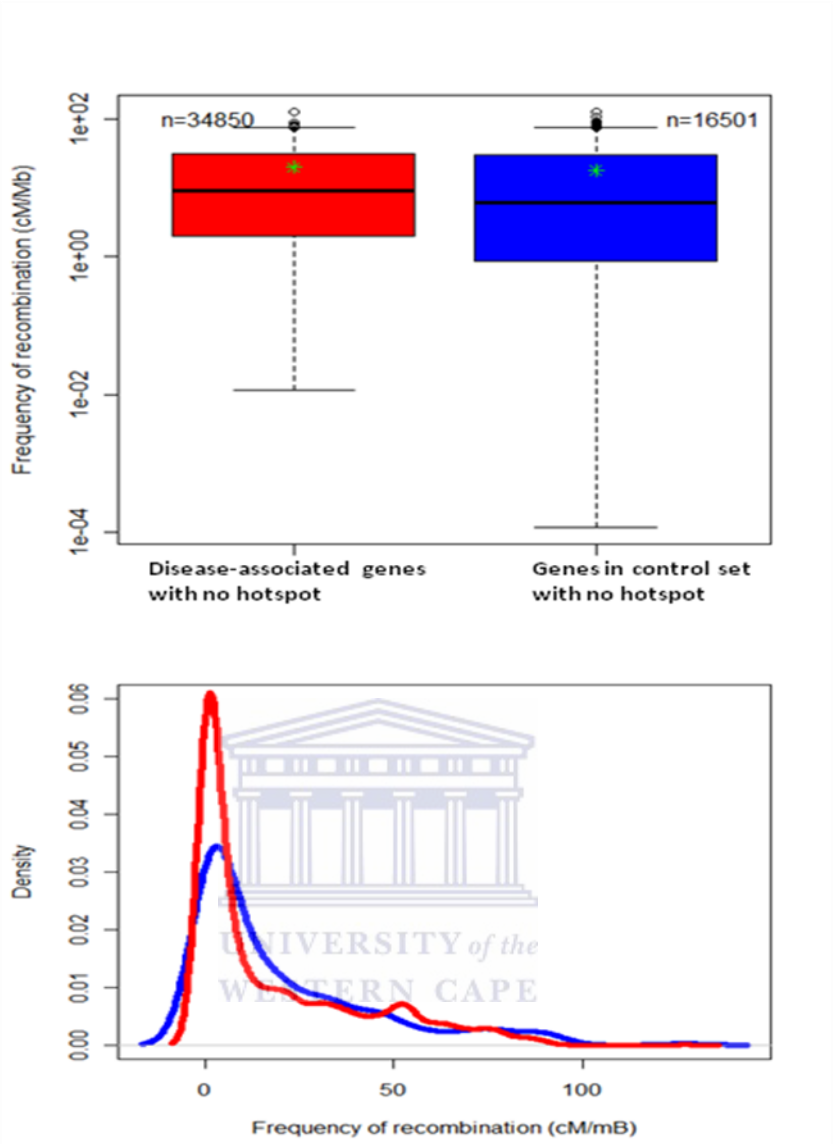


Figure 23 Frequency of recombination of the highest scoring hotspot for each disease-associated gene compared to the highest scoring hotspot for each gene in the control set

This boxplot figure, and its associated density plot, illustrates the frequency of recombination of the highest scoring hotspot for each disease-associated gene compared to the frequency of recombination for the highest scoring hotspot for each gene in the control set. It shows that there is a slight increase in recombination of the highest scoring hotspot for each disease-associated gene.

3.3.4. Analysis of the frequency, distance and overall scoring metric of hotspots nearest to disease-associated genes compared to the frequency, distance and overall scoring metric of hotspots nearest to the genes in the control set.

It is important fact to mention here is that the scoring system was developed because I had not anticipated such a high percentage of genes would contain a hotspot. The scoring system, however, was only used to analyze a small subset of disease-associated genes and genes in the control set that did not contain an internal hotspot but rather lie near a hotspot. This was because the score metric was not relevant to genes that did contain an internal hotspot because "distance from hotspot" was not a factor.

Once each gene in the human genome had a score assigned to it, the overall difference in scoring of genes in the disease-associated gene was compared to the overall scoring of genes in the control set.

A visual assessment of the boxplot in Figure 24 reveals that the distribution of the score metric of those hotspots that lie closest to disease-associated genes with no internal hotspot compared to the score metric of those hotspots that lie closest to genes in the control set that have no internal hotspot is, again, fairly similar with a mean of 0.0185 and 0.0124 respectively. The distribution for both datasets is skewed to the right with a number of visible outliers. From this boxplot we may determine that there is only a slight difference in the dataset, with the maximum scores of the hotspots that lie closest to genes in the control set slightly higher than the maximum scores of the hotspots closest to disease-associated genes.

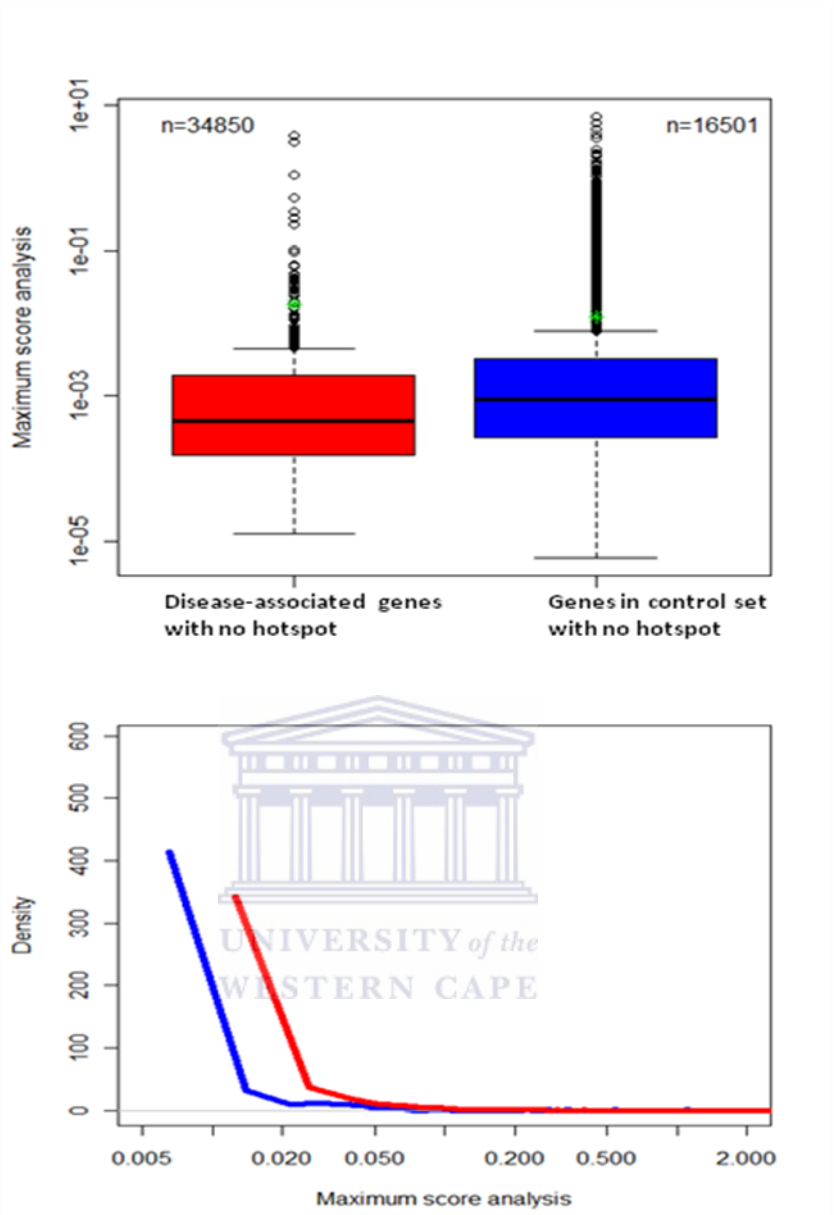


Figure 24 Score Metric of disease-associated genes compared to the score metric of genes in the control set

This boxplot figure, and its associated density plot, illustrates the score metric of disease genes compared to the score metric of the genes in the control set. It shows there is little difference between the disease-associated genes when compared to the genes in the control set when one looks at the subset of genes that do not contain an internal hotspot but rather lie near to a hotspot.

CHAPTER FOUR

Discussion and Conclusion

In this thesis, I have reviewed a set of features that represent general properties of human genes. The control set of genes, assembled using the ~38 000 known genes, pseudogenes and RNA genes from Ensembl, and which have no known association with any human disease, were evaluated against the ~13 000 disease-associated genes listed in Online Mendelian Inheritance in Man (OMIM).

Other researchers have examined these same features independently in various studies but, to my knowledge, there have been no comprehensive studies investigating all the features examined in this study. The data presented in this thesis collates analysis of all known disease-associated genes as well as all known human genes that have not thus far been identified as being involved in disease.

Table 4 gives a comparison of mean values and standard deviations for the various features of 13095 disease-associated genes and 38256 genes in the control set. The data clearly shows that disease-associated genes are longer than genes in the control set, while base composition (GC content), position on the chromosome and genetic variation (SNPs) do not differ significantly between disease-associated and genes in the control set.

The finding that disease-associated genes are longer than genes in the control set is not novel and moreover it is consistent with what has previously been described in literature.

Table 4 Tabulated displays of the mean and measure of standard deviation (S.D) for the reviewed features of the 13095 disease genes and 38256 non-disease genes.

	Disease-associated genes		Control set		Unit
	Mean	**S.D**	**Mean**	**S.D**	
GC content	46.83	8.02	45.60	7.91	%
SNP density	1.60	1.40	1.62	1.97	SNPs per 10kb
Gene length	75518	1522173	17402	5574779	bp
Position on chromosome					
Nearest hotspot	34745196	26813339	37630704	27392564	bp
Number of hotspots					
Internal hotspots	95.36		58.45		%
Nearest hotspot	4.64		41.55		%
Frequency of recombination					
All hotspots	10.40	17.20	10.88	18.38	cM/Mb
Internal hotspots	9.94	16.60	5.88	12.50	cM/Mb
Nearest hotspot	20.02	24.20	18.01	22.55	cM/Mb
Score					
Nearest hotspot	0.0185	0.21	0.0124	0.12	

It can be reasoned that this is due to the fact that the longer the gene is, the more likely it is to undergo a disease-causing mutation based purely on the increased length of gene available to undergo random mutations. In this study I was able to confirm these results with an average disease-associated gene length of 75kb compared to the average gene length of 17kb in the control set of genes.

The plot in Figure 15 shows, that even at the long end of the scale, disease-associated genes are generally longer. The plot in Figure 19 clearly shows that disease-associated genes are more likely to contain an internal hotspot than genes in the control set.

It is important to note, however, that even though the gene length results support previous findings; it might be a result of the fact that the control gene list used in this study has not been curated and contains a substantial amount of smaller genes compared to the well-curated disease-associated gene list from OMIM. Given that disease-associated genes are longer than genes in the control set it could be purely by chance alone that disease-associated gene are more likely to contain a hotspot, in other words the increase in hotspot frequency may be a result of longer gene length.

Question: if a disease-associated gene contains a hotspot is this purely due to the fact that it is longer, in other words the hotspot does not contribute to disease gene status, or is it because the hotspot is contributing to the likelihood of being a disease-associated gene?

If one compares the length of genes containing a hotspot compared to the length of genes not containing a hotspot there is definitely a negative correlation (Figure 25).

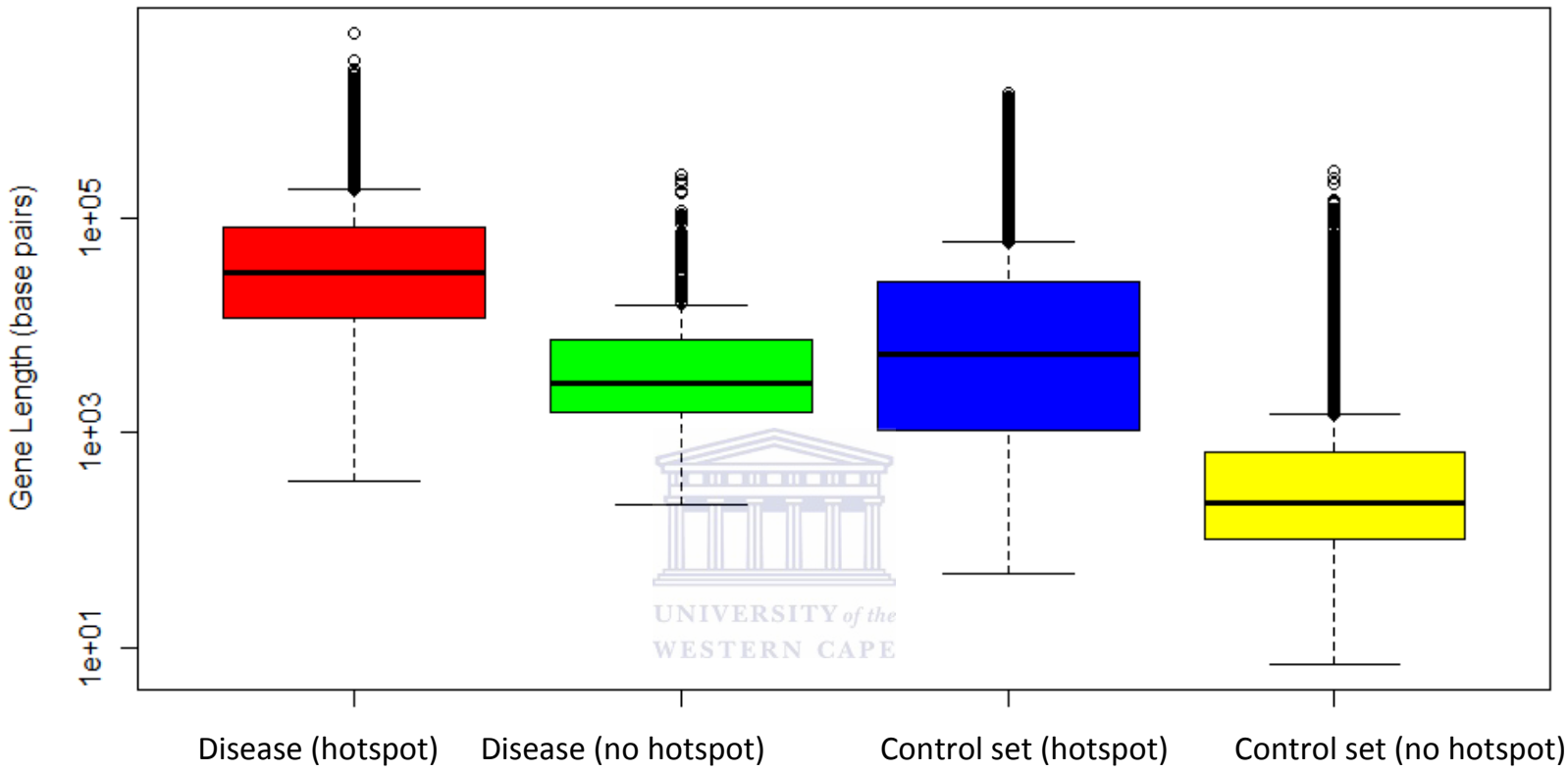


Figure 25 Relationship between gene length and the presence or absence of recombination hotspots

This figure illustrates the relationship between disease-associated genes and the genes in the control set and the likelihood that they will contain an internal hotspot. It shows that the presence of a hotspot is not sufficient on its own to cause disease gene status.

The box represents the data distribution between the 25th percentile and the 75th percentile, or the interquartile range (IQR). The horizontal line in the box shows the median value. The outliers (1.5 x IQR) are shown as circles.

Disease genes that contain a hotspot are longer than genes in the control set that contain a hotspot but disease-associated genes that do not contain a hotspot are also longer than genes in the control set that do not contain a hotspot and this result gives some indication that the presence of a hotspot is not sufficient on its own to cause disease gene status.

One could speculate that the number of extreme outliers could be a consequence of the varying distance that a recombination hotspot lies from a gene. The distance that the hotspot lies from the gene and the size of the gene could both play a role in the fact that there are so many extreme outliers, especially in the genes in the control set where the genes are smaller in size.

When one assesses the middle of the range dataset, in other words, only the disease-associated gene and non-disease gene lengths between 2000bp and 1500000bp, and excluding the smallest genes, this observation changes, as shown in Figure 26. The percentage of disease-associated genes with a hotspot increases to 97.02% while the percentage of genes in the control set with a hotspot increases to 91.12%. This demonstrates that the presence of a hotspot in a gene is a consequence of gene length rather than a feature that establishes gene status.

This result could be due to many factors, the most obvious being that the large numbers of very short genes, which are possibly genes which have not as yet been reported as having a function in disease, could be a consequence of poor annotations or pseudogenes; however future research would need to be done to elucidate this possibility.

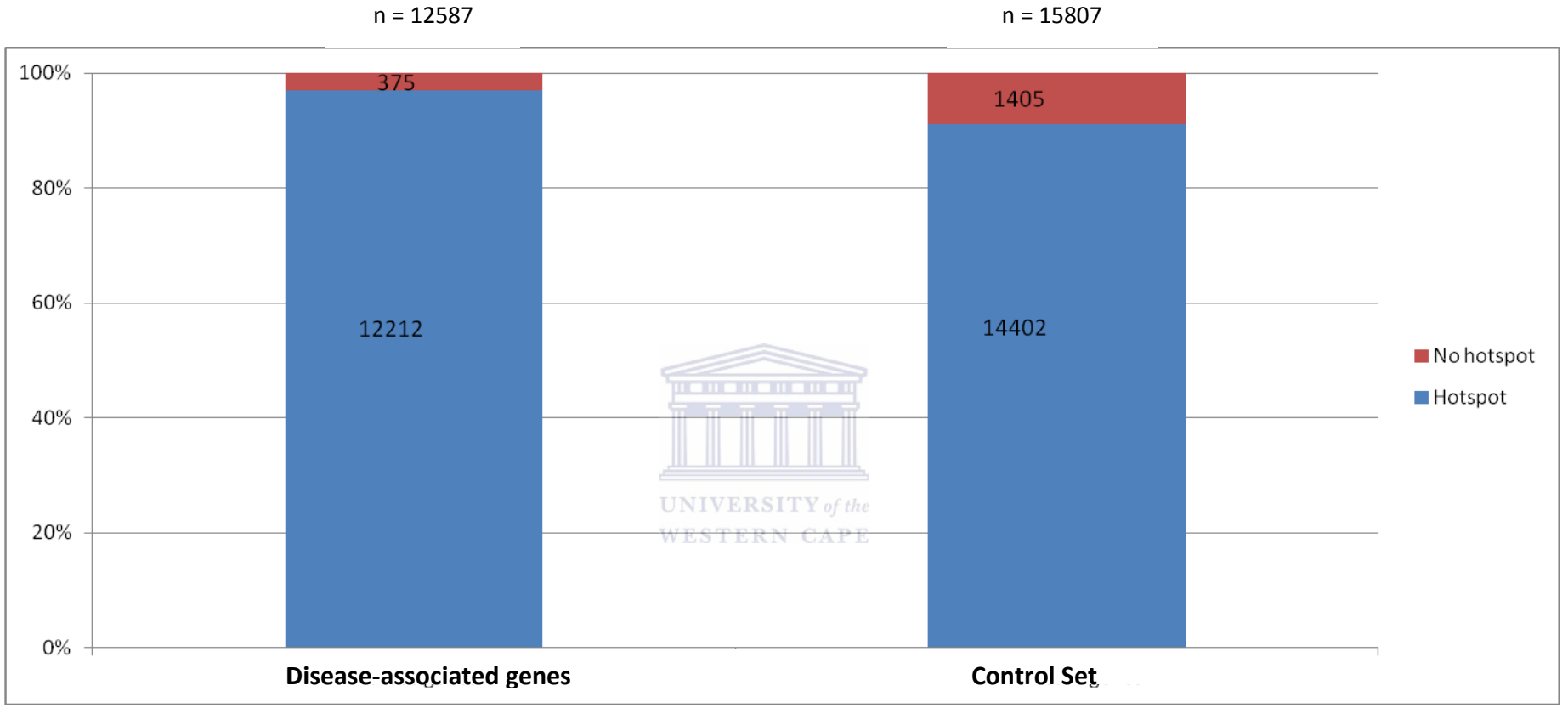


Figure 26 Comparison of the median gene lengths of disease-associated genes and genes in the control set

A graphical representation of the distribution of recombination hotspots within disease-associated genes and genes in the control set using a subset of the total genes, i.e. the middle of the range dataset excluding the smallest genes. This demonstrates that the presence of a hotspot in a gene is a consequence of gene length rather than a feature that establishes gene status.

The observation that a higher proportion of disease-associated genes contain hotspots when compared to the proportion of genes in the control set that contain hotspots is a novel finding, while the other results are not entirely consistent with the findings that have been previously published Reference. The reason for these discrepancies may be because no study has systematically evaluated gene length, base composition (GC content), positional effect (shortest distance to chromosome end) and genetic variation (SNP density) of all known genes from Ensembl (coding and non-coding genes, psuedogenes, RNA gene etc.) as well as all the known disease-associated genes in OMIM in a single study.

This study also makes special reference to the relationship between recombination hotspots and their presumed function in disease-associated genes. Most other studies make use of a randomly selected subset of genes from Ensembl, as well as other data sources, as a representative model, with statistical analysis to extrapolate their findings.

It may be interesting to note that the GC content in genes is different compared to the GC content of the whole genome. In genes, the average GC content is between 45% and 50%, and is evenly distributed. In the genome, however the GC content is ~41% GC with an uneven distribution skewed to the left. As a consequence, regions of high GC content (62-68%) have higher relative gene density than regions of lower GC content (Lander *et al*., 2001, International Human Genome Sequencing Consortium, 2001).

If one refers to the data in Table 4, the results show that the study found no substantial differences between the percentages GC content in disease-associated genes when compared to genes in the control set. These findings may be a result of the larger control set and consequently the emergence of more outliers or simply the fact that there are many more genes in the dataset. This study also concludes that there is no substantial difference

between the percentage GC content in genes that contain an internal hotspot and genes that do not contain an internal hotspot suggesting that GC content does not influence where recombination hotspots occur. This is in contradiction to what has been published in previous literature as, to my knowledge, no previous literature has been published specifically on the GC content of coding regions of disease-associated genes and genes in the control set. However, there have been publications that support a causal relationship between GC-content and recombination rate in humans as well as identifying that recombination hotspots are associated with local increases in GC content (Fullerton *et al*., 2001, Freudenberg *et al*., 2009). It has been proposed that this effect might result from biased gene conversion (BGC). In this process gene conversion, in other words, the copy/pasting of one allele onto the other heterozygous loci during meiotic recombination, is biased towards GC-alleles, and leads to an increased probability of GC-rich regions compared to AT-rich regions (Duret *et al*., 2008). This BGC should render enrichment of GC-content in regions of high recombination compared to regions of low recombination.

The findings about SNP density in disease-associated genes compared to genes in the control set was not expected. Previous publications have confirmed the average SNP density as 8.33 SNPs per 10kb in the human genome and this study showed the average SNP density as 1.61 SNPs per 10kb within genes. The difference in these findings are due to the fact that the whole genome will have a higher SNP density because intergenic regions, in other words "non-gene DNA", can tolerate changes much more than coding or gene regions and therefore there will be many more SNPs outside of gene regions. Also, since it is known that disease-associated genes are generally longer than genes in the control set logically they should have more SNPs. These results show that whilst total SNP count per gene is higher in disease-associated genes according to their length, there is no

large variation in SNP density in disease-associated genes compared to genes in the control set. While attempting to identify genome-wide genes likely to be involved in human genetic diseases Lopez-Bigas *et al* concluded that genes involved in disease tend to be situated in conserved regions of the genome exposed to strong evolutionary constraints and that these genes therefore had not had the opportunity to accumulate many variations (Lopez-Bigas *et al*., 2004). This finding is supported by earlier studies that have hypothesised that the most severe "disease" genes are human essential genes, or housekeeping genes, and any disruption to their function will cause fatal consequences. It is thus proposed that these genes are less likely to tolerate sequence changes and therefore will contain fewer polymorphisms (Tu *et al*., 2006). These findings are however, not supported in this study and this may be because of the SNP data used. At any point in time, SNP records are certainly unlikely to catalogue all known SNPs. The current data for human variome is continuously and rapidly expanding as many more genomes from diverse ethnic origins, especially in Africa, are sequenced. For this study, SNP data from Ensembl SNPMart 65 was used. While this database is comparatively current, SNP density will continue to be limited by our limited knowledge for some time yet and a real conclusion about SNP density in disease-associated genes and the control set cannot be conclusively drawn.

Disease-associated genes appear randomly distributed across the genome, which corresponds with the random distribution of all genes across the genome. Similarly, when examining the distribution of recombination hotspots and their frequencies, a random distribution in seen across the genome. This relationship is illustrated in Figure 27. It shows where the genes lie on the chromosome as well as the distribution of hotspots and their frequency of recombination and that these are random distributions and not skewed

towards the centers or ends of the chromosome. The analysis of the distance of a gene to the chromosome end corroborates this finding that recombination is not facilitated more easily near the chromosome ends as would be suggested by the hypothesis that the integrity of a chromosomes tertiary structure is more flexible towards chromosome ends. This may be due to the telomere, a region of repetitive nucleotide sequences at the end of a chromosome. The telomere protects the end of the chromosome from deterioration or from fusion with neighboring chromosomes by maintaining a well-packed tertiary structure. Telomere regions deter the degradation of genes near the ends of chromosomes by allowing chromosome ends to shorten, which necessarily occurs during chromosome replication (Nachman, 2002, Maddar *et al*., 2001).

On initial observation of Figure 27, it appears that on some chromosomes, gene dense regions seem to undergo fewer recombination events, for example on chromosome 3 at p21.31 and chromosome 16 at q22.1 and q22.2, and this could warrant further investigation in future studies.

To summarize, this study has determined that even though disease-associated genes are longer than genes in the control set and more likely to contain an internal recombination hotspot this does not indicate that these are traits that *cause* disease gene status. However, the fact that many genes in the control set do in fact contain hotspots, does not rule out the possibility that these genes may be found to underlie disease in the future. In general, it is difficult to make definitive conclusions about the control set if genes not implicated in disease as genes can only be defined as such within the limits of our current knowledge. This issue is discussed in Tiffin *et al*., 2009 (review article). In this study I have observed such differences between disease-associated genes and the control set. It is not possible,

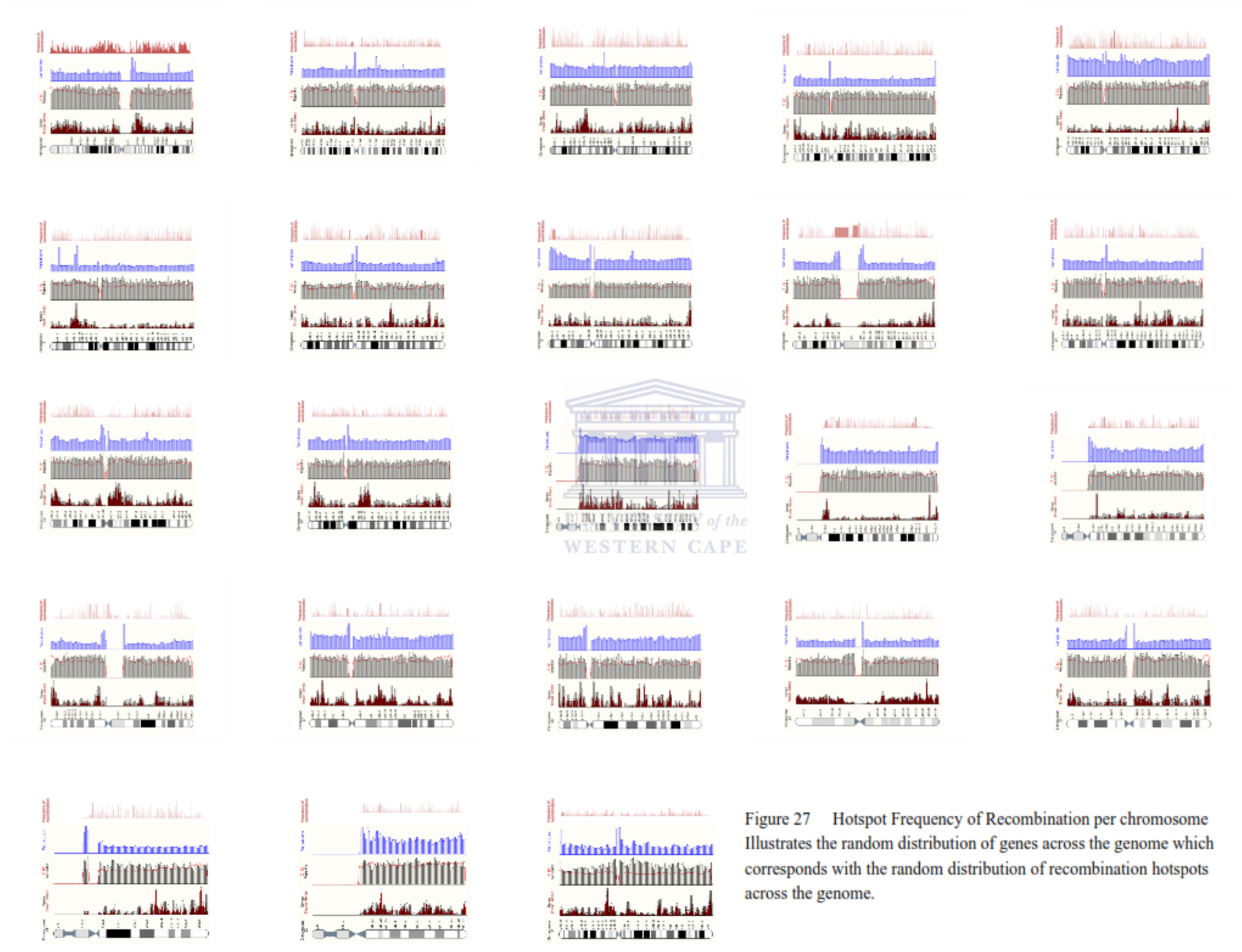


Figure 27 Hotspot Frequency of Recombination per chromosome Illustrates the random distribution of genes across the genome which corresponds with the random distribution of recombination hotspots across the genome.

however, to demonstrate a causal relationship between these factors within the current study. I propose possible relationships between disease-associated genes and the control set and their characteristics, but further work is necessary to test whether causal relationships do exist. It is possible that the increase in number of hotspots indicates more variability being introduced and therefore more likelihood of causing altered phenotype, including disease phenotype. It could also be that containing an internal recombination hotspot predisposes a gene to cause disease and that longer genes are more likely to contain a hotspot and therefore longer genes are more likely to be disease-associated genes. In order to prove either hypothesis, further investigation is required.

Co-occurring traits in disease-associated genes can contribute as supporting evidence for a hypothesis but they do not determine causality. In order to determine causality one could, for example, look at different ethnicities where the distribution of hotspots differ to established recombinant hotspot sets which are Eurocentric and identifying if the link between disease and the occurrence of hotspots still holds true or whether it is in fact only gene length that is the defining factor.

In conclusion, I propose that genes that are longer are more likely to be disease-associated genes because they have more SNPs that might have disease-associated alleles and because they are more likely to contain a recombination hotspot which leads to greater variation and consequently more chance of a disease phenotype arising.

One should keep in mind that Next Generation Sequencing is aiding in the discovery of disease genes by sequencing the entire protein-coding sequence, or exome and this will make studies, such as the one done here, even more valuable in the future.

Future Direction

This approach gives us the potential to identify "most likely" disease-associated genes in a population-specific manner, and to look at the effect of distance/frequency of recombination hotspots on aetiological genes for diseases that have varying prevalence between populations. Given that, to date, so much disease-associated gene research is conducted within the Northern Hemisphere in a predominantly Caucasian environment, and then translated to the African populations, this would have great implications for prediction and analysis of candidate disease-associated genes exclusively for indigenous Africans. If we can link population structure to prediction of disease-associated genes, we pave the way to disease-associated gene prediction specifically adapted for indigenous Africans, with a methodology that is equally applicable to other populations.

African natives are the most ethnically diverse population in the world (Tishkoff *et al*., 2002). This is because the African continent is regarded as the "*cradle of mankind*" where the first human remains were discovered from 200 000 years ago. For this reason initiatives like HapMap (URL www.hapmap.org/) and 1000 Genome Project (URL http://www.1000genomes.org/) have focused on African-based studies for disease-related genetic and genomic research. To date, the complete genomes of two individuals as well as the exome sequence of three individuals from indigenous populations in Southern Africa have been sequenced (Schuster *et al*., 2010). The 1000 Genome project has samples from ~500 DNA samples from West African Ancestry while HapMap is working with ~480 from East and West Africa. The problem is that even though there is a significant amount of data becoming available about African genomes, there is not a lot of information about genes that cause disease in Africans as well as disease prevalence in African populations.

Recent genetic studies on African Americans have uncovered similar methods to creating detailed genome maps of African American DNA and this new technique will enable Scientists to locate the genes that cause disease (Hinch *et al*., 2011). As information becomes known about the genetics of African diseases, it may become easier to test whether the increased frequency of recombination hotspots in Africans results in different disease-associated genes in these populations. This will, however, also require the assembly of reliable test sets of African-specific disease-associated genes and the associated control sets of non disease-associated genes.

Additional studies surrounding known translocation breakpoints that have been identified in cancers could also prove an interesting avenue for further computational analysis of sequences. The methods used in a recent study on translocations in lymphomas and leukemias in humans (Hakim *et al*., 2012) may potentially assist with identifying other motifs or sequence characteristics responsible for the predisposition to translocations in other forms of cancer.

References

1. A haplotype map of the human genome. (2005).Nature, 437(7063), 1299-320.
2. A map of human genome variation from population-scale sequencing. (2010).Nature, 467(7319), 1061-73.
3. Abeysinghe, S. S., Chuzhanova, N., & Cooper, D. N. (2006). Gross deletions and translocations in human genetic disease. Genome dynamics, 1, 17-34.
4. Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., & Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. BMC bioinformatics, 6, 55.
5. Amberger, J., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic acids research, 37(Database issue), D793-6.
6. Ames, D., Murphy, N., Helentjaris, T., Sun, N., & Chandler, V. (2008). Comparative analyses of human single- and multilocus tandem repeats. Genetics, 179(3), 1693-704.
7. Auton, A. (2007). Thesis: The Estimation of Recombination Rates from Population Genetic Data.
8. Baird, D. M. (2008). Mechanisms of telomeric instability. Cytogenetic and genome research, 122(3-4), 308-14.
9. Balakirev, E. S., & Ayala, F. J. (2003). Pseudogenes: are they "junk" or functional DNA? Annual review of genetics, 37, 123-51.

10. Barnes, M. R. (2006). Navigating the HapMap. Briefings in bioinformatics, 7(3), 211-24.

11. Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics (Oxford, England), 21(2), 263-5.

12. Blixt, S. (1975). Why didn't Gregor Mendel find linkage? Nature, 256(5514), 206-206.

13. Brown, T. (2002). Mutation, Repair and Recombination. Wiley-Liss. (URL: http://www.ncbi.nlm.nih.gov/books/NBK21114/#A8454)

14. Capper, R., Britt-Compton, B., Tankimanova, M., Rowson, J., Letsolo, B., Man, S., Haughton, M., et al. (2007). The nature of telomere fusion and a definition of the critical telomere length in human cells. Genes & development, 21(19), 2495-508.

15. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lanc, C. R., et al. (1999). (.pdf) Nature Genetics, 22(July), 231-238.

16. Carrington, M., & Cullen, M. (2004). Justified chauvinism: advances in defining meiotic recombination through sperm typing. Trends in genetics : TIG, 20(4), 196-205.

17. Chang, H., Chuang, W.-Y., Sun, C.-F., & Barnard, M. R. (2012). Concurrent acute myeloid leukemia and T lymphoblastic lymphoma in a patient with rearranged PDGFRB genes. Diagnostic pathology, 7, 19.

18. Chemistry of the cell and genetics. (Retrieved from http://www.ucl.ac.uk/~ucbhjow/bmsi/bmsi-lectures.html)

19. Chen, J.-M., Cooper, D. N., Férec, C., Kehrer-Sawatzki, H., & Patrinos, G. P. (2010). Genomic rearrangements in inherited disease and cancer. Seminars in cancer biology, 20(4), 222-33.

20. Chen, J.-M., Férec, C., & Cooper, D. N. (2006). A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview. Human genetics, 120(1), 1-21.

21. Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature genetics, 38(11), 1251-60.

22. Coop, G., & Przeworski, M. (2007). An evolutionary view of human recombination. Genetics, 8(January).

23. Clark, AG, Wang, X, Matise, T., 2010. Contrasting Methods of Quantifying Fine Structure of Human Recombination. Annual Review of Genomics and human genetics. September 22; 11: 45–64.

24. Duret, L., & Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS genetics, 4(5), e1000071.

25. Farazi, T. A., Spitzer, J. I., Morozov, P., & Tuschl, T. (2011). miRNAs in human cancer. The Journal of pathology, 223(2), 102-15.

26. Fearnhead, P. (2006). SequenceLDhot: detecting recombination hotspots. Bioinformatics, 22(24), 3061-3066.

27. Fisher, C. (2010). Soft tissue sarcomas with non-EWS translocations: molecular genetic features and pathologic and clinical correlations. Virchows Archiv : an international journal of pathology, 456(2), 153-66.

28. Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., et al. (2011). Ensembl 2011. Nucleic acids research, 39(Database issue), D800-6.

29. Folkersen, L., van't Hooft, F., Chernogubova, E., Agardh, H. E., Hansson, G. K., Hedin, U., Liska, J., et al. (2010). Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. Circulation. Cardiovascular genetics, 3(4), 365-73.

30. Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164), 851-61.

31. Freudenberg, J., Wang, M., Yang, Y., & Li, W. (2009). Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome. BMC bioinformatics, 10 Suppl 1, S66.

32. Fullerton S, Carvalho AB, C. A. (2001). Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol, 18, 1139-42.

33. Fullerton, S. M., Carvalho, A. B., & Clark, A. G. (2000). Letter to the Editor Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. Molecular Biology, 1139-1142.

34. Goldstein, D. B., & Weale, M. E. (2001). Population genomics: linkage disequilibrium holds the key. Current biology : CB, 11(14), R576-9.

35. Greenwood, T. a, Rana, B. K., & Schork, N. J. (2004). Human haplotype block sizes are negatively correlated with recombination rates. Genome research, 14(7), 1358-61.

36. Haber, J. E., Ira, G., Malkova, A., & Sugawara, N. (2004). Repairing a double-strand chromosome break by homologous recombination: revisiting Robin

Holliday's model. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 359(1441), 79-86.

37. Hakim, O., Resch, W., Yamane, A., Klein, I., Kieffer-Kwon, K.-R., Jankovic, M., Oliveira, T., et al. (2012). DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature.

38. Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N., & Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic acids research, 33(8), 2374-83.

39. He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., et al. (2005). A microRNA polycistron as a potential human oncogene. Nature, 435(7043), 828-33.

40. Hey, J. (2004). What's so hot about recombination hotspots? PLoS biology, 2(6), e190.

41. Hinch, A. G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C. D., Chen, G. K., et al. (2011). The landscape of recombination in African Americans. Nature, 476(7359), 170-5.

42. Holliday, R. (1964). A mechanism for gene conversion in fungi. Genetical Research, 89(5-6), 285-307.

43. Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. Genetical research, 50(3), 245-50.

44. Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature, 451(7181), 998-1003.

45. Jeffreys, A. J., & Neumann, R. (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Human molecular genetics, 14(15), 2277-87.

46. Johnson, R. E., Washington, M. T., Prakash, S., & Prakash, L. (2000). Fidelity of human DNA polymerase eta. The Journal of biological chemistry, 275(11), 7447-50.

47. Kann, M. G. (2010). Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefings in bioinformatics, 11(1), 96-110.

48. Kauppi, L., Jeffreys, A. J., & Keeney, S. (2004). Where the crossovers are: recombination distributions in mammals. Nature reviews. Genetics, 5(6), 413-24.

49. Kim, H., Gillis, L. C., Jarvis, J. D., Yang, S., Huang, K., Der, S., & Barber, D. L. (2011). Tyrosine kinase chromosomal translocations mediate distinct and overlapping gene regulation events. BMC cancer, 11, 528.

50. Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. a, Richardsson, B., Sigurdardottir, S., et al. (2002). A high-resolution recombination map of the human genome. Nature genetics, 31(3), 241-7.

51. Kristin G. Ardlie, L. K., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3, 3(4), 299-309.

52. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001a). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921.

53. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001b). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921.

54. Li, J., Zhang, M. Q., & Zhang, X. (2006). ARTICLE A New Method for Detecting Human Recombination Hotspots and Its Applications to the HapMap ENCODE Data. Journal of Human Genetics, 79(October).

55. Li, N., & Stephens, M. (2003a). Using Single-Nucleotide Polymorphism Data. Genetics, 2233(December), 2213-2233.

56. Li, N., & Stephens, M. (2003b). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics, 165(4), 2213-33.

57. Lichten, M., & Goldman, A. S. (1995). Meiotic recombination hotspots. Annual review of genetics, 29, 423-44.

58. López-Bigas, N., & Ouzounis, C. A. (2004a). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic acids research, 32(10), 3108-14.

59. López-Bigas, N., & Ouzounis, C. a. (2004b). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic acids research, 32(10), 3108-14.

60. Maddar, H., Ratzkovsky, N., & Krauskopf, A. (2001). Role for telomere cap structure in meiosis. Molecular biology of the cell, 12(10), 3191-203.

61. Maindonald, J. (2008). Using R for Data Analysis and Graphics Introduction , Code and Commentary. Australian Journal of Zoology, (January).

62. Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008). Science in medicine A HapMap harvest of insights into the genetics of common disease. Genome Research, 118(5).

63. McEente, G., Minguzzi, S., O'Brien, K., Ben Larbi, N., Loscher, C., O'Fágáin, C., & Parle-McDermott, A. (2011). The former annotated human pseudogene dihydrofolate reductase-like 1 (DHFRL1) is expressed and functional. Proceedings of the National Academy of Sciences of the United States of America, 108(37), 15157-62.

64. McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science (New York, N.Y.), 304(5670), 581-4.

65. Mcvean, G. A. T., Myers, S. R., & Hunt, S. (2011). The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Society, 581(2004).

66. Micklos, D. A., & Freyer, G. A. (2003). DNA Science: A First Course (2nd editio., p. 285).

67. Mitelman, F., Mertens, F., & Johansson, B. (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. Nature genetics, 15 Spec No, 417-74.

68. Mucha, M., Lisowska, K., Goc, a, & Filipski, J. (2000). Nuclease-hypersensitive chromatin formed by a CpG island in human DNA cloned as an artificial chromosome in yeast. The Journal of biological chemistry, 275(2), 1275-8.

69. Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science (New York, N.Y.), 310(5746), 321-4.

70. Myers, S., Bottolo, L., Freeman, C., Mcvean, G., & Donnelly, P. (2011). R EPORTS A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. October, 321(2005).

71. Myers, S., Freeman, C., Auton, A., Donnelly, P., & McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nature genetics, 40(9), 1124-9.

72. Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. Current opinion in genetics & development, 12(6), 657-63.

73. Nambiar, M., Kari, V., & Raghavan, S. C. (2008). Chromosomal translocations in cancer. Biochimica et biophysica acta, 1786(2), 139-52.

74. Need A.C. and Goldstein D.B., 2009. Next generation disparities in human genomics: concerns and remedies. Trends in Genetics, Volume 25, Issue 11, November 2009, Pages 489–494.

75. Novo, F. J., & Vizmanos, J. L. (2006). Chromosome translocations in cancer: computational evidence for the random generation of double-strand breaks. Trends in genetics : TIG, 22(4), 193-6.

76. Ohshima K, Masahira H, Yada T, Gojobori T, Sakaki Y, et al. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol 4: R74

77. Ortiz de Mendíbil, I., Vizmanos, J. L., & Novo, F. J. (2009). Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. PloS one, 4(3), e4805.

78. Osada, N., Mano, S., & Gojobori, J. (2009). Quantifying dominance and deleterious effect on human disease genes. Proceedings of the National Academy of Sciences of the United States of America, 106(3), 841-6.

79. Paigen, K., & Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. Nature reviews. Genetics, 11(3), 221-33.

80. Parton, M. J. (2003). Online Mendelian Inheritance in Man OMIM: www.ncbi.nlm.nih.gov/entrez. Journal of Neurology, Neurosurgery & Psychiatry, 74(6), 703-703.

81. Petes, T. D. (2001). Petes-RecHotSpot-NRG-2001[1].pdf. Nature Reviews Genetics, 2, 360-369.

82. Pickering, B. M., & Willis, A. E. (2005). The implications of structured 5' untranslated regions on translation and disease. Seminars in cell & developmental biology, 16(1), 39-47.

83. de Pontual, L., Yao, E., Callier, P., Faivre, L., Drouin, V., Cariou, S., Van Haeringen, A., et al. (2011). Germline deletion of the miR-17~92 cluster causes skeletal and growth defects in humans. Nature genetics, 43(10), 1026-30.

84. Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. American journal of human genetics, 69(1), 1-14.

85. Purandare, S. M., & Patel, P. I. (1997). Recombination Hot Spots and Human Disease. Genome Research, 773-786.

86. Rabbitts, T. H. (1994). Chromosomal translocations in human cancer. Nature, 372(6502), 143-9.

87. Reich, D. (2009). Using genetics to study human history and natural selection. Harvard Medical School Department of Genetics Broad Institute.

88. Reich, D. E., Gabriel, S. B., & Altshuler, D. (2003). Quality and completeness of SNP databases. Nature genetics, 33(4), 457-8.

89. Schneider, J. a, Peto, T. E. a, Boone, R. a, Boyce, A. J., & Clegg, J. B. (2002). Direct measurement of the male recombination fraction in the human beta-globin hot spot. Human molecular genetics, 11(3), 207-15.

90. Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., et al. (2010). Complete Khoisan and Bantu genomes from southern Africa. Nature, 463(7283), 943-7.

91. Shaw, Z 2012. Learn Python The Hard Way Second Edition

92. Slatkin, M. 2008. Linkage disequilibrium – understanding the evolutionary past and mapping medical future. Nature Reviews Genetics 9, 477-485.

93. Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart--biological queries made easy. BMC genomics, 10, 22.

94. Smith, N. G. C., & Eyre-Walker, A. (2003). Human disease genes: patterns and predictions. Gene, 318, 169-175.

95. Spencer, C. C. a, Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., et al. (2006). The influence of recombination on human genetic diversity. PLoS genetics, 2(9), e148.

96. Strachan, T., & Read, A. (1999). Identifying human disease genes - Human Molecular Genetics - NCBI Bookshelf. Human Molecular Genetics. 2nd edition.

97. Stringer CB, A. P. (1988). Genetic and fossil evidence for the origin of modern humans. Science, (239), 1263-1268.

98. Stumpf, M. P. H., & Goldstein, D. B. (2003). Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Current biology : CB, 13(1), 1-8.

99. Svensson, O., Arvestad, L., & Lagergren, J. (2006). Genome-wide survey for biologically functional pseudogenes. PLoS computational biology, 2(5), e46.

100. Teo YY, Small KS, Kwiatkowski DP Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 11(2): 149-160. The International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437(7063): 1299-1320.

101. The International HapMap Project. (2003).Nature, 426(6968), 789-96.

102. The International HapMap Consortium. (2007). NIH Public Access. October, 437(7063), 1299-1320.

103. The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature, 467(7319), pp.1061-73.

104. Theuns, J., & van Broeckhoven, C. (2000). Hum. Mol. Genet.pdf. Human Molecular Genetics, 9(16), 2383-2394.

105. Thum, T., Galuppo, P., Wolf, C., Fiedler, J., Kneitz, S., van Laake, L. W., Doevendans, P. A., et al. (2007). MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure. Circulation, 116(3), 258-67.

106. Tiffin, N., Adie, E., Turner, F., Brunner, H. G., van Driel, M. A., Oti, M., Lopez-Bigas, N., et al. (2006). Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic acids research, 34(10), 3067-81.

107. Tiffin, N., Andrade-Navarro, M. a, & Perez-Iratxeta, C. (2009). Linking genes to diseases: it's all in the data. Genome medicine, 1(8), 77.

108. Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., et al. (2009). The genetic structure and history of Africans and African Americans. Science (New York, N.Y.), 324(5930), 1035-44.

109. Tsai, A. G., & Lieber, M. R. (2010). Mechanisms of chromosomal rearrangement in the human genome. BMC genomics, 11 Suppl 1, S1.

110. Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T., & Sun, F. (2006). Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC genomics, 7, 31.

111. Via, M., Ziv, E., & Burchard, E. G. (2009). Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. Clinical genetics, 76(3), 225-35.

112. Vinogradov, A. E. (2003). DNA helix: the importance of being GC-rich. Nucleic acids research, 31(7), 1838-44.

113. Wang, D. G. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. Science, 280(5366), 1077-1082.

114. Wang, W. Y. S., & Todd, J. A. (2003). The usefulness of different density SNP maps for disease association studies of common variants. Human molecular genetics, 12(23), 3145-9.

115. Willaert, A. (n.d.). Human Gene Mapping and Disease Gene Identificaton.

116. Zelent, A., Greaves, M., & Enver, T. (2004). Role of the TEL-AML1 fusion gene in the molecular pathogenesis of childhood acute lymphoblastic leukaemia. Oncogene, 23(24), 4275-83.

117. Zhang, Y., Gostissa, M., Hildebrand, D. G., Becker, M. S., Boboila, C., Chiarle, R., Lewis, S., et al. (2010). The role of mechanistic factors in promoting chromosomal translocations found in lymphoid and other cancers. Advances in immunology, 106, 93-133.

118. Zhao, Z., Fu, Y.-X., Hewett-Emmett, D., & Boerwinkle, E. (2003). Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene, 312, 207-213.

119. Zheng, J., Khil, P. P., Camerini-Otero, R. D., & Przytycka, T. M. (2010). Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome. Genome biology, 11(10), R103. BioMed Central Ltd.

120. Zöllner, S., & von Haeseler, a. (2000). A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. American journal of human genetics, 66(2), 615-28.

Electronic references

1. Basic concepts: Linkage disequilibrium (URL: http://blogs.discovermagazine.com/gnxp/2007/01/basic-concepts-linkage-disequilibrium)
2. Chemistry of the cell and genetics (URL: http://www.ucl.ac.uk/~ucbhjow/bmsi/bmsi-lectures.html).
3. Meiosis image (URL: http://t1.gstatic.com/images)
4. Genetics (URL: http://www.ucl.ac.uk/~ucbhjow/bmsi/).
5. Genetic Association Database (URL: http://geneticassociationdb.nih.gov)
6. Genetic Home Reference handbook (URL: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp)
7. Genetic / Physical Map image (URL: http://t1.gstatic.com/images)
8. GWA Studies (URL: http://www.genome.gov/gwastudies)
9. Human Genetics – Linkage and Mapping (URL: http://www.uic.edu/classes/bms/bms655/lesson12.html 21/11/2011).
10. Human Gene Mutation Database (URL: http://www.hgmd.cf.ac.uk/ac/index.php)
11. Linkage disequilibrium image (URL: http://www.biometrics.wur.nl)
12. NCBI Resources (URL: http://www.ncbi.nlm.nih.gov)

13. Online Mendelian Inheritance in Man (URL: http://omim.org)

14. Population Genetics (URL: http://bio.classes.ucsc.edu/bio107/).

15. R project (URL: http://cran.r-project.org/mirrors.html)

16. SNP image (URL: http://learn.genetics.utah.edu/)

17. The Ensembl Project (URL: http://www.ensembl.org)

18. The International HapMap Project (URL: http://www.hapmap.org)

19. The 1000 Genome Project Consortium (URL: http:// www.1000genomes.org/)

Appendix 1

cleanup_hotspot_data.py

This program reads through (hotspotdata.txt) file. This program was designed to extract the chromosome name, hotspot co-ordinates and frequency of recombination data and write these results into a tab delimited output file (File 2 – hotspots_data_clean.txt).

```
# open file and read lines into list
directory = "C:/Students/tracey/analysis/"
hotspots_infile = open(directory + hotspots_filename)
hotspots_lines = hotspots_infile.readlines()
outfile = file(directory + 'hotspots_data_clean.txt', 'a')
# for each line, split into list at tabs
x = 0
print 'there are', len(hotspots_lines), 'lines in the file'
for hotspot_line in hotspots_lines:
#         print hotspot_line
    line_list = hotspot_line.split('\t')
    if len(line_list) == 4 and  line_list[0].startswith('chr'):
        print>>outfile, hotspot_line.strip()
    else:
        print 'line', x, 'is an exception'
#       print hotspot_line
    x = x+1
outfile.close()
```

get_gene_coordinates.py

This program reads through multiple .ftp files stored in a single working directory and extracts the gene co-ordinates. This program was designed to extract the ENSEMBL gene ID, gene name, chromosome name, and gene co-ordinates (gene start and gene end) result per gene and write these results into two tab delimited output files (File 3 – disease_gene_coords.txt and File 4 – nondisease_gene_coords.txt).

```
latest_database = "ensembl_mart_65"
nondisease_outfile = file("nondisease_gene_coords.txt", 'a')
disease_outfile = file(disease_gene_coords.txt", 'a')
# END USER INPUTS
#_____
# collect list of disease gene IDs
disease_genes_infile = "disease_genes.txt"
dg_ids = [ ]
infile = open(disease_genes_infile)
lines = infile.readlines()
for line in lines:
    line = line.split('\t')
    dg_id = line[0]
    dg_id = dg_id.strip()
    if dg_id not in dg_ids:
        dg_ids.append(dg_id)
print "there are", len(dg_ids), "IDs in the disease gene ID list"
# connect to ~~local~~ Ensembl database
```

```
import MySQLdb as dbi

dbc = dbi.connect(host="martdb.ensembl.org",

                  db= latest_database,

                  user = "anonymous",

                  port = 5316)

print 'connected'

# get all records one at a time.

x=0

cursor = dbc.cursor()

cursor.execute("""select distinct  stable_id_1023, display_label_1074, name_1059, seq_region_start_1020, seq_region_end_1020

                  from  hsapiens_gene_ensembl__gene__main;""")

row = cursor.fetchone()

while row is not None:

#   print 'row is', row

    ensembl_id =row[0]

    ensembl_id.strip()

    gene_name= row[1]

    gene_name.strip()

    chrom_name = row[2]

    chrom_name.strip()

    gene_start =row[3]

    gene_end = row[4]

    if ensembl_id in dg_ids:
```

```
    print>>disease_outfile, ensembl_id, '\t', gene_name,'\t', chrom_name,'\t', gene_start,'\t', gene_end

    else:

    print>>nondisease_outfile, ensembl_id, '\t', gene_name,'\t', chrom_name,'\t', gene_start,'\t', gene_end

    row = cursor.fetchone()

   x=x+1

   print x

cursor.close()

disease_outfile.close()

nondisease_outfile.close()

dbc.close()
```

hotspot_vs_gene_script.py

This program reads through gene filename (disease_gene_coords.txt and nondisease_gene_coords.txt) files and hotspots filename (hotspots_data_clean.txt) file. This program was designed to extract hotspot co-ordinates per chromosome and create a directory for disease-associated genes and genes in the control set. Then the script measures the distance and frequency of recombination for each gene/hotspot combination on each of the chromosomes for the disease-associated gene list and non-disease gene list. The program then selects the top scoring hotspot data for each gene and writes these results into two tab delimited output files (File 5 – score_output_disease_gene_coords.txt and File 6 – score_output_nondisease_gene_coords.txt).


```
from __future__ import division
# define source files
x = 0
gene_filename =['nondisease_gene_coords.txt', 'disease_gene_coords.txt'] #
hotspots_filename = "hotspots_data_clean.txt"
chromosomes = ["1", "2", "3", "4", "5", "6", "7", "8","9","10","11","12", "13",
            "14", "15", "16", "17", "18", "19", "20","21" ,"22", "X"]
hotspots = {}
for chromosome in chromosomes:
   hotspots[chromosome] = []
print hotspots
# Get hotspot coordinates and create dictionary
# open file and read lines into list
```

```
hotspots_infile = open(directory + hotspots_filename)

hotspots_lines = hotspots_infile.readlines()

# for each line, split into list at tabs

x = 0

print 'there are', len(hotspots_lines), 'lines in the file'

for hotspot_line in hotspots_lines[27:]:

    x = x+1

    hotspot_line = hotspot_line.split('\t')

    chrom = hotspot_line[0]

    chrom = chrom.strip('chr')

    chrom = chrom.strip()

    hp_position = hotspot_line[1]

    hp_rate = hotspot_line[2]

    if chrom in chromosomes:

        hotspots[chrom].append((hp_position,hp_rate))

for gene_file in gene_filename:

    print 'gene_file is', gene_file

    genes = {}

# Get gene coordinates and create dictionary

    gene_infile = open(directory + gene_file)

    score_outfile = file(directory + "score_output_" + gene_file, 'a')

    print>>score_outfile,
"gene_ensembl_id\tmax_score\thp_position\thp_frequency\tdistance\tgene_chrom\tgene_name\tgene_start\tgene_end"

    gene_lines = gene_infile.readlines()
```

```
for gene_line in gene_lines:
    gene_line = gene_line.split('\t')
    gene_ensembl_id = gene_line[0]
    gene_ensembl_id = gene_ensembl_id.strip()
    gene_name = gene_line[1]
    gene_name = gene_name.strip()
    gene_chrom = gene_line[2]
    gene_chrom = gene_chrom.strip()
    gene_start = int(gene_line[3])
    gene_end = int(gene_line[4])
    if gene_chrom in chromosomes:
        hotspot_list = hotspots[gene_chrom]
        scores = {}
        for item in hotspot_list:
            hp_position = int(item[0])
            hp_frequency = float(item[1])
            if hp_position < gene_start:
                distance = gene_start - hp_position
            if hp_position > gene_end:
                distance = hp_position - gene_end
            if hp_position > gene_start and hp_position <gene_end:
                distance = 0.1
            hotspot_gene_score = float(hp_frequency/distance)
            scores[hotspot_gene_score] = (hp_position, hp_frequency, distance)
```

```
        max_score = max(scores)

        values = scores[max_score]

#       print gene_name, gene_chrom, gene_start, max_score, values

        hp_position = values[0]

        hp_frequency = values[1]

        distance = values[2]

        print>>score_outfile, gene_ensembl_id,'\t', max_score,'\t',
        hp_position,'\t',hp_frequency,'\t',distance,'\t', gene_chrom, '\t', gene_name, '\t',
        gene_start, '\t', gene_end

    score_outfile.close()
```

---

get_all_snp_info.py

This program reads through multiple .ftp files stored in a single working directory and downloads the SNP co-ordinates and chromosome name. This program was designed to extract the SNP ID, chromosome name, and SNP position result per gene and write these results into separate tab delimited output files (snp_coords_(chrom_x).txt).

```
latest_database = "snp_mart_65"
# END USER INPUTS
#_____________________________________________________________
# connect to local database
import MySQLdb as dbi
dbc = dbi.connect(host="martdb.ensembl.org",
                  db= latest_database,
                  user = "anonymous",
                  port = 5316)
print 'connected'
# create directory of all data
chromosomes = ["1", "2", "3", "4", "5", "6", "7", "8","9","10","11","12", "13",
               "14", "15", "16", "17", "18", "19", "20","21" ,"22", "X", "Y"]
# get all records one at a time.
for chrom_x in chromosomes:
    outfile = file("snp_coords_" + chrom_x + ".txt", "a")
    x=0
    print 'chrom', chrom_x
```

```
    cursor = dbc.cursor()
    cursor.execute("""select distinct variation_name_2026, name_1059, seq_region_start_2026
                    from  hsapiens_snp__variation_feature__main
                    where name_1059 = '%s';""" % chrom_x)
    row = cursor.fetchone()
    while row is not None:
        snp_id = row[0]
        chrom_name = row[1]
        snp_pos =row[2]
        print>>outfile, snp_id, "\t", chrom_name, "\t", snp_pos
        x =x + 1
        row = cursor.fetchone()
    print 'completed for chromosome', chrom_x,"; there are", x, " snps"
    cursor.close()
dbc.close()
outfile.close()
```

---

split_genes_by_chrom.py

This program reads through gene filename (disease_gene_coords.txt) and (nondisease_gene_coords.txt) files and separates the disease-associated gene and genes in the control set by chromosome name and writes these results into separate tab delimited output files (filename[disease_gene/nondisease_gene] _gene_chrom.txt).

```
disease_genes = [disease_gene_coords.txt", 'disease_genes']
nondisease_genes  = [nondisease_gene_coords.txt", 'nondisease_genes']
files = [disease_genes, nondisease_genes]
for filename in files:
    infile = open(filename[0])
    lines = infile.readlines()
    for line in lines:
        line = line.strip()
        linedata = line.split('\t')
        gene_chrom = linedata[2].strip()
        outfile = file("filename[1] + "_" + gene_chrom + ".txt", 'a')
        print>>outfile, line
        outfile.close()
```


---

get_gene_snp_count.py

This program reads through gene filenames (disease_gene_coords.txt and nondisease_gene_coords.txt) files. This program was designed to extract the ENSEMBL gene ID, gene name, chromosome name, gene co-ordinates and SNP count per gene and write these results into separate tab delimited output files (snp_count_disease_genes_chrom_(x).txt).

```
gene_coords_dg = ("disease_genes_", 'disease_genes')

gene_coords_ndg = ("nondisease_genes_", 'nondisease_genes')

outfile1 = file("snp_count_disease_genes.txt", 'a')

print>>outfile1,
"ensembl_id\tgene_name\tgene_chrom_name\tgene_start\tgene_end\tsnp_count"

outfile2 = file(snp_count_nondisease_genes.txt", 'a')

print>>outfile2,
"ensembl_id\tgene_name\tgene_chrom_name\tgene_start\tgene_end\tsnp_count"

outfile1.close()

outfile2.close()

# END USER INPUTS

#_____________________________________________________________

# create directory of all data

print "creating directory of all snps"

chromosomes = ["1", "2", "3", "4", "5", "6", "7", "8","9","10","11","12", "13",

        "14", "15", "16", "17", "18", "19", "20","21" ,"22", "X", "Y"]

for chrom in chromosomes:

    infile_name = "snp_coords_" + chrom + ".txt"

    print 'chrom is', chrom

    snp_ids = []
```

```
    snp_data = []
# get all records one at a time from file
    infile = open(infile_name)
    print 'infile open for chrom ', chrom
    lines = infile.readlines()
    print 'there are', len(lines), 'lines'
     for line in lines[:100000]:
        #print 'line is', line
        snp_coords = line.split("\t")
        snp_id = snp_coords[0]
        snp_id = snp_id.strip()
        chrom_name = snp_coords[1]
        chrom_name = chrom_name.strip()
        snp_pos =snp_coords[2]
        snp_pos = int(snp_pos)
        if snp_id not in snp_ids:
            snp_ids.append(snp_id)
            snp_data.append((snp_id, snp_pos))
    print 'there are', len(snp_ids), 'in snp_ids'
#_____________________________________________________________________
# Get gene coordinates and find snps that fall in the coordinate
    print "Processing gene coordinates"
    gene_coords = [gene_coords_dg, gene_coords_ndg]
    for item in gene_coords:
        outfile = file(snp_count_" + item[1].strip() + ".txt", 'a')
```

```
    #print>>outfile,
"ensembl_id\tgene_name\tgene_chrom_name\tgene_start\tgene_end\tsnp_count"
    for chromosome in chromosomes:
        infile_genes = open(item[0] + chromosome + '.txt' )
        lines_genes = infile_genes.readlines()
        print 'there are', len(lines_genes),'gene lines'
        for line_genes in lines_genes:
          line_genes = line_genes.split("\t")
          ensembl_id = line_genes[0].strip()
          gene_name = line_genes[1].strip()
          gene_chrom_name = line_genes[2].strip()
          gene_start = line_genes[3].strip()
          gene_start = int(gene_start)
          gene_end = line_genes[4].strip()
          gene_end = int(gene_end)
           if gene_chrom_name == chrom:
              snp_count = 0
              for data in snp_data:
                 snp_pos = int(data[1])
                 if snp_pos >= gene_start and snp_pos <= gene_end:
                    snp_count = snp_count +1
                    #print data, snp_count
            print>>outfile, ensembl_id,"\t", gene_name,"\t", gene_chrom_name,"\t",
gene_start,"\t", gene_end,"\t", snp_count
    infile_genes.close()
     outfile.close()
    dbc.close()
```