The Genomics of Type 1 Diabetes Susceptibility Regions and the Effect of Regulatory SNPs

Sylvia Enobong Beka

Submitted to the University of Hertfordshire in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

University of Hertfordshire

March, 2015

Abstract

Human complex diseases, like Diabetes and Cancer, affect many people worldwide today. Despite existing knowledge, many of these diseases are still not preventable. Complex diseases are known to be caused by a combination of genetic factors, as well as environmental and life style factors. The scope of this investigation covered the genomics of Type 1 Diabetes (T1D). There are 49 human genomic regions that are known to carry markers (disease-associated single nucleotide mutations) for T1D, and these were extensively studied in this research. The aim was to find out in how far this disease may be caused by problems in gene regulation rather than in gene coding. For this, the genetic factors associated with T1D, including the single point mutations and susceptibility regions, were characterised on the basis of their genomic attributes. Furthermore, mutations that occur in binding sites for transcription factors were analysed for change in the conspicuousness of their binding region, caused by allele substitution. This is called SNP (Single nucleotide polymorphism) sensitivity. From this study, it was found that the markers for T1D are mostly non-coding SNPs that occur in introns and non-coding gene transcripts, these are structures known to be involved in gene regulatory activity. It was also discovered that the T1D susceptibility regions contain an abundance of intronic, non-coding transcript and regulatory nucleotides, and that they can be split into three distinct groups on the basis of their structural and functional genomic contents. Finally, using an algorithm designed for this study, thirty-seven SNPs that change the representation of their surrounding region were identified. These regulatory mutations are non-associated T1D-SNPs that are mostly characterised by Cytosine to Thymine (C-T) transition mutations. They were found to be closer in average distance to the disease-associated SNPs than other SNPs in binding sites, and also to occur frequently in the binding motifs for the USF (Upstream stimulatory factor) protein family which is linked to problems in Type 2 diabetes.

Acknowledgement

I wish to express my profound gratitude to my principal supervisor, Dr Rene te Boekhorst for his continual support, guidance and enthusiasm throughout this project. I am also thankful to my second supervisors, Dr Maria Schilstra and Dr Colin Egan, for their friendship and helpful contribution to this work.

The idea for this project originated with Dr Irina Abnizova of the Wellcome Trust Sanger Institute in Hinxton, Cambridge. I would like to acknowledge her invaluable contribution, advice and support towards this research, and to thank her for the many opportunities to attend seminars and courses at the Wellcome Trust Sanger Institute, and at the University of Cambridge.

This research was funded by a University of Hertfordshire research studentship. For this, my gratitude goes to Dr Volker Steuber and Prof. Bruce Christianson for placing me in a PhD. position and for being supportive throughout my study. I am also thankful to Dr Martina Doolan, Dr Ela Bryson, Dr Katie Graeme-Cook and Prof. Robert Slater for giving me opportunities to grow as a female researcher by involving me in the Athena Swan Committee, the Research Development Working Group and Bioinformatics projects in the School of Life and Medical sciences. Also, sincere thanks to Prof. Conrad Bessant, Dr Yi Sun, Prof. Daniel Polani and Dr Andreas Kukol for their roles as external and internal PhD examiners.

I acknowledge my dearest friends, Miss Nkoyo E. Ekop, Fr. Charles Berebon and Mrs Minna Browne, who have been very kind and showed a lot of care and support during my studies, especially by staying in touch and visiting whenever they could. Also, a big thank you to the brilliant new friends I have met, especially Mrs Lorraine Nichols in the administrative office and my colleagues, with whom I worked in the Science and Research Technology Institute (STRI) Room IE114, for their kindness and for good times.

On a personal note, I am very grateful to my dear parents, Dr F.T. Beka and Mrs Shawn Beka, who have supported me in every way through out my life and education, have been a great source of encouragement and have prayed for me endlessly. This work is dedicated to both of you. I thank my dear siblings, Dr Jason Beka, Patrice Beka, Francis Beka and especially dear "Dr to be!" Nathan Beka, for their care, support and friendship.

Most of all, I am thankful to God for life, health and the ability to have carried out this research.

Table of Contents

List of Tables

List of Figures

Chapter 1

Motivation

Human complex diseases, including diabetes, cancers, and neuro degenerative disorders, are major challenges because they still affect many people worldwide today and are difficult to decipher. Diabetes affects 347 million people worldwide, and Type 1 Diabetes is the third most prevalent chronic disease of childhood, affecting up to 0.4% of children by the age of thirty. Although much research has been done towards finding the cause of T1D over the last four decades, the exact mechanism leading to its onset is still not known (Noble and Erlich, 2012). Despite existing knowledge, the disease is not preventable (WHO Diabetes fact sheet 312, 2015).

Complex diseases are caused by a combination of genetic, environmental and life style factors, and describing the aetiology (biological mechanism that leads to disease onset) of such diseases is not an easy undertaking (Noble and Erlich, 2012). In order to delve into the intricacies of any complex disease, the underlying genomics have to be understood; especially as there is now sufficient evidence that genetic variation plays an important role in the determination of individual susceptibility to disease (Knight, 2010). Genome Wide Association Studies (GWAS) involve examining common genetic variation, particularly Single Nucleotide Polymorphism variants (SNPs) (see chapter 2) between different individuals to see if any particular variant is associated with a certain phenotype. It usually involves comparing the genomes (DNA sequences) of a large number of two groups of individuals. One group is the people with the phenotype of interest (cases) and the second is similar people without the phenotype (control). If a form of the SNP (called an allele) is more frequent in people with the phenotype, then the SNP is said to be "associated" with the disease. The associated SNP is considered to mark a region of the human genome which influences the risk of disease. This region is referred to as a disease susceptibility region or a disease risk locus (Burren et al., 2011). Each region may contain other correlated SNPs, genes and biologically active DNA sequences, like binding sites, that are linked with the associated SNP.

GWAS have led to the discovery of SNPs that are significantly associated with a number of complex diseases including T1D, Rhuematoid arthritis, Crohn's disease etcetera (Manolio et al., 2009; Hindorff et al., 2009; Barrett et al., 2007). However, these studies cannot specify which genes are causal. Generally, studies of disease-associated SNPs tend to be strongly gene-oriented. This means that the common focus is on the associated SNPs that occur within coding regions of genes, and their possible effects on gene products. This approach has been quite successful in studies of Mendelian or monogenic diseases (i.e. caused by problems in a single gene). But this approach has been rather unsuccessful in the study of complex diseases which are typically polygenic, and are which are more common in people than monogenic diseases.

Recent studies suggest that SNPs involved in the regulation of gene activity, may actually contribute more to the aetiology of complex diseases than those in coding sequences (Burton et al., 2007, Djebali et al., 2012, Ward and Kellis, 2012). It is therefore worthwhile to move beyond the conventional "one SNP-one gene" approach to probe for the effect of SNPs in other functional parts of the genome that are known to be involved in regulation.

1.1 INTRODUCTION

This research project grows out of interest in the genetics and genomics of complex diseases, particularly Type 1 Diabetes (T1D). The field of genomics has provided the first systematic approaches to discovering genes and cellular pathways underlying a number of diseases (Lander, 2011.). My research is focused on SNP variants that occur in susceptibility regions for T1D. The main aim of the research is to study the impact of SNPs on the regulation of gene transcription, particularly identifying and analysing the effects of SNPs that occur in transcription factor binding sites (TFBS). This study is inspired by and extends unpublished work by Abnizova et al. (2007) which suggests a computational approach for identifying regulatory elements and variants thereof that may affect gene expression particularly through the binding of transcription factors (TFs) to DNA.

The suggestion that the genetic determinants of complex diseases are perhaps better sought in problems associated with gene regulation is due to findings that many of the disease associated variants occur in non-coding DNA sequences within the genome (ENCODE, 2012; Schuab et al., 2012; Hindhorff, 2009; Barrett et al., 2007). Recent efforts by the **Enc**yclopedia **of** **D**NA **e**lements (ENCODE) consortium, to characterise the human genome, have revealed that most of the non-coding part of the genome is not inactive but is associated with different forms of regulatory activity (ENCODE, 2012; Thurman, 2012). One important regulatory process that takes place within the genome is the (in-) activation of gene expression through the interaction of a particular type of protein complex called a Transcription Factor (TF) (Lewin, 2008). The process of gene expression takes place in two main steps, transcription and translation. Transcription involves transcribing genetic information (in this case, the DNA sequence of gene parts called "codons") to a messenger RNA template.

The mature messenger RNA is either biologically active itself or is translated into a chain of amino acids that eventually form a protein product (Lodish et al., 2000; Jacob and Monod, 1961).

The process of transcription is finely regulated by different factors, among which TFs play a vital role. These proteins bind to specific DNA sequences near the transcription start site of genes in what is called a promoter region (Whitfield et al., 2012) and control the rate of transcription. The TFs can also bind to DNA within the region to be transcribed or to other distal ("upstream") elements called enhancers or silencers (Lewin, 2008). The region of DNA that interacts with and is bound by a single TF is the Transcription Factor Binding Site (TFBS), which usually ranges in size from 8-10 to 16-20 nucleotides, the building blocks of DNA (Zambeli et al., 2012). TFBSs have characteristic motifs which the TFs recognise and bind to. These motifs are usually similar but not always identical. The binding of TFs to TFBSs will either activate gene expression or block it (Zambeli et al., 2012).

The DNA sequences of any two unrelated people are the same at about 99.9%, yet of the remaining 0.1% there is considerable variation between different individuals. These DNA polymorphisms are usually defined as variation present at more than 1% frequency in the population. They are quite important because they influence how people differ in their risk of disease or even response to medication. SNPs are the most common variants in the genome, with about 10 million thought to exist in human DNA (Laurilla and Lahdeshmaki, 2009). SNPs occur when one of four possible nucleotide bases (A, C, G or T) is substituted by another at a single position. There are other forms of sequence polymorphisms. They include insertions and deletions of one or more nucleotides; short tandem repeated motifs of one to six nucleotides called "microsatellites" or longer repeating "minisatellites" as well as other sequence rearrangements (Knight, 2005). Most of the DNA sequence polymorphisms are of no functional importance. Yet, some that occur in the coding sequence of a gene (i.e. those parts that are translated into protein) have been found to cause disease (Choi et al., 2009). This is because they alter the structure of the encoded protein by changing the identity of an amino acid in the peptide chain of the protein (Cargill et al., 1999), leading to the formation of an aberrant protein. These functional coding polymorphisms can be identified through experimental follow up studies after GWAS by methods such as "exome sequencing [1]" (Ng et al., 2010; Bamshad et al., 2011; Rabbani et al., 2014) and RNA-Seq (a method to measure levels of RNA transcripts including mRNAs, non-coding RNAs and small RNAs) (Cirulli et al., 2010; Wang et al., 2009). While coding polymorphisms have been successfully implicated in some monogenic Mendelian disorders like familial Alzheimer's disease (AD) (Barber, 2012; Betram and Tanzi, 2009; Tanzi and Betram, 2005) and Maturity Onset Diabetes of the Young (MODY) (Peltonen, 2006; Knight, 2005), this approach has not been successful in the study of complex diseases (Sanghera and Blackett, 2012; Ng et al., 2010; Petretto et al., 2007). In those cases, more than one SNP is linked to the disease and these may be positioned in both coding and non-coding parts of the genome. Also, more than one gene contributes to the condition and environmental and lifestyle elements may trigger disease onset. For many complex diseases, these triggers are still unknown (Silverman and Loscalzo, 2013).

[1] Exome sequencing is a popular strategy using exon capture methods to identify rare variants in exons candidate susceptibility genes

Polymorphisms that occur in non-coding parts of the genome, especially in regulatory regions have long been suggested to be important modulators of gene expression. They are also thought to be associated with evolutionary change (Wray et al., 2003; King and Wilson, 1975). Today, non-coding polymorphisms within regulatory regions are receiving increased interest from the scientific community, especially because SNPs associated with complex diseases appear to occur more in non-coding DNA than in coding DNA (Encode, 2012). This suggests their possible influence on gene regulation. There are two important types variation that can occur in a key regulatory sequences, the cis-acting variants and the trans-acting variants. The cis-acting variants occur in cis-acting regulatory elements like TFBSs in promoters and response elements. These regulatory elements occur in the vicinity (near the locus/chromosomal position) of the structural portion of the target gene to be regulated or within the sequence of the gene itself. Cis-acting variants include SNPs in TFBSs. Each TFBS has a characteristic sequence motif which is modelled by what is called a consensus sequence. This term is defined as a sequence of DNA that has similar structure and function in the same or in different organisms. For a particular TFBS, the consensus sequence is determined by calculating the order of most nucleotides found at each position in an alignment of multiple DNA sequences[2]. It shows positions where a nucleotide identity is highly conserved and also where the nucleotide identities are variable. A mutated nucleotide in a regulatory region can impact the consensus sequence of a TFBS in such a way that it alters the affinity with which a TF is recruited or binds to the region. This in turn affects the level of gene expression.

There are two types of mutation that can take place. An up-mutation occurs when the mutated nucleotide causes a sub-sequence in the promoter to look more like the consensus sequence of a binding site. This triggers transcription by making the motif of the binding site more conspicuous. It increases binding intensity of transcription factors, where a tighter bind leads to an up-regulation of gene transcription. In contrast, a down-mutation destroys a conserved nucleotide in a consensus sequence causing it to look less like a binding motif. This reduces binding at the core sequence leading to a down-regulation of transcription.

The second type of regulatory variant, the trans-acting variant, affects a protein that binds to the cis-acting elements to control gene expression. These proteins are referred to as trans-acting elements. A mutation in the gene of a trans-acting element could affect the expression its target gene (Schadt et al., 2003; Yvert et al., 2003; Brem et al., 2002). However, the current focus of complex disease studies and of this thesis is on potential cis-acting variants in regulatory modules.

[2] Sequence alignment is a way of arranging sequences of DNA (or RNA or protein) from the same or different organisms, in order to identify regions of similarity that may be as a consequence of functional, structural, or evolutionary relationships between the sequences. It is a pattern of writing one sequence on top of another where the residues in one position are deemed to have a common evolutionary origin. If the same letter occurs in both sequences then this position has been conserved in evolution. If the letters differ, then it is assumed that the two derive from an ancestral letter (which could be one of the two or neither)

The experimental identification of functional regulatory variants (i.e. SNP alleles that impact regulatory processes like gene expression, gene regulation and post translational modification), especially those in binding sites, is not a straightforward task (Knight, 2014.). Firstly because, it previously involved a slow and costly laboratory process (but which does yield accurate results). Secondly, because there are many variants occurring in non-coding DNA, and it is quite difficult to pinpoint actual functional regulatory variants that may contribute to the phenotype under study. Computational prediction of candidate functional regulatory variants can be quite helpful in identifying regulatory variants, by narrowing down the non-coding variants to a considerable number of candidates for onward experimental verification. A number of computational methods have been previously developed for the identification of candidate functional regulatory variants that are likely to play an important biological role (Laurilla and Lahdesmaki, 2009; Xu and Taylor, 2009; Andersen et al., 2008; Laurilla and Lahdesmaki, 2008; Abnizova et al., 2007). These methods make use of computationally predicted regulatory regions and binding sites for the identification of regulatory variants that may affect function by influencing binding. However, the presence of a binding motif in the genome does not indicate that a transcription factor necessarily binds it in vivo.

Recently, high-throughput methods[3] have boosted experimental detection functional binding sites in the genome. Laboratory methods are combined with massively parallel[4]* DNA sequencing, which is the process of determining the order of nucleotides in a molecule of DNA. One of such methods is **Ch**romatin **I**mmuno-**P**recipitation assays followed by **seq**uencing (**ChIP-seq**) method (Jothi et al., 2008; Johnson et al., 2007, Kim and Ren, 2006). Chromatin immune-precipitation (ChIP) is a method used to survey interactions between proteins and DNA as well as proteins and RNA. The process is aimed at determining whether particular proteins interact with specific regions in the genome that could be promoters and enhancers or binding sites. As the name implies, ChiP-seq combines chromatin immune-precipitation (ChIP) with massively parallel DNA sequencing to identify the protein binding sites in DNA.

Another method is DNase-seq (DNase I-hypersensitive site identification by sequencing), a molecular biology that is used for identification of regulatory regions especially promoters in the genome. It is based on genome-wide sequencing of regions that are super sensitive to cleavage by the DNase I enzyme. DNase I hypersensitive sites are thought to be characterized by open, accessible chromatin (Tsompana and Buck, 2014; Boyle et al., 2008).

[3] Methods involving the use of automation equipment with classical biology techniques to address biological questions that cannot be achieved using conventional methods

[4] This means high-throughput approaches to DNA sequencing also called next generation sequencing.

Also known as open chromatin, accessible chromatin regions are identified as **n**ucleosome[5]-**d**epleted **r**egions (NDRs) (Giresi et al., 2007; Kim et al., 2007; Hogan et al., 2006), which are often associated with regulatory factor binding. They have been shown to be associated with all known classes of active DNA regulatory elements, including promoters, enhancers, silencers, insulators, and locus control regions (Cockerill, 2011; Gross and Garrard 1988). Also, 30 years of research have shown that DNase I hyper-sensitive (HS) sites are markers for these different types of genetic regulatory elements (Felsenfeld and Groudine, 2003; Gross and Garrard, 1988; Stalder et al., 1980). FAIRE-seq (**F**ormaldehyde-**A**ssisted Isolation of **R**egulatory **E**lements followed by **seq**uencing), the successor of DNase-seq is also a molecular biology method used to determine the sequence of a DNA region in the genome that is associated with regulatory activity (Giresi et al., 2007). These experimental methods enable accurate and reliable interpretation of regulatory events in the genome that are central to biological processes as well as diseases (Illunina help pages). Information gained from these methods are stored in online repositories including the Encode project, the Ensembl genome browser (Cunningham et al., 2015) and the UCSC genome browser (Rosenbloom et al., 2015). Research based on these methods have provided evidence that the presence of SNPs in these regulatory regions of the genome can lead to differences in transcription factor binding between individuals (Chen et al., 2014; Gagliano et al., 2014; Schuab et al., 2012; Kasowski et al., 2010).

For my project, the SNPs that occur in the susceptibility regions for T1D (referred to as T1D-SNPs) will be identified from T1Dbase, a dedicated database for the genomics of T1D (Burren et al., 2011) (see section 2.11). Those that occur in experimentally verified regulatory regions (REG-SNPs) and binding sites (TFBS-SNPs) will be also be identified and accepted as given in the Ensembl genome browser (Cunningham et al., 2014) (see section 2.11). The local neighbourhood (adjacent sequence of nucleotides) of these SNPs will be analysed for change in sequential properties that occurs when the reference allele of the SNP is substituted with its alternate allele, this is referred to as **SNP sensitivity**. The alternate allele may alter the signal strength of the binding site in which the SNP occurs by causing it to become significantly over-represented (more pronounced) or under-represented (less pronounced). A computational method will be developed and implemented to measure the change in representation caused by the presence of the alternate allele of the SNP. Biologically, this process can lead to change in binding affinity of a transcription factors. The outcome of the SNP sensitivity method will be the identification of T1D-SNPs that significantly change the representation of their surrounding

[5] DNA is packaged into the cell nucleus by special proteins called histones. The basic unit of DNA packaging in eukaryotes is a segment of DNA wound around eight histone protein cores. This complex of DNA and proteins forms the chromatin (Kornberg, 1974). The structure of chromatin depends on the stage of the cell cycle. Parts of DNA in chromatin that are under active gene transcription, are structurally loose (or more loosely packed) allowing access to polymerase enzymes and transcription factors. This form of chromatin is called Euchromatin and makes up 92% of the human genome (IHGCS, 2004; Cooper, 2000). The DNA of less active genes are more tightly packed and referred to as heterochromatin (Dame, 2005; Cooper, 2000)

sequential neighbourhood. These SNPs will be suggested as candidate functional regulatory SNPs that may influence gene expression by causing alteration of TF binding.

1.2 Research Framework

1.2.1 Aim

The aim of this project is to elucidate the impact of SNPs on the regulation of Type 1 Diabetes (T1D). The set of SNPs to be studied are those that occur in the T1D susceptibility regions, which have been mapped by GWAS and SNP genotyping studies (Barrett et al., 2009). These will be referred to as T1D-SNPs. There are two sub-sets of the T1D-SNPs, the disease-associated and non-associated SNPs. The former are those that have been identified by GWAS as having a statistically significant high occurrence in individuals that have T1D compared to those who do not.

1.2.2 Research Questions

The key research question is to find out in how far T1D can be considered as a disease caused by disruptions in gene regulation rather than disruptions in protein coding. This theme is addressed by the following enquiries:

a) What proportion of T1D-SNPs is located in various genomic parts, such as introns, exons and upstream regions?

b) What proportion of the T1D-SNPs is located in regulatory modules?

c) How many of the T1D-SNPs in regulatory regions are located in transcription factor binding sites (TFBS)?

d) Does the over- or under-representation of a binding motif containing a SNP variant differ significantly from that of its alternate allele?

Questions (a) to (c) are concerned with the possibility of a SNP being in a regulatory region, while question (d) is explicitly about its possible effect on regulatory mechanism, in other words a SNP or its variant being a ‘recognition beacon’ for a transcription factor. Crucial to tackling the research questions are (a) a complete as possible, reliable dedicated and easy to work with database, and (b) to build and apply an appropriate statistical method test for SNP sensitivity.

1.2.3 Objectives

Using available information about T1D, this project will involve first of all establishing the distribution of disease and non-disease associated SNPs over the various genomic parts. This will be done to investigate if the T1D-SNPs occur more frequently in certain genic positions, particularly the non-coding genic positions, than non-associated SNPs. To do this, the genomic regions that confer susceptibility to T1D as well as the SNPs that occur within these regions will be identified from two online databases, T1Dbase and Ensembl genome browser, which are described in the literature review.

Subsequently, the susceptibility regions themselves will be characterised by their genomic properties, which will include factors such as total region size, number of genes and SNPs contained. This will be done to find out if the T1D susceptibility regions differ strikingly in genomic content among each other, and also if such eventual differences are related to the presence of loci associated with other autoimmune diseases[6].

Finally, the SNPs that occur in TFBS will be identified using an online software called the Variant Effect Predictor tool (VEP) (McLaren et al., 2010). This tool searches the Ensembl genome browser to locate SNPs that occur in regulatory sequences, including binding motifs. VEP has been chosen because the binding motifs used in prediction of TFBS-SNPs are from the Jasper database (Mathelier et al., 2014). This is the largest and freely accessible online resource that contains information for transcription factor binding motifs in genomes of different organisms. Also, the identified TFBS-SNPs will be those that have been verified by experimentation using such methods as DNase-seq and FAIRE-seq mentioned in Section 1.1. Subsequently, the test for SNP sensitivity will be performed.

The overall intention of my study is to analyse the consequence of SNPs that are located in TFBS and regulatory regions and to produce a list of those T1D-SNPs that significantly change the over- or under-representation of their surrounding sequential neighbourhood. A further intention is also to see if this signal enhancing or reducing effect is stronger than that of other T1D-SNPs that are in regulatory regions but not in binding sites.

1.3 Contribution To Knowledge

Through this research, the following facts have been discovered about the genomics of T1D:

1) Characterisation of the disease-associated and non-associated T1D-SNPs

[6] An illness that occurs when the body tissues are attacked by its own immune system. In an autoimmune disorder, the immune system does not distinguish between healthy tissue and antigens. As a result, the body sets off a reaction that destroys normal tissues

a) T1D-SNPs occur more in non-coding DNA than in coding DNA, although it must be noted that non-coding DNA is much more abundant than coding DNA.

b) Disease-associated T1D-SNPs occur relatively more in some non-coding DNA parts than non-disease associated T1D-SNPs. The genic profiles of disease-associated SNPs show that they occur most often in introns overlapping with non-coding gene transcripts.

c) The genic profiles of the non-associated SNPs show that they also occur most often in introns, but not overlapping with non-coding gene transcripts.

d) T1D-SNPs may affect more than one process because many of them occur in overlapping alternative transcripts of the same gene or in transcripts of overlapping genes.

e) The disease-associated T1D-SNPs occur significantly more in overlapping transcripts than non-associated T1D-SNPs.

2) Characterisation of the T1D susceptibility regions

a) T1D susceptibility regions, characterised by features reflecting genomic content, they can be grouped into three clusters

b) One cluster of regions is characterised by high counts of intronic nucleotides and non-coding transcript nucleotides. A second cluster of regions is characterised by a high occurrence of intergenic nucleotides and high SNP counts. It contains the **HLA** (**H**uman **L**eukocyte **A**ntigen) region, which is the largest T1D region and most associated with the disease (see section 2.3). The third cluster is characterised by high gene density as well as non-coding transcript nucleotides.

c) Twenty-five T1D regions carry markers for fourteen other autoimmune diseases. These regions are dispersed across all three clusters. The cluster of regions characterised by high gene density and high counts non-coding transcript nucleotides has the strongest degree of sharing. The regions mostly carry markers for Multiple sclerosis, Irritable bowel disease and Crohn's Disease.

d) The cluster (first) of regions with high intronic and non-coding transcript nucleotide counts has the second strongest degree of sharing, with most regions carrying markers for Rheumatoid arthritis, Ulcerative colitis and Crohn's disease as well.

3) SNP sensitivity

a) Of all 260,000 SNPs in the T1D susceptibility regions, only 92 occur in TFBS. None of these regulatory SNPs is a disease-associated T1D-SNP.

b) 37 of the 92 TFBS-SNPs test positive for SNP-sensitivity. These regulatory SNPs change the sequential properties of the surrounding region in which they occur. Biologically, this implies a possible influence on TFBS recognition and binding by transcription factors.

c) The regulatory SNPs are significantly closer in proximity to the disease associated SNPs than the TFBS-SNPs that were negative for SNP sensitivity.

d) 16 of the 37 regulatory SNPs occur in the HLA region, the susceptibility region with the highest association (odds ratio = 7) to T1D.

e) 37% of the regulatory SNPs are C-T transition mutations, which are thought to reduce binding affinity

f) The regulatory SNPs are most often found within binding motifs for the USF family of regulatory proteins, which have previously been associated with Type 2 Diabetes.

4) Constant change/update to accessible information in biological databases

Our current understanding of the molecular events that functionally characterize cellular biology continues to be revised (Weinberg and Morris, 2013) especially with the advent of technological enhancements that have boosted experimental techniques. Hence, there is need to keep with regular updates that are made to biological databases due to constant revision of genomic information. Experimental techniques, recently boosted by next generation DNA sequencing methods, typically supersede in-silico (computational) methods by yielding more accurate results because they are carried out in-vivo.

Data previously generated by computational methods are now vastly being validated and replaced by experimentally confirmed information. Due to this practice, massive changes to genomic information were seen during the course of this project.

Changes have been made to the number of T1D susceptibility regions as well as region coordinates. The numbers and identities of the disease associated SNPs have also been revised over time. At the beginning of this project, there were 51 susceptibility regions, in the second year the number increased to 55, and by then end of the project the number had reduced to 49.

Considerable changes have also been made to the numbers of SNPs that occur in regulatory regions and TFBSs. In the second year of my project, 973 TFBS-SNPs were identified from the Ensembl genome browser as occurring in binding sites in the T1D susceptibility regions. The following year, this number was reduced ten-fold to 97.

1.4 Thesis Overview

The following is an overview of the subsequent chapters in this thesis:

Chapter 2 is a summary of the genomics of T1D as a complex autoimmune disease. It also outlines the intricacy of gene expression in disease susceptibility regions, as well as the identification of regulatory SNPs that could influence gene expression.

Chapter 3 describes a characterisation study of the associated and non-associated T1D-SNPs. This study involves distinguishing between the associated and non-associated T1D-SNPs on the basis of the types of genomic parts in which they occur. This chapter includes a brief introduction to the work, followed by a presentation of the statistical analysis done in order to differentiate between both SNP groups, and a short summary.

Chapter 4 presents a study of the genomic make up of T1D susceptibility regions. In this study, the regions are classified on the basis of their structural genomic features. Functional genomic attributes are also related to the classes formed on the structural features. The susceptibility region groups are also studied for level of association with other autoimmune diseases. This chapter also contains an introduction, results, and a brief summary.

Chapter 5 entails the identification of regulatory SNPs that may cause change in the presentation of the binding site within which they occur. This is referred to as SNP sensitivity and will be done using an algorithm developed for this project.

Chapter 6 concludes the dissertation. It includes an outline and a review of the main findings of this thesis. It also highlights potential avenues that can be explored for future research.

There are two additional sections after Chapter 6, the references and appendices. Finally, it is important to mention that, in this dissertation, two types of brackets have been used to distinguish between table references within the text. The brackets, **()** and **{}**, are used to denote tables that are within the main body of the text, and tables that are in the appendix, respectively.

CHAPTER 2

LITERATURE REVIEW

The sequencing of the human genome (Collins et al., 2003, IHGSC, 2004), which involved determining the exact order of nucleotide base pairs that make up human DNA, and attempting to identify and map the function of genes of the human genome, has led to significant development in the field of complex disease genomics. The following sub-sections outline the genomics of T1D, the intricacy of gene expression in disease susceptibility regions, as well as the identification of regulatory SNPs that could influence gene expression.

2.1 COMPLEX DISEASE STUDIES

Complex diseases are multifactorial conditions caused by a combination of genetic, environmental, and sometimes lifestyle factors. They are also multi-genic, meaning that more than one gene contributes to disease susceptibility. This is in contrast with Mendelian diseases, which are caused by defects in just one gene. Therefore, complex diseases do not obey the single-gene dominant or single-gene recessive Mendelian pattern of inheritance that is characteristic of single-gene (monogenic) diseases (Davey and Ebrahim, 2004). They are more common than single-gene (monogenic) disorders, yet defining the risk patterns underlying complex diseases is still problematic (Ward and Kelis, 2012; Craig, 2008; Hirschhorn et al., 2002). The study of monogenic diseases (like Huntington's disease (Johnson, 2012), Cystic Fibrosis (O'Sullivan and Freedman, 2009) and Sickle cell anaemia (Gabriel and Przybylski, 2010; Diggs et al., 1933) has been quite successful, and has contributed a great deal towards the current understanding of many forms of genetic diseases (Duncan et al., 2014; Peltonen and McKusick, 2001) including underlying disease molecular mechanisms.

However, the Human Genome Project, which set out to determine the sequence of chemical base pairs that make up the DNA and map all the genes of the human genome (IHGSC, 2004; Lander, 2001), has dramatically accelerated biomedical research, and changed the approach to understanding complex diseases (Bell and Spector, 2011; Craig, 2008). The completion of the project has allowed for precise inference of gene structure and detection of mutations across the genome (Lander, 2011). It has also led to vast improvements in DNA sequencing technologies including high throughput next generation sequencing platforms, RNA sequencing and SNP genotyping methods. As a result, new insights into the genetic pathogenesis of disease continue to be revealed from on-going research projects carried out by consortiums such as the ENCODE project (Bernstein et al., 2012; Birney et al., 2007), the Type 1 Diabetes Genetics consortium

(T1DGC) (Hitner, 2010; Rich et.al., 2009; Rich et al., 2006), and the Wellcome Trust Case Control Consortium (WTCCC et al., 2007; WTCCC et al., 2007). Other genomes, like those of yeast *(Saccharomyces cerevisiae)* (Hong et al., 2008; Dwight et al., 2002; Cherry et al., 1997), mouse (Qi et al., 2005; Shaw, 2004), chimpanzee (Mikkelsen et al., 2005), and bovines (Elsik et al., 2009) genomes have also been sequenced, and have helped the interpretation of the human genome through comparative analysis. These have enabled experimental studies of genes associated with disease susceptibility in model systems (Lander, 2011; Todd, 2010).

2.2 TYPE 1 DIABETES

Diabetes (or Diabetes mellitus) is a set of disorders characterized by either an absolute or a relative deficiency of insulin and/or insulin resistance. T1D accounts for about 10% of all diabetes cases (Maahs et al., 2010). It has been reported to be the second most prevalent chronic disease of childhood, with a peak onset at about twelve years (Imkampe and Gulliford, 2011). The disease affects up to 0.4% of children by the age of 30, with an overall lifetime risk of nearly 1% (Qiao, 2007; Concannon et al., 2005). Both genetic and environmental factors are thought to contribute to T1D susceptibility. Although this disease has been studied since the 1970s (Singal and Blajchman, 1973), its aetiology has not yet been elucidated. However, knowledge about its biochemistry and genetics has increased significantly (Noble et al., 2012). Animal (mouse) models and human studies have shown T1D to be a chronic immune-mediated disease manifested by an autoimmune attack on the pancreatic β-cells in the islets of Langerhans (Heras et al., 2010; Knip and Siljandera, 2008). It is characterized by selective loss of insulin-producing β-cells in the pancreatic islets in genetically susceptible persons (Knip and Siljandera, 2008), which is due to the presence of antibodies wrongly directed against the β-cells and insulin (Gilliam et al., 2004). This condition leads to complete dependence on exogenous insulin to regulate blood glucose levels (Noble and Erlich, 2012). The first indications of an association between T1D and a particular genomic region were reported for the **H**uman **L**eukocyte **A**ntigen (HLA) locus (Cudworth and Woodrow, 1974; Nerup et al., 1974; Singal and Blajchman, 1973), which is described in the following section. Since that discovery, a lot of research into the biochemistry of T1D has been done.

2.3 THE GENETICS OF TYPE 1 DIABETES

The study of the genome to map disease-susceptibility regions for T1D and other multifactorial diseases has been facilitated by recent advances in next generation DNA sequencing methods. Genome wide scans for the identification of SNPs linked with T1D susceptibility have been carried out on large cohorts including collections of families with affected sibling pairs (Pociot et al., 2010). These studies have provided evidence for over forty T1D susceptibility regions, but the exact mechanisms by which the variation found in these regions confer susceptibility to T1D is still not clear (Noble and Erlich, 2012). The most important genes contributing to T1D susceptibility are located in the MHC class II region, also referred to as the Human Leukocyte Antigen (HLA) locus (Burren et al., 2011). The HLA region is located on chromosome 6, and is a system of 240 genes that encode for proteins on the surface of cells that are responsible for regulation of the immune system in humans (Gale and Gillespie, 2014; Noble and Erlich, 2012; Cano, 2007; Hurley et al., 1997). The proteins encoded by the HLA genes, particularly the HLA class I (A, B, and C) and class II (DR, DQ, and DP) antigens, are unique to each individual (Noble and Erlich, 2012).

The HLA genes are highly polymorphic with up to 6500 unique allelic sequences reported as of July 2011, and increasing to 12500 as of February 2015 (*http://www.ebi.ac.uk/imgt/hla/stats.html*). The HLA has been implicated in the aetiology of more than 100 diseases (Delves, 2014; Noble and Erlich, 2012). The risk of disease is determined by specific combinations of alleles referred to as haplotypes, where certain mutated HLA proteins, called antigens[7] are more likely to develop particular diseases. These include complex autoimmune diseases like T1D, Coeliac disease, Systemic lupus erythaematosus (SLE), Sjögren syndrome, Narcolepsy and Ankylosing spondylitis (Delves, 2014; Gonzalez-Galarza et al., 2013; Noble and Erlich, 2012; Apanius et al., 1997).

Presently, 48 other genomic regions, referred to as susceptibility regions, have been found to also confer susceptibility to T1D (Burren et al., 2011; Steck and Rewers, 2011; Yang et al., 2011; Bluestone et al. 2010; Poicot et al., 2010; Todd et al., 2010; Todd et al., 2007). But their contribution is minimal in comparison to the HLA locus (Gillespie, 2014). Also, research has shown that less than 10% of individuals with HLA-conferred diabetes susceptibility actually progress to clinical disease (Knip and Siljandera, 2008, Wenzlau et al., 2008). This implies that additional factors are needed to trigger and drive β-cell destruction in genetically predisposed persons (Knip and Siljandera, 2008). Environmental factors are believed to influence the expression of T1D. The reason being that in the case of identical twins, if one twin has T1D, the other twin only has it 30%–50% of the time, despite having the same genome. This means that other factors contribute to the prevalence or onset of this disease (Knip et al., 2005). Indications of environmental influence include the presence of a 10-fold difference in occurrence among Caucasians living in different areas of Europe. In addition, people who move to these destinations tend to acquire the rate of disease of the destination country. Other theories surrounding environmental factors include a virus-triggered autoimmune response in which the

[7] A toxin or other foreign substance which induces an immune response in the body, especially the production of antibodies. It is recognized as non-self by the adaptive immune system triggers an immune response, stimulating the production of an antibody that specifically reacts with it (Albert et al., 2002).

immune system attacks virus-infected cells along with the beta cells in the pancreas. The Coxsackie virus family is implicated (Fairweather and Rose, 2002). Also, a rodenticide (Pyrinuron) and an antibiotic (Streptozotocin) used in chemotherapy for pancreatic cancer are thought to selectively destroy pancreatic cells, leading to T1D onset (Mandal, 2013; Changrani et al., 2006). But evidence given for this is inconclusive. Furthermore, life style factors including psychological stress are also thought to have a negative effect on diabetes (American Diabetes Association, 2014). The symptoms of Diabetes are depicted in Figure 1.

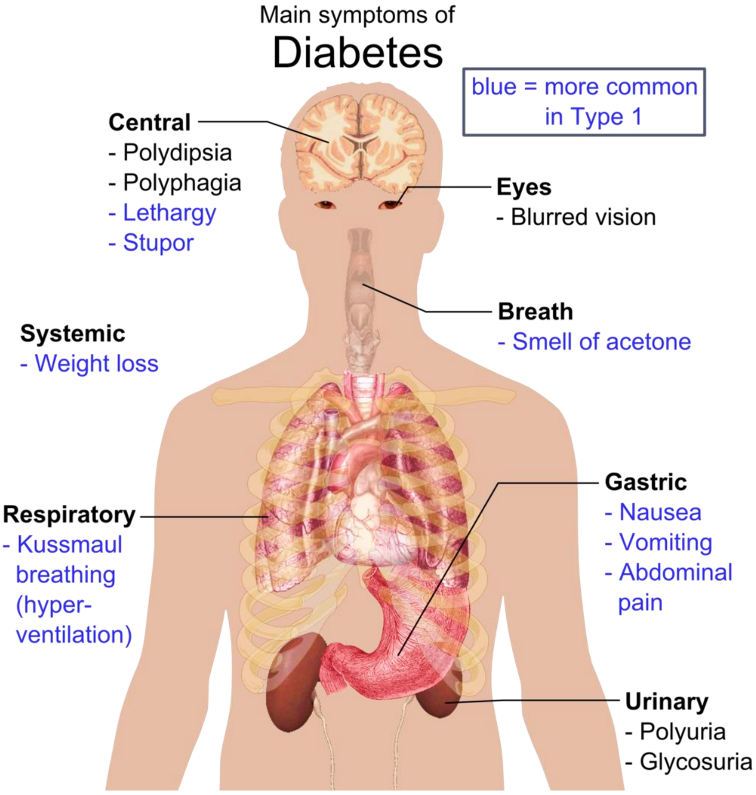


Figure 1. An overview of the symptoms of Diabetes. (Source; Häggström, Mikael).

2.4 Single Nucleotide Polymorphisms (SNPs)

There are numerous variations within the human genome, including SNPs, insertions, deletions, and copy number variations. SNPs are the most prevalent. They are common variations (occurring with a frequency of at least 0.1%) that occur in DNA where a single nucleotide differs between individuals or paired chromosomes of an individual (Barreiro et al., 2008). SNPs have been implicated in a number of diseases and are therefore, essential to the investigation of genetic disorders (Barreiro et al., 2008). The locations of SNPs are studied to assess disease risk and are also used as markers for the identification of disease-associated mutations. SNPs that occur in a coding sequence of a gene are particularly of much interest to scientists, especially if that gene may be involved in the susceptibility to a disease. If the SNP disrupts the production of a functional gene product (protein), then there is a high probability that this SNP will demonstrate a phenotypic effect.

2.5 Genome Wide Association Studies (GWAS)

Genome-wide association studies of large cohorts have been successful in identifying SNPs associated with a large number of phenotypes (Schaub et al., 2012). These range from physical differences between individuals to susceptibility to certain diseases including complex diseases (Hirshhorn et al., 2002). In the 1990s, extensive family-based studies were applied and were quite successful in uncovering the basis of monogenic diseases (Glazier et al., 2002; Beavis, 1998). But they were largely unsuccessful for common complex diseases that afflict most people. In order to study the latter, geneticists conceived principles for genetic mapping based on populations rather than families. This gave birth to the genome wide association study (GWAS).

The aim of a GWAS is to identify variants (like SNPs) that are significantly associated with a particular phenotype. The study involves testing a comprehensive catalogue of common genetic variants in cases (affected individuals) and controls (unaffected individuals) from a population to find those variants associated with a disease (Zhao et al., 2007; Yang et al., 2011). A typical study involves the comparison of common genetic variants in a large collection (up to 200, 000) of individuals to find out if any variant is associated (i.e. occurs more often than expected by chance) with a particular trait. A SNP that occurs more frequently (statistically significant) in the cases is said to be associated with the disease. The SNPs found to be associated with a particular disease are then used as markers for genomic regions that predispose one to susceptibility for that disease. Since the first GWAS in 2005, subsequent studies using high density SNP genotyping platforms soon followed providing evidence for susceptibility regions for T1D and a number of other diseases (Burton et al., 2007; Steck and Rewers, 2011). Currently, there is about 200,000 identified SNPs in T1D regions, and 86 of these have been found to be significantly associated with T1D which are listed in Table 2.

2.6 Type 1 Diabetes Susceptibility Regions

GWAS have uncovered about 100 genomic regions that confer susceptibility to autoimmune diseases including T1D, Rheumatoid arthritis, Multiple sclerosis and Coeliac disease. Forty-nine susceptibility regions for T1D have been mapped by genotyping the most significant T1D associated SNPs (Barrett et al., 2009). SNP genotyping involves the measurement of single nucleotide polymorphisms (SNPs) between individuals. Typically, after SNP genotyping, the pattern of linkage disequilibrium (LD) of the nucleotides surrounding the SNP is assessed. LD is the occurrence of a combination nucleotide variants (SNP and gene alleles or genetic markers) in a population more often or less often than would be expected from a random formation of haplotypes. LD is derive from genetic linkage which is the tendency of alleles that are located in proximity to each other on a chromosome to be inherited together. The DNA sequence that contains the cluster of tightly-linked alleles that are likely to be inherited together then is the haplotype[8] (Lewin, 2008).

Genes in strong LD (within the same haplotype block) as the disease-associated SNPs are assessed for possible functional relevance to T1D. (Bradfield et al., 2011; Burren et al., 2011). The LD block may be further studied for additional SNPs, some of which may be even stronger associated with disease than those identified by the original GWAS. Such extended haplotype investigations also allow scientists to establish whether an association is due to one or more causal variants (Todd et al., 2007). Figure 2, taken from the T1Dbase website, shows the human chromosomes with T1D susceptibility regions indicated by blue bars.

Of the 49 T1D susceptibility region, the HLA association is the strongest with Odd Ratios (ORs) ranging from 0.02 to >11 for specific haplotypes (Noble and Erlich, 2012; Todd et al., 2010). This region contributes to about 50% of genetic susceptibility to T1D, specifically the HLA class II DR-DQ haplotypes (Erlich et al., 2008). Particularly, the DR4-DQ8 and DR3-DQ2 haplotype combinations are present in about 90% of children with T1D (Held et al., 1999; Tait and Boyle, 1986; Deschamps et al., 1980). A genotype containing both haplotypes (DR4-DQ8/DR3-DQ2) carries the highest risk of diabetes, and is commonly seen in 5% of early-onset disease (Gale and Gillespie, 2014). Other strong associations to T1D susceptibility come from polymorphisms in the insulin INS gene (OR = 3.5), the PTPN22 gene (OR = 3.8), the IL2RA and COBL genes (OR = 2.5; 2.4, respectively) (Gillespie, 2014; Pociot et al., 2010; Todd et al., 2010). The rest of the genomic regions that confer susceptibility to T1D have smaller effects with ORs between

[8] Put together, the haplotype is the group of genes that a progeny inherits from one parent

1.1 and 1.9 (Gillespie, 2014; Todd et al., 2010). The names of the T1D susceptibility regions are listed in Table 1 along with the names of the disease associated SNPs and genes. T1D has also been shown to be associated with some other autoimmune conditions like Rheumatoid arthritis, Graves' disease and Malignant anaemia (Heras et al., 2010; Knip and Siljandera, 2008). Markers for these other diseases can be found within the susceptibility regions forT1D. The names of diseases that share T1D susceptibility regions are shown in Table 2.

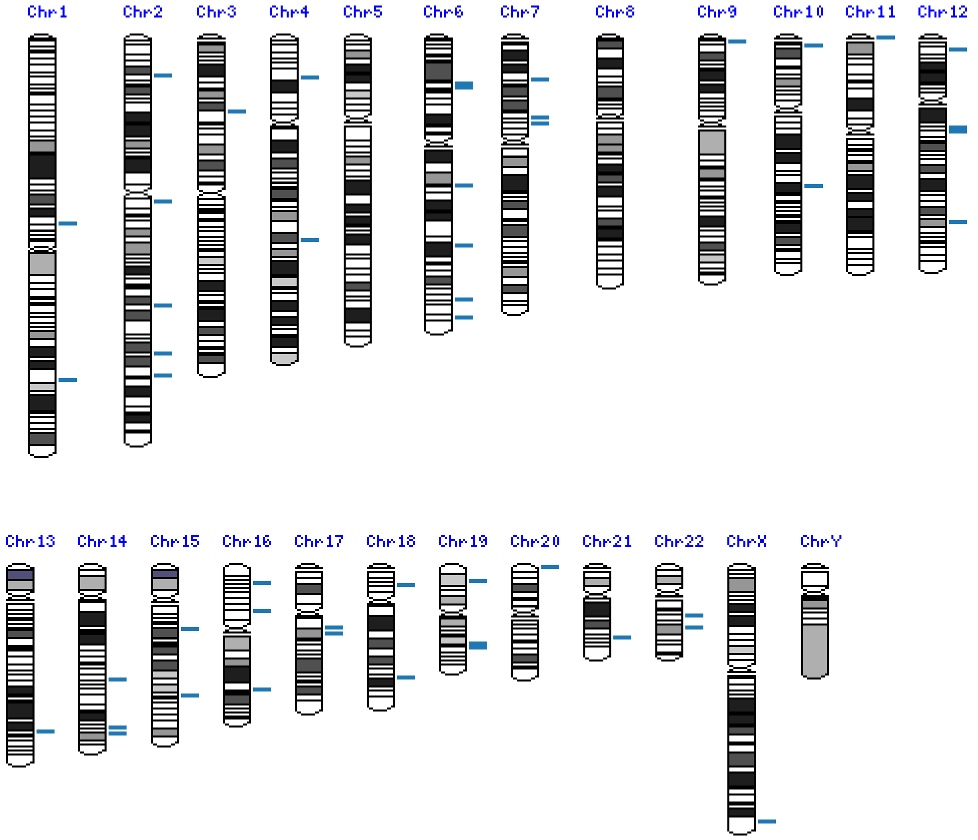


Figure 2. The human T1D susceptibility regions are depicted as blue bars at their respective chromosomal positions (Source: T1Dbase)

Table 1. The names of the T1D susceptibility regions, the disease-associated SNPs and candidate susceptibility genes (Source: T1Dbase).

T1D-Region	Associated SNPs	Candidate Susceptibility genes	Candidate Causal SNP
1p13.2	rs6679677, rs2476601	PTPN22	rs2476601
1q32.1	rs3024493, rs3024505	IL10	rs3024505
2p23.3	rs478222		
2q11.2	rs13415583, rs6740838, rs9653442		rs9653442
2q24.2	rs2111485, rs1990760	IFIH1	rs1990760
2q32.3	rs7574865		
2q33.2	rs3087243, rs11571316	CTLA4	rs3087243, rs11571316
3p21.31	rs333	CCR5	rs333
4p15.2	rs10517086, rs11933540		
4q27	rs4505848, rs6827756	IL2	rs2069762
MHC/HLA	rs6916742, rs9268645	HLA-DQB1, HLA-DRB1, HLA-B, HLA-A	
6q15	rs597325, rs72928038, rs11755527	BACH2	rs11755527
6q22.32	rs9388489	C6orf173	rs9388489
6q25.3	rs1738074	TAGAP	rs1738074
6q27	rs924043		
7p15.2	rs7804356		
7p12.2	rs10272724	IKZF1	rs10272724
7p12.1	rs4948088	COBL	rs4948088
9p24.2	rs10758593, rs7020673	GLIS3	rs7020673
10p15.1	rs10795791, rs12251307, rs7090530	IL2RA	rs12722495, rs11594656, rs2104286
10p15.1	rs11258747		rs947474
10q23.31	rs10509540	C10orf59	rs10509540
11p15.5	rs7928968		
11p15.5	rs689, rs7111341, rs689	INS	rs689
12p13.31	rs10492166, rs917911, rs4763879	CD69	rs4763879
12q13.2	rs705704, rs2292239	ERBB3	rs2292239
12q14.1	rs10877012		
12q24.12	rs17696736, rs653178, rs3184504	SH2B3	rs3184504
13q32.3	rs9585056		
14q24.1	rs1465788		
14q32.2	rs941576, rs4900384		rs4900384
14q32.2	rs941576		rs941576
15q14	rs12908309		rs17574546
15q25.1	rs34593439, rs3825932	CTSH	rs3825932
16p13.13	rs12927355, rs12708716	CLEC16A	rs12708716
16p11.2	rs9924471, rs4788084	IL27	rs4788084
16q23.1	rs8056814, rs7202877		rs7202877
17q12	rs12453507, rs2290400	ORMDL3	rs2290400
17q21.2	rs7221109		
18p11.21	rs2542151, rs1893217, rs478582	PTPN2	rs1893217, rs478582
18q22.2	rs1615504, rs763361	CD226	rs763361
19p13.2	rs2304256		
19q13.32	rs425105		
19q13.33	rs516246, rs601338, rs602662		
20p13	rs2281808		
21q22.3	rs11203203, rs3788013	UBASH3A	rs3788013
22q12.2	rs5753037		
22q12.3	rs229541		
Xq28	rs2664170		

Table 2. Autoimmune diseases that have markers in T1D susceptibilty regions (Source: T1Dbase).

T1D-Region	Other Diseases associated with region
1p13.2	Juvenile Rheumatoid Arthritis, Rheumatoid Arthritis, Crohn Disease, Systemic Lupus Erythematosus, Autoimmune Thyroiditis, Vitiligo
1q32.1	Ulcerative Colitis, Crohn Disease, Inflammatory Bowel Disease, Systemic Lupus Erythematosus
2p23.3	
2q11.2	Juvenile Rheumatoid Arthritis, Rheumatoid Arthritis, Celiac Disease
2q24.2	Ulcerative Colitis, Inflammatory Bowel Disease, Psoriasis, Vitiligo
2q32.3	Rheumatoid Arthritis, Biliary Liver Cirrhosis, Systemic Lupus Erythematosus
2q33.2	Rheumatoid Arthritis, Celiac Disease, Autoimmune Thyroiditis
3p21.31	Celiac Disease
4p15.2	Rheumatoid Arthritis
4q27	Celiac Disease
MHC/HLA	Rheumatoid Arthritis, Celiac Disease, Multiple Sclerosis
6q15	Rheumatoid Arthritis, Celiac Disease, Multiple Sclerosis, Autoimmune Thyroiditis
6q22.32	
6q25.3	Celiac Disease, Multiple Sclerosis
6q27	
7p15.2	
7p12.2	
7p12.1	
9p24.2	
10p15.1	Rheumatoid Arthritis, Vitiligo
10p15.1	
10q23.31	
11p15.5	
11p15.5	
12p13.31	Multiple Sclerosis
12q13.2	
12q14.1	Multiple Sclerosis
12q24.12	Juvenile Rheumatoid Arthritis, Rheumatoid Arthritis, Celiac Disease, Primary Sclerosing Cholangitis, Biliary Liver Cirrhosis, Vitiligo
13q32.3	
14q24.1	
14q32.2	
14q32.2	
15q14	
15q25.1	Celiac Disease, Narcolepsy
16p13.13	Biliary Liver Cirrhosis, Multiple Sclerosis
16p11.2	Crohn Disease
16q23.1	
17q12	Rheumatoid Arthritis, Biliary Liver Cirrhosis
17q21.2	
18p11.21	Celiac Disease, Ulcerative Colitis, Crohn Disease, Inflammatory Bowel Disease
18q22.2	Celiac Disease, Multiple Sclerosis
19p13.2	Rheumatoid Arthritis, Biliary Liver Cirrhosis, Multiple Sclerosis, Psoriasis
19q13.32	
19q13.33	Crohn Disease, Inflammatory Bowel Disease
20p13	
21q22.3	Rheumatoid Arthritis, Vitiligo
22q12.2	
22q12.3	
Xq28	

2.7 Genes and Transcripts

A gene is a molecular unit of heredity in a living organism. A modern working definition of a gene is: a particular region of the genomic sequence, corresponding to a unit of inheritance, of which parts (called exons) are involved in the synthesis of proteins. Proteins, in turn, play an important role in the development and functioning of all known living organisms. Within a gene, exons are interrupted by introns, parts that are not directly contributing to the synthesis of proteins (but may be involved in the regulation of gene activity). Exons are made up of series of three letter nucleotide sequences (codons), also called a reading frame[9]. These are transcribed to a complementary single strand (but with Thymine replaced by Uracil, and the introns "spliced out") called messenger RNA (mRNA). In turn, mRNA functions as a template to which RNA (transfer- or tRNA) temporarily attaches. Each tRNA molecule consists of two functional sites. The first one, called "anticodon", operates as a docking site; it is a sequence of three bases that are complementary to a codon in the messenger RNA. The second functional part attaches to one of the 20 possible amino acids, as specified by the sequence of nucleotides in the (anti-) codon. After dissolving the bonds between tRNA and mRNA, the amino acids link to form a polypeptide sequence which after intricate folding leads to a protein (Figure 3).

The sequence of nucleotides in exons thus determines the string of amino acids and in this way the function and structure of a protein. From this it follows that changes in the succession of nucleotides by mutations (including SNPs, deletions, insertions and copy number variations) may lead to a (often disruptive) change in the protein. These mutations include single nucleotide changes (SNPs, insertions, deletions (together called "indels")) and multiple nucleotide polymorphisms ("micro-satellites", copy number variations and large sequence variations). A particular locus[10] may be occupied by any one of the alleles (transcripts) of a gene or other functional DNA sequences, where an allele is one of several gene transcripts. Gene overlap occurs when the overlaying genes share the same DNA sequence perhaps in a different reading frame or on the opposite DNA strand, and yet do not share regulatory elements or any exons (Sanna et al., 2008; Gerstein, 2007).

[9] A reading frame is a way of splitting the sequence of nucleotides that make up the exon into a set of consecutive, non-overlapping triplets. The triplets equate to amino acids or stop signals during translation and are referred to as codons. An open reading frame (ORF) is the part of a reading frame that has the potential to code for a protein. It is a continuous stretch of DNA that begins with a start codon, usually methionine (ATG), and ends with a stop codon (TAA, TAG or TGA in most genomes) (Brown, 2010)

[10] Nowadays the definition of locus also entails the location of other DNA sequences than just genes, i.e. although a gene has a locus, a locus can contain more than just a gene.

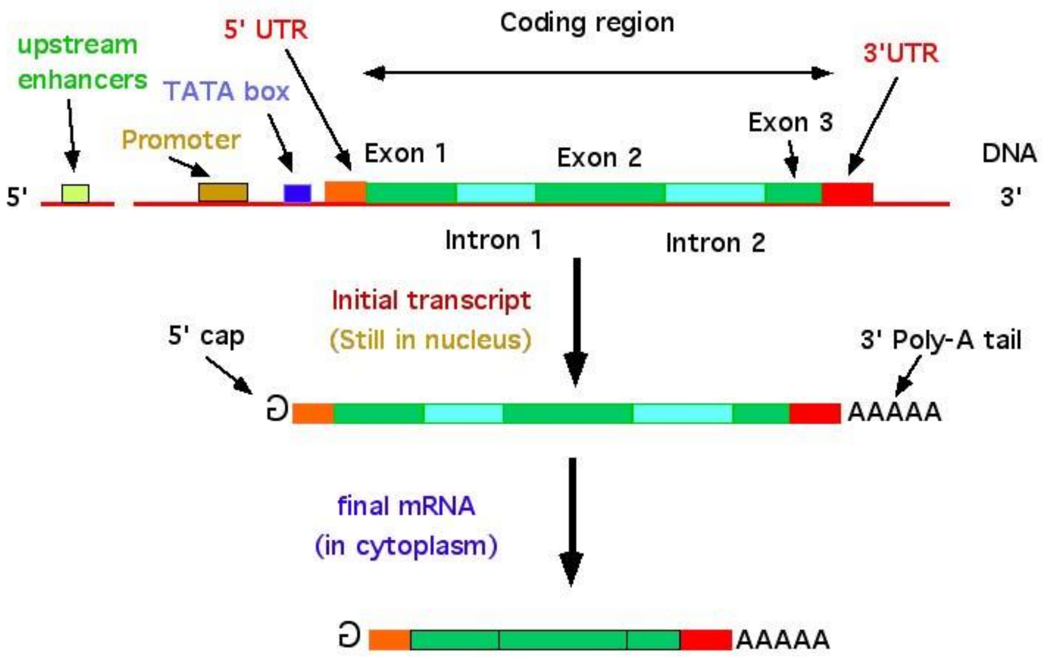


Figure 3. An illustration of transcription of a protein coding gene. The gene is first transcribed to an initial transcript, and the introns are spliced out to form a final mature mRNA. (Source: British Journal of Anaesthesia)

There are also non-coding genes that are not translated into protein, but which produce functional RNA. They differ from the protein-coding genes in that they do not have an open reading frame. They are abundant in the genome and are involved in regulatory activity (Pique-Regi et al., 2011). They include biologically active RNA genes like transfer RNA (tRNA) or small nuclear RNA (snRNA), ribosomal RNA (rRNA), long non-coding RNA (lincRNA), microRNA and silencing RNA (siRNA).

Transcripts are generated by a process, in which a DNA stretch is transcribed to an initial transcript unit called a pre-messenger RNA. The pre-mRNA may be involved in protein synthesis by acting as template for transfer-RNA (or t-RNA). In this case the transcript is called a “coding transcript”. Alternatively, the pre-mRNA is a template for other types of RNA genes that are non-coding. Unlike coding transcripts, the non-coding transcripts do not result in a protein product, but instead are biologically active molecules that play other important roles in the genome such as chromatin[11] maintenance and regulation of gene activity (Pique-Regi et al., 2011). Known products of “non-coding transcripts” include lincRNA, microRNA, ribosomal RNA and other small nuclear RNA genes. Their structure and function is summarised in Table 3. A gene can have more than one transcript. Many genes contain numerous exons and introns,

[11] Chromatin is a complex of macromolecules consisting of DNA, protein and RNA. It packages DNA into a smaller volume to fit in the cell and prevents DNA damage. It is involved in the control gene expression and DNA replication. Chromatin maintenance is any regulatory activity that involves of the preservation of the physical structure of eukaryotic chromatin.

and the exons can be spliced together in more than one pattern to generate multiple, distinct mRNA transcripts. This process is referred to as alternative splicing (Lewin, 2008). These distinct mRNA transcripts, which are referred to as alternative transcripts, transcript variants, splice variants, or isoforms, in turn produce different variants of a protein from the same gene (Guttmacher, 2002) (see Figure 4). The creation of a protein from its gene is called gene expression.

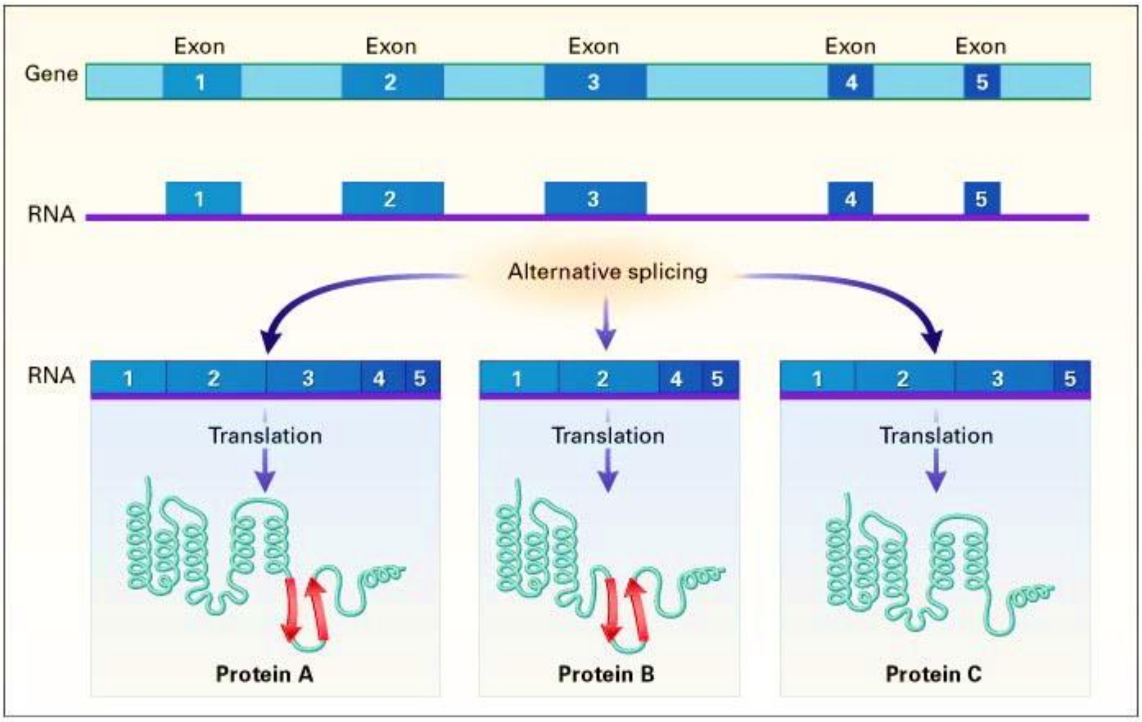
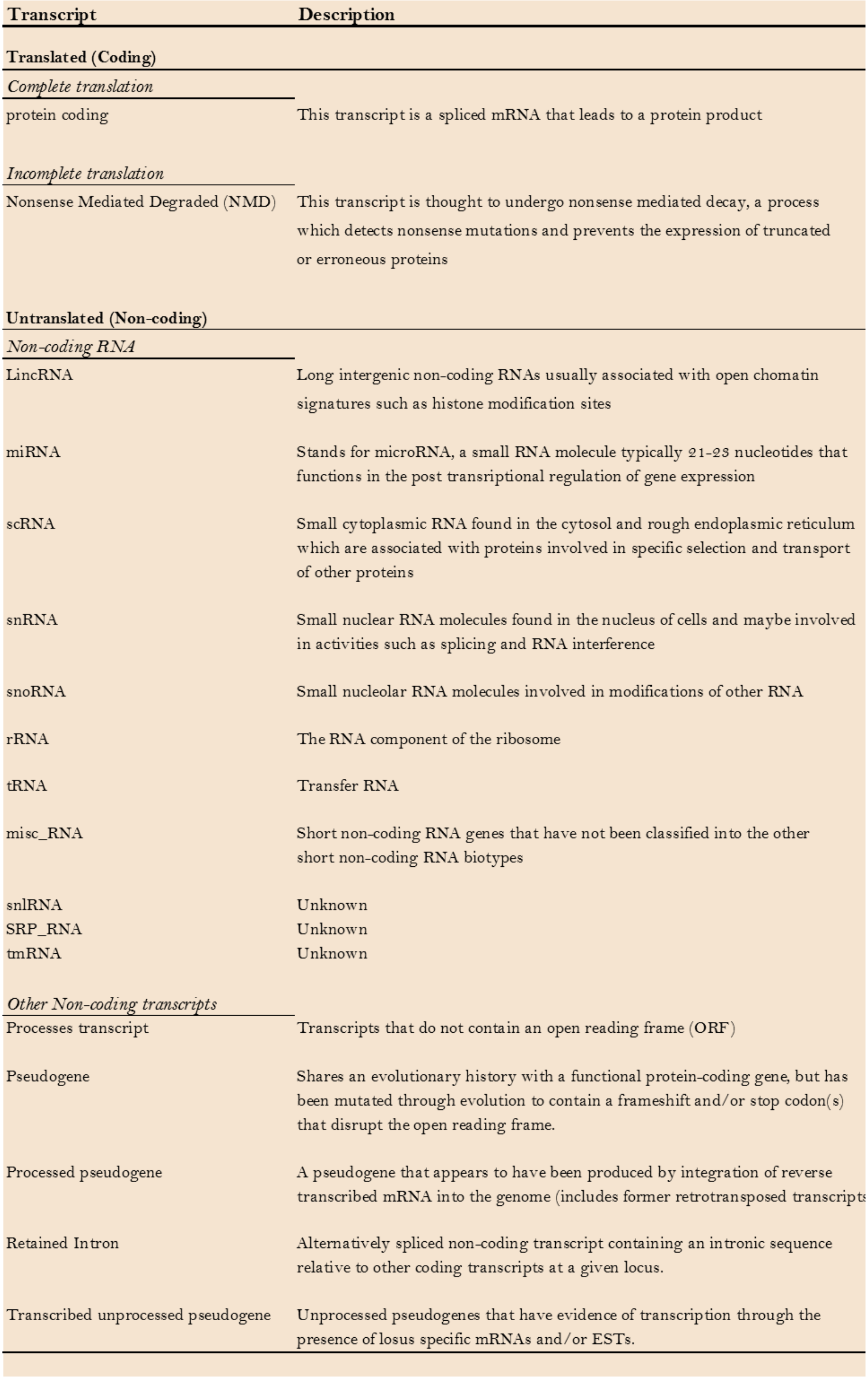


Figure 4. An Illustration of alternative splicing. The exons from a single gene are spliced together in three different patterns which give rise to three variants of the same protein. (Source: Guttmacher and Collins, 2002).

Table 3. Brief definitions of coding and non-coding transcripts (Source: Information given has been extracted from Ensembl Documentation, December, 2014).



Transcript	Description
Translated (Coding)	
Complete translation	
protein coding	This transcript is a spliced mRNA that leads to a protein product
Incomplete translation	
Nonsense Mediated Degraded (NMD)	This transcript is thought to undergo nonsense mediated decay, a process which detects nonsense mutations and prevents the expression of truncated or erroneous proteins
Untranslated (Non-coding)	
Non-coding RNA	
LincRNA	Long intergenic non-coding RNAs usually associated with open chomatin signatures such as histone modification sites
miRNA	Stands for microRNA, a small RNA molecule typically 21-23 nucleotides that functions in the post transriptional regulation of gene expression
scRNA	Small cytoplasmic RNA found in the cytosol and rough endoplasmic reticulum which are associated with proteins involved in specific selection and transport of other proteins
snRNA	Small nuclear RNA molecules found in the nucleus of cells and maybe involved in activities such as splicing and RNA interference
snoRNA	Small nucleolar RNA molecules involved in modifications of other RNA
rRNA	The RNA component of the ribosome
tRNA	Transfer RNA
misc_RNA	Short non-coding RNA genes that have not been classified into the other short non-coding RNA biotypes
snlRNA	Unknown
SRP_RNA	Unknown
tmRNA	Unknown
Other Non-coding transcripts	
Processes transcript	Transcripts that do not contain an open reading frame (ORF)
Pseudogene	Shares an evolutionary history with a functional protein-coding gene, but has been mutated through evolution to contain a frameshift and/or stop codon(s) that disrupt the open reading frame.
Processed pseudogene	A pseudogene that appears to have been produced by integration of reverse transcribed mRNA into the genome (includes former retrotransposed transcripts
Retained Intron	Alternatively spliced non-coding transcript containing an intronic sequence relative to other coding transcripts at a given locus.
Transcribed unprocessed pseudogene	Unprocessed pseudogenes that have evidence of transcription through the presence of losus specific mRNAs and/or ESTs.

2.8 Regulation of Gene Expression

Apart from the protein coding sequences, there are other biologically relevant nucleic acid sequences that play other important roles in the genome such as regulation of gene expression and maintenance of the chromatin structure (Pique-Regis et al., 2011). Regulation of gene expression involves a process that leads to increase or decrease in the production of specific proteins (Jacob and Monod, 1961). It is an important aspect of the cell because it increases the versatility and adaptability of an organism by allowing the cell to produce proteins only when they are needed (Payankaulam, 2010; Jacob and Monod, 1961). Gene expression is regulated at the level of transcription (described in 2.8), which can only occur if transcription factors bind to the DNA. Binding occurs within special nucleotide sequences called regulatory regions that are usually several hundred base pairs long (Lodish et al., 2000). Regulatory regions surround transcription start sites (TSSs) of genes apart from some sequences called enhancers that are located far upstream or downstream of their target gene (Birney et al., 2007; Dineen et al., 2007). Regulatory regions contain transcription factor binding sites (TFBSs) which are short sequences of DNA nucleotides that have distinctive motifs (Zhang et al., 2014). These TFBSs are recognised by the transcription factors which bind preferentially to distinct motifs and activate gene expression (Whitfield et al., 2012). Accurate functional annotation of regulatory elements is therefore important for understanding the basic process of gene regulation (Pique-Regis et al., 2011). Yet, this is still a challenge in modern genomics.

2.9 Identification of Regulatory Regions and TFBS

The genetic basis of gene expression has been investigated across tissues (Dimas et.al, 2012) and populations (Stranger et al., 2012). Variation in genomic regions involved in regulation of gene expression is vital to evolution and disease (Pique-Regi et al., 2011; Nicolae et al., 2010). Computational approaches to the prediction of regulatory sequences have been encouraged through improvements in high throughput DNA sequencing techniques. These methods side step the ultimately more reliable but slow and expensive route of experimental verification (Abnizova et al., 2006). Computational methods have developed significantly in recent years (Chan et al., 2010; Dineen et al., 2010; Huang et al, 2004; Ohler and Niemann, 2001; Stormo, 1990; Jensen and Knudsen, 2000; Vanet et al., 2000; van Helden et al., 2000; Hughes et al., 2000; Workman and Stormo, 2000; Zhu and Zhang, 1999; van Helden et al., 1998; Bailey and Elkan, 1995; Lawrence et al., 1993). Most of the time, the models accurately predict in vitro binding motifs for transcription factors (Andersen et al., 2008; Tronche et al., 1997). Results from computationally identified binding motifs can be found in databases like TRANSFAC (Matys et al., 2006), JASPAR (Mathelier et al., 2014; Bryne et al., 2008), and SCPD (Zhu and Zhang, 1999).

However, there is still considerable need for reliable detection, in vivo, of regulatory regions and biologically relevant sites they contain (Dineen et al., 2010; Guhathakurta, 2006; Hoogendoorn et al., 2004; Hoogendoorn et al., 2003). Current experimental techniques like DNase-seq and its successor FAIRE-seq are applied to human cell lines to verify sequences associated with regulatory activity by detection of DNase hypersensitive sites (Song et al., 2011; Crawford et al., 2006). Combining data from computation and experimental methods can lead to accurate identification of true regulatory regions in the genome.

2.10 Variation in regulatory regions and TFBS

Genetic variation in regulatory regions can influence gene expression. There is now increased interest in regulatory SNPs, which have been suggested to have significant contribution to the aetiology of some complex diseases (Stranger et al., 2012; Wellcome Trust, 2007). However, their identification and evaluation is not effortless due to difficulty in identifying regulatory region prediction (Mariño-Ramírez et al., 2009). Most of the non-coding genome is yet to be deciphered (in terms of function), and the process of regulation is not yet fully described (Altshuler et al., 2008). This makes it difficult to predict the functional effect of regulatory variants (Pique-Regi et al., 2011). The effect of mutations on TF binding have been studied computationally. A number of in-silico (computational) methods to predict candidate regulatory variants that may affect function have been developed (Laurilla and Lahdesmaki, 2009; Xu and Taylor, 2009; Andersen et al., 2008; Laurilla and Lahdesmaki, 2008; Abnizova et al., 2007). These methods make use of computationally recognised regulatory regions to identify candidate regulatory variants. More recent methods make use of data from experimental methods like ChIP-seq and FAIRE-seq (Chen et al., 2014; Gagliano et al., 2014; Landt et al., 2012; Schuab et al., 2012).

2.11 Biological data and databases

The fundamental data for my research are the human T1D susceptibility regions and the SNPs that occur in the susceptibility regions. Three online biological databases, T1Dbase (Burren et al., 2011) Ensembl (Cunningham et al., 2014) and Entrez dbSNP (Sherry et al., 2001) will be used for data collection.

T1Dbase

T1Dbase is a bioinformatics resource of the International Type 1 Diabetes Genetics Consortium (T1DGC) (available at *www.t1dbase.org)*. The 49 regions that affect risk of T1D are listed in this database. Data that can be obtained from T1Dbase include names and chromosomal coordinates of the susceptibility regions. Also available are the identifiers for disease associated SNPs, candidate susceptibility genes and other autoimmune diseases associated with each of the susceptibility regions. For my research, the data I obtained from T1Dbase included a compilation of SNP variants in T1D susceptibility regions, the alleles[12] (reference and mutant) of each SNP locus, the exact location of each SNP in the genome (chromosomal coordinates), and the genic position of each SNP based on gene structure. Genic positions can be classified on the basis of function and structure. The functional classification establishes if the SNP is in a gene or not; and if it is in a gene, whether or not the gene is protein coding or non-coding. The SNPs can also be classified structurally according to their genic position, i.e. whether they are in an intra-genic, gene flanking or inter-genic part of the susceptibility region.

Advantages and Limitations of T1Dbase: It is completely dedicated to the genetics of T1D and as such the database is focused on T1D susceptibility loci data. This eliminates the need to search the much larger human genome data set. The limitation of T1Dbase at the time of this study was that one could not comprehensively retrieve the genic positions (genomic region in which a SNP is located) of the T1D- associated SNPs. This has however changed with a database update in early 2014.

Ensembl

Ensembl is a publicly available web resource that contains automatically annotated genomes. It is integrated with other available biological databases like Jasper for binding motifs. It is a much larger web resource than T1Dbase, and contains general information about the human genome including variants. These include SNPs, insertions, deletions and somatic mutations (Alterations in DNA that occur after conception, meaning that they are not inherited) for several species. Data from Ensembl can be accessed in a number of ways. The names of all the SNPs that occur in the T1D susceptibility regions can be collected from Ensembl using the Biomart tool (Kinsella et al., 2011). To achieve this, the coordinates of the T1D regions obtained from T1Dbase are uploaded to the biomart query page which allows one to search the genome browser and retrieve data like the names, chromosomal positions, and genic positions (referred to as "consequence to transcript", in Ensembl) of the SNPs. The SNP genic positions tell if a SNP is located within a gene, adjacent to a gene or whether they occur in inter-genic positions between gene coding regions, as well as the particular genes in which they are located. Information about genes, including gene names, chromosomal coordinates, biotype (coding or non-coding), and number of splice variants, can also be retrieved from Ensembl.

[12] Allele, in this case, refers to one of two or more forms of the variant. For SNPs, the original (non-mutated) nucleotide is referred to as the reference allele, while its variant form is called the alternate or mutant allele. Although most SNPs have only one mutant allele, some have more than one

Advantages of Ensembl: There is a number of advantages to using Ensembl. (i) It is a larger web resource than T1Dbase and integrates data from a wide range of biological research sources into its database. Therefore, available information is quite comprehensive. (ii) Genic positions for 99% of the variants obtained from T1Dbase could be retrieved. (iii) Ensembl contains quality checks for genetic variants in its variation pipeline. A variant is flagged as failed if certain quality criteria are not met, for instance if none of the variant alleles match the reference allele of the variant. Generally, Ensembl was found to give more detailed information regarding the genic positions of variants compared to T1Dbase.

NCBI-dbSNP

dbSNP (Database of Single Nucleotide Polymorphisms) (Sherry et al., 2001) is a large database for single nucleotide variants. It contains information about single nucleotide variant alleles and the sources of experimental data, and is available at (*http://www.ncbi.nlm.nih.gov/SNP/*). dbSNP was used to cross check information about SNP alleles retrieved from T1Dbase and Ensembl, and also to clarify any obscurities or incompleteness (especially missing alleles) encountered with the retrieved SNP data from both databases. dbSNP is incorporated into NCBI's Entrez system of databases. It contains information about variations in the human, mouse, rat, chimpanzee and the malaria parasite species (Sherry et al., 2001). The database is mainly devoted to single nucleotide substitutions, the rest includes information about insertion/deletion polymorphisms, microsatellite and minisatellite repeats and other uncharacterized heterozygous assays.

Ravendbase.

All data retrieved from T1Dbase and Ensembl were incorporated into an own database, called Ravendbase. This database was designed and implemented by myself at the beginning of this project and completed through an unpublished MSc project (Beka, 2012). This database was created so as to link supplementary information about the T1D SNPs taken from Ensembl with the information retrieved from T1Dbase. This included such information that included the numbers and names of genes and transcripts that the SNPs intersect, the biotypes of the gene transcripts, if a SNP is in a regulatory region or not. At present, some of this information is not available from T1Dbase. Altogether, information collected about the T1D SNPs, for this project, from both biological resources was stored in linked tables in Ravendbase for quick and easy access. Genomic information for 300,707 variants that occur in the susceptibility regions for T1D can be retrieved from this database, with SNPs forming the largest subset. Ravendbase is available online at (*http://ravendbase.com*), the structure of Ravendbase is described in Appendix D.

CHAPTER 3

CHARACTERISTICS OF ASSOCIATED AND NON-ASSOCIATED T1D-SNPS

This chapter is about the categorization of SNPs in the T1D susceptibility regions. Both associated and non-associated SNPs were classified by the type of genomic part in which they occur. The main aim of the work described here is to find characteristics that separate associated T1D-SNPs from non-associated T1D-SNPs. It was found that the associated SNPs occur more in multiple and different genic positions than the non-associated SNPs and most frequently in a combination of intronic regions and non-coding transcripts. In contrast, many non-associated SNPs are frequent in just intronic regions, as well as in gene flanking regions.

3.1 INTRODUCTION

The human genome is littered with millions of SNPs (Christley et al., 2008). SNPs are important as markers for certain diseases or as causative agents. Although the majority of SNPs has minimal effects, some of them have been shown to have detrimental consequences (Zhang et al., 2014; Bush and Moore, 2012). The type of genomic structure in which a SNP is positioned is important because of its possible impact on the biological system.

SNPs in genes may influence protein synthesis by affecting (a) the amino acid sequence of that protein, (b) affecting mRNA transcript stability (lifetime duration[13]) through processes like nonsense mediated decay (Isken and Maquat, 2007), (c) translation rate (like causing translation pausing) due to change in RNA secondary structure (Sacchetti, 2001; Zama, 1999) or through mutations in translation initiation factors (Schwartz and Parker 1999), and (d) alternative splicing by altering the consensus sequence of a splice site (Zhang et al., 2014; Griffith et al., 2008).

SNPs in non-coding regions can alter gene expression by modulating the activity of cis-regulatory elements (Zhang et al., 2014) e.g. transcription factor binding affinity (Griffith et al., 2008), and possibly the activity of RNA genes involved in regulation (see Sections 2.7 and 2.8)

[13] The greater the stability of an mRNA the more protein may be produced from that mRNA. A limit to the lifetime of mRNA enables a cell to alter protein synthesis rapidly in response to its changing needs.

(Chen et al., 2014; Schaub et al., 2012; Ward and Kellis, 2012; Laurilla and Lahdesmaki, 2009; Andersen et al., 2008; Abnizova et al., 2007; Knight, 2005; Stranger and Dermitzakis, 2005).

Currently, there is seventy-nine SNPs have been linked with susceptibility to T1D (Burren, et al., 2011). These disease-associated SNPs have been identified by GWAS as occurring significantly more in individuals who have T1D than in individual who do not have the condition. These disease-associated SNPs are markers for the forty-nine T1D susceptibility regions, and within these regions are over 250,000 other non-associated SNPs that are in linkage with the disease-associated SNPs.

In this chapter, I will classify and contrast the associated and non-associated SNPs by the structural part of the genome in which they occur. The aim of this is to investigate if the disease-associated SNPs occur in other genic positions than non-associated SNPs. Non-coding genic positions are particularly of interest because they may be involved in a variety of gene regulatory activity, and this relates to the main thesis of my research. Characterisation will be done by first establishing the structural part of the genome in which each associated and non-associated SNP is positioned. The term "structural part" refers to coding (exonic) and non-coding (intronic, 5' UTR, 3'UTR, upstream, downstream, non-coding transcript, NMD transcript and intergenic) genic positions (see Figure 5).

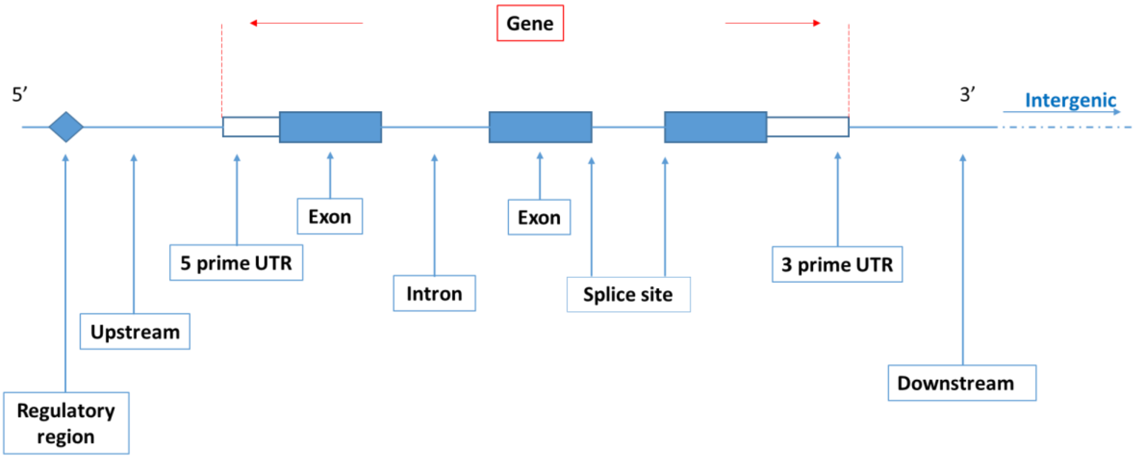


Figure 5. An illustration of the structure of a protein coding gene with possible genic positions of a SNP depicted.

The following sections outline how the associated and non-associated SNPs were the classified by the genic positions in which they occur in the genome, and by the number of gene transcripts that they affect. SNPs that were found to affect multiple transcripts were further characterised by the numbers and different types of genic positions in which they occur. These features were analysed in order to distinguish between the associated and non-associated SNPs.

3.2 T1D SNP Genic Positions

The genic position of a SNP refers to the type of genic structure in which it occurs. Nine main genomic structures are distinguished in my work. They are grouped into three general classes, (i) intra-genic, (ii) gene flanking and (iii) inter-genic (Figure 6).

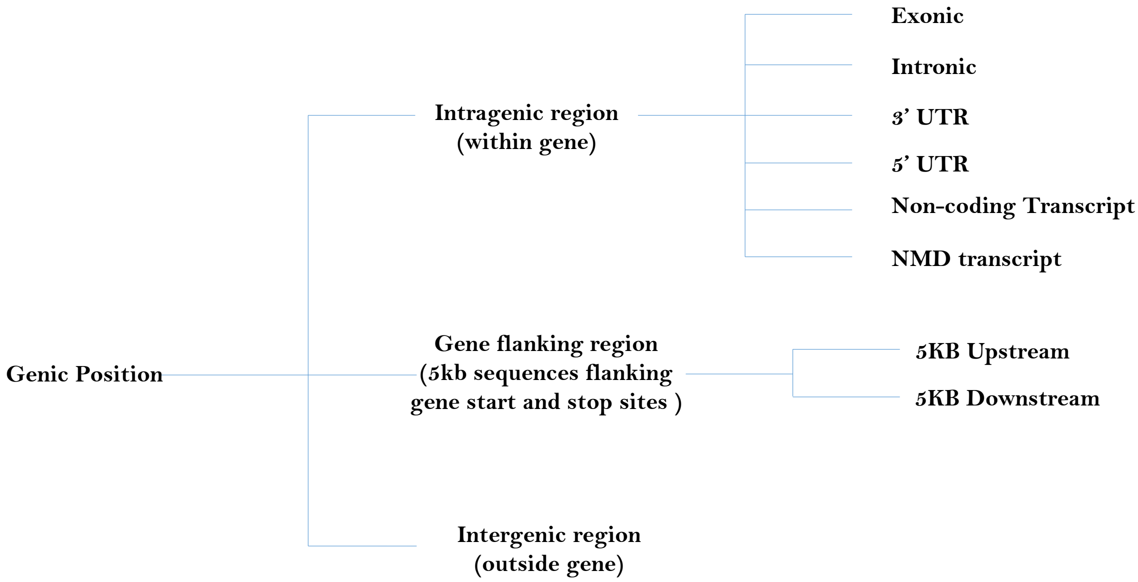

Genic Position
Intragenic region (within gene)
Exonic
Intronic
3' UTR
5' UTR
Non-coding Transcript
NMD transcript
Gene flanking region (5kb sequences flanking gene start and stop sites)
5KB Upstream
5KB Downstream
Intergenic region (outside gene)

Figure 6. A grouping of the nine types of genic positions in which a SNP can occur in the genome

(i) Intra-genic Genic positions: are within the transcripts of genes. A SNP's position may be in any structure of a coding transcript including exons, introns, 3' and 5' UTR sequences. SNPs may also be in a non-coding gene transcripts such as NMD transcripts, and pseudogenes, as well as transcripts of functional non-coding RNA genes like miRNA, lincRNA, snoRNA (see Table 3).

(ii) Gene-flanking Genic positions: include up to 5000 nucleotide base pairs (5 kilo-bases) adjoining the transcription start (upstream) and end (downstream) sites of functional coding and non-coding gene sequences.

(iii) Inter-genic Genic positions: lie between the downstream and upstream sequences of neighbouring genes. These nucleotide sequences are assumed not to represent genes or any other functional (non-coding) sequences.

For this work, the names (variant identifiers) of 260,302 T1D-associated (N = 79) and non-associated (N = 260,223) SNPs were obtained from T1dbase. Their chromosomal coordinates and genic positions were retrieved from the Ensembl genome browser by means of the Biomart tool.

3.3 SNP Distribution in Genic Positions

The frequencies of associated and non-associated SNPs in various genic positions SNPs were determined. The results are presented as pie charts in Figures 7 and 8. Both appear to have a high frequency of occurrence in intronic, intergenic, upstream and downstream regions, and in non-coding transcripts. Remarkably, all these genic positions are non-coding structures, but it must also be noted that most of the genome (approximately 98%) is non-coding. What distinguishes the associated-SNPs from the non-associated SNPs is the fact that the associated-SNPs seem to occur twice as often in non-coding transcripts than the non-associated SNPs (Figures 7 and 8).

It should be taken into account that one and the same SNP may be located within more than one genic position and also in more than one type of genic position. This occurs if the SNP is within a genomic region that gives rise to more than one gene (gene-overlap) or to multiple splice isoforms of a single gene (transcript-overlap). Gene overlap occurs if overlaying genes share the same DNA sequence (i.e. in a different reading frame or on the opposite DNA strand) and yet do not share regulatory elements or any exons (Gerstein, 2007). Multiple splice isoforms are generated by alternative splicing. This mechanism entails the differential removal of introns from a primary RNA into a variety of possible mature mRNAs. Multiple transcripts overlaying a SNP locus can either be splice variants of only one gene (Figure 9), or splice variants of two or more overlapping genes.

Many nucleotides in the T1D susceptibility regions are associated with multiple transcripts of one or more genes. On average, a SNP occurs in 7.8 transcripts. Consequently, many T1D associated and non-associated SNPs affect multiple transcripts. Each SNP occur within the same type of genic position in all the affected transcripts or could be in a different genic positions in each overlapping transcript. 71% of the SNPs are found in multiple transcripts. Approximately 43% of these, which include most of the associated SNPs, occur not only in more than one genic position, but also in different types genic positions. Figure 10 depicts an actual example that shows how the associated-SNP 'rs281417' occurs in five different genic positions in transcripts of two genes. The SNP affects an exon and 5'UTR region in ZGLP1 transcripts, and an intron, NMD transcript and non-coding transcript in FDX1L.

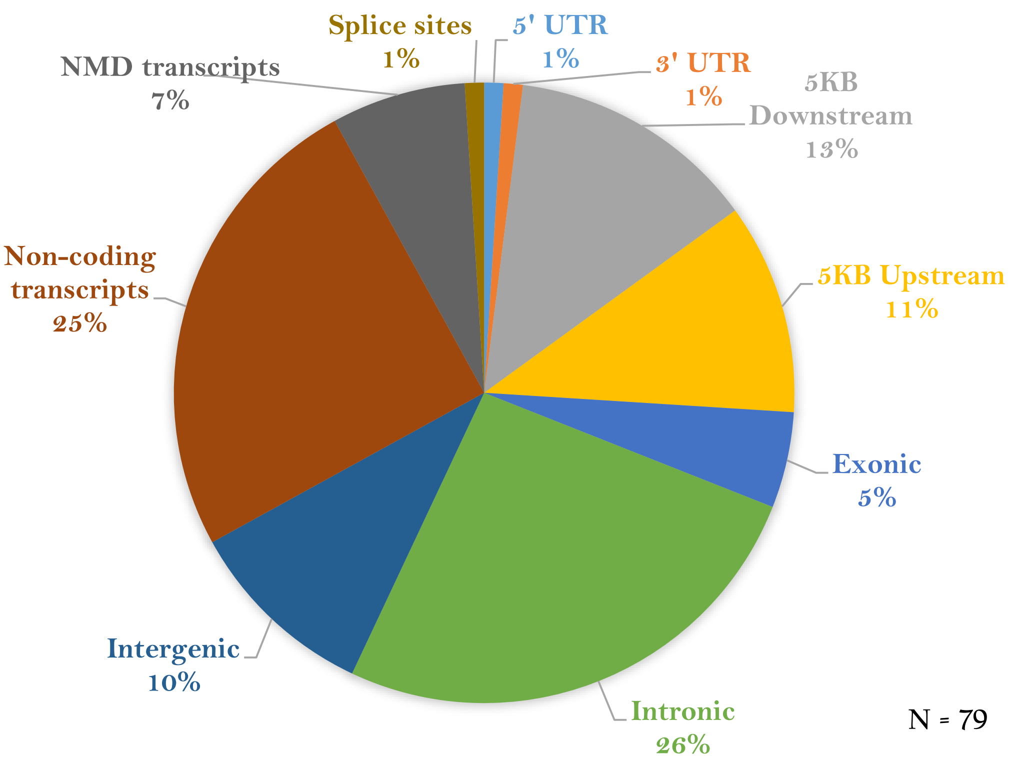


Figure 7. A pie chart showing the distribution of associated T1D-SNPs in the various genic positions

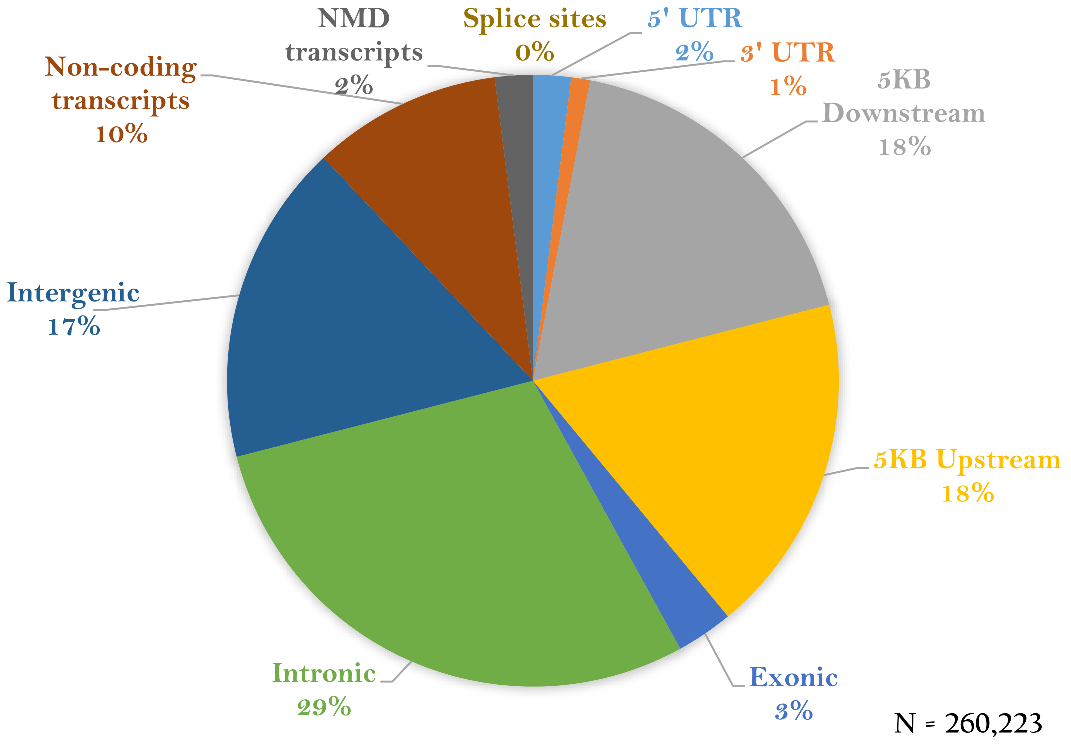


Figure 8. A pie chart displaying the distribution of non-associated T1D-SNPs in the various genic positions

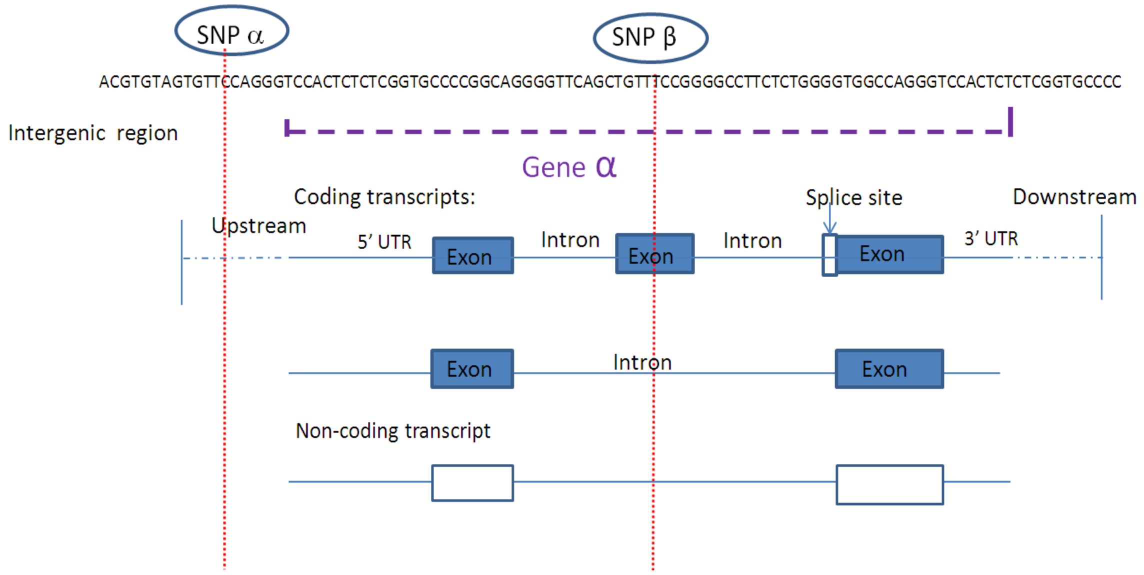


Figure 9. An illustration depicting how a SNP may affect different structural parts in overlapping transcripts that overlay the SNP locus.

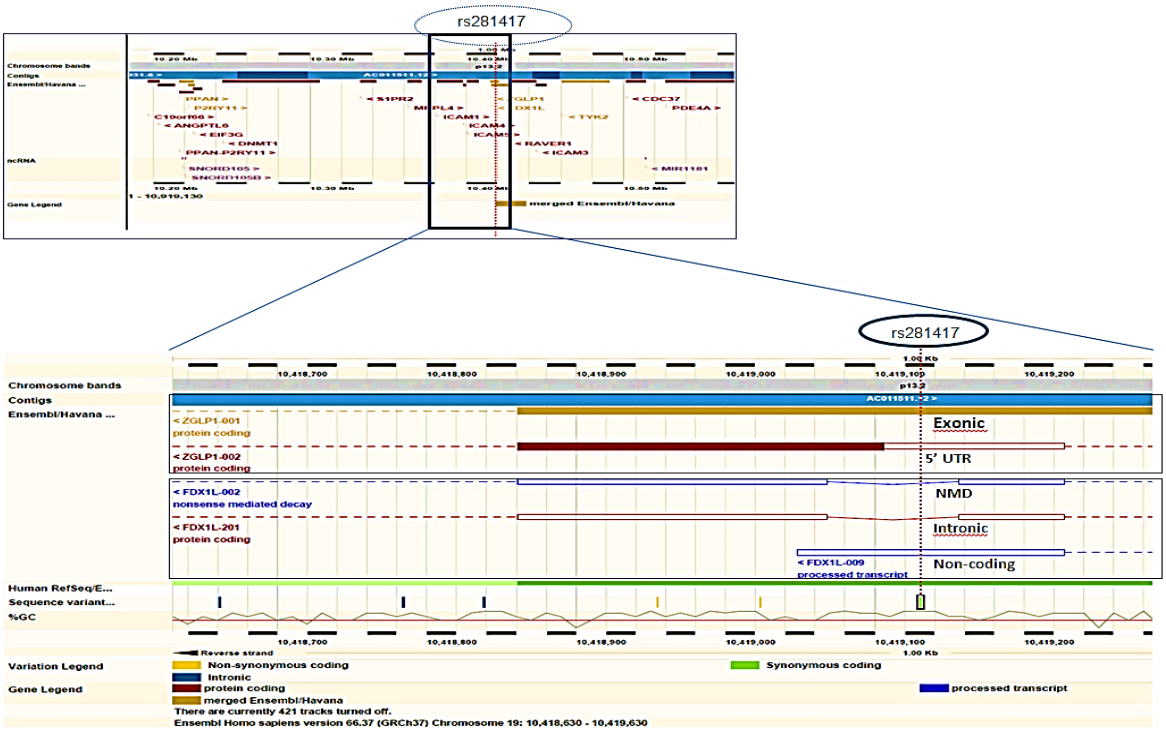


Figure 10. A genome browser illustration of the associated T1D-SNP rs281417 in overlapping transcripts of 2 overlapping genes (Source: Ensembl genome browser V63)

If a locus is characterised by **transcript overlap**, the alternative transcripts of the same gene may be read differently in the alternative splicing process. Each transcript may differ due to exclusion or inclusion of exons from processed mRNA (Sammeth et al., 2008), because of intron retention, or as a result of alternative start sites (Matlin, 2005; Black, 2003). In case of **gene overlap**, the same applies for alternative transcripts of each of the overlapping genes. Furthermore, the overlapping genes themselves may have different start sites (and are therefore read differently) or may be on opposite strands of the DNA molecule, and consequently are read in opposite directions to each other.

The current discovery and knowledge about extensive transcriptional activity within the human genome can be attributed to research by the ENCODE (**ENC**yclopedia **O**f **D**NA **E**lements) consortium (Becker et al., 2011). Previously, it was thought that a gene was mostly transcribed to a single mRNA transcript, which in turn is translated to a functional protein. However, an initial ENCODE study, which aimed to characterise 1% of the human genome (ENCODE, 2012; Birney et al., 2007), revealed that the human genome is much more pervasively transcribed than was previously thought. They found that most nucleotides in the genome are associated with at least seven alternative transcripts (Birney et al., 2007). This extensive transcriptional activity is also characteristic of loci within many of the T1D susceptibility regions, especially the HLA locus on chromosome six.

3.4 SNP Occurrence in Multiple Genic Positions

A SNP's ability to mutate multiple sites runs counter to the classic one-SNP-one-gene approach to disease studies and necessitates a deeper investigation into the genic properties of SNPs. Since many of the T1D-SNPs occur in more than one type of genic position in overlapping transcripts, Figures 7 and 8 may not be an entirely accurate representation because such these SNPs will be recounted in every category in which they occur. Thus the percentages of SNPs in some categories are likely to be inflated. This prompted me to look into more detail at the number and types of genic positions occupied by SNPs and the type of overlap occurs in the region surrounding the SNP.

The "Genic position count" refers to the number of unique genic positions in which a SNP occurs in overlapping transcripts. If a SNP is in the same genic part in more than one transcript, it is counted as one occurrence. An example is illustrated in Figure 11 for the associated-SNP 'rs2476601'. This SNP maps to the 1p13.2 T1D locus and affects eleven alternative transcripts of the PTPN22 gene. However, it does not occur in one and the same but in four distinct genic parts (exon, intron, NMD transcript and non-coding transcript). Therefore, SNP 'rs2476601' has a unique genic position count of 4 (Table 4). This analysis was carried out for each T1D-SNP (*see Appendix A* {*Table* 7}).

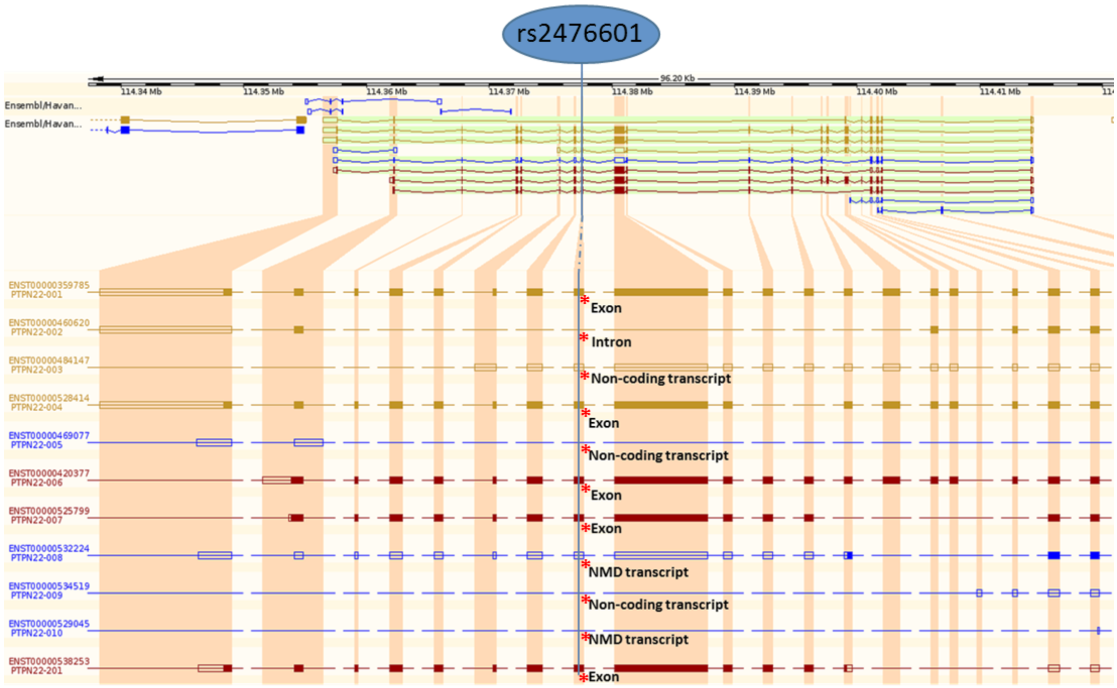

rs2476601
96.20 Kb
114.34 Mb
114.35 Mb
114.36 Mb
114.37 Mb
114.38 Mb
114.39 Mb
114.40 Mb
114.41 Mb
Ensembl/Havan...
Ensembl/Havan...
ENST00000359785 PTPN22-001
ENST00000460620 PTPN22-002
ENST00000484147 PTPN22-003
ENST00000528414 PTPN22-004
ENST00000469077 PTPN22-005
ENST00000420377 PTPN22-006
ENST00000525799 PTPN22-007
ENST00000532224 PTPN22-008
ENST00000534519 PTPN22-009
ENST00000529045 PTPN22-010
ENST00000538253 PTPN22-201
*Exon
*Intron
*Non-coding transcript
*Exon
*Non-coding transcript
*Exon
*Exon
*NMD transcript
*Non-coding transcript
*NMD transcript
*Exon

Figure 11. The associated SNP rs2476601 affects different genic positions in splice isoforms of the PTPN22 gene. (Source: Ensembl genome browser V63)

Table 4. Genic positions of SNP rs2476601 in transcripts of the PTPN gene retrieved from Ensembl.

Gene Name	Transcript name	Transcript biotype	SNP Alleles	Genic position
PTPN22	PTPN22-001	Protein coding	G (C)	Exon
PTPN22	PTPN22-002	Protein coding	G (C)	Intron
PTPN22	PTPN22-003	Retained intron	G (C)	Non-coding transcript
PTPN22	PTPN22-004	Protein coding	G (C)	Exon
PTPN22	PTPN22-005	Non-coding transcript	G (C)	Non-coding transcript
PTPN22	PTPN22-006	Protein coding	G (C)	Exon
PTPN22	PTPN22-007	Protein coding	G (C)	Exon
PTPN22	PTPN22-008	Nonsense mediated decay	G (C)	NMD transcript
PTPN22	PTPN22-009	Non-coding transcript	G (C)	Non-coding transcript
PTPN22	PTPN22-010	Nonsense mediated decay	G (C)	NMD transcript
PTPN22	PTPN22-201	Protein coding	G (C)	Exon

Associated SNPs have a significantly higher genic position counts than non-associated SNPs (Kolmogorov-Smirnov test, D = 0.29, Dcrit = 0.15 for α = 0.05%) (Figures 12 and 13).

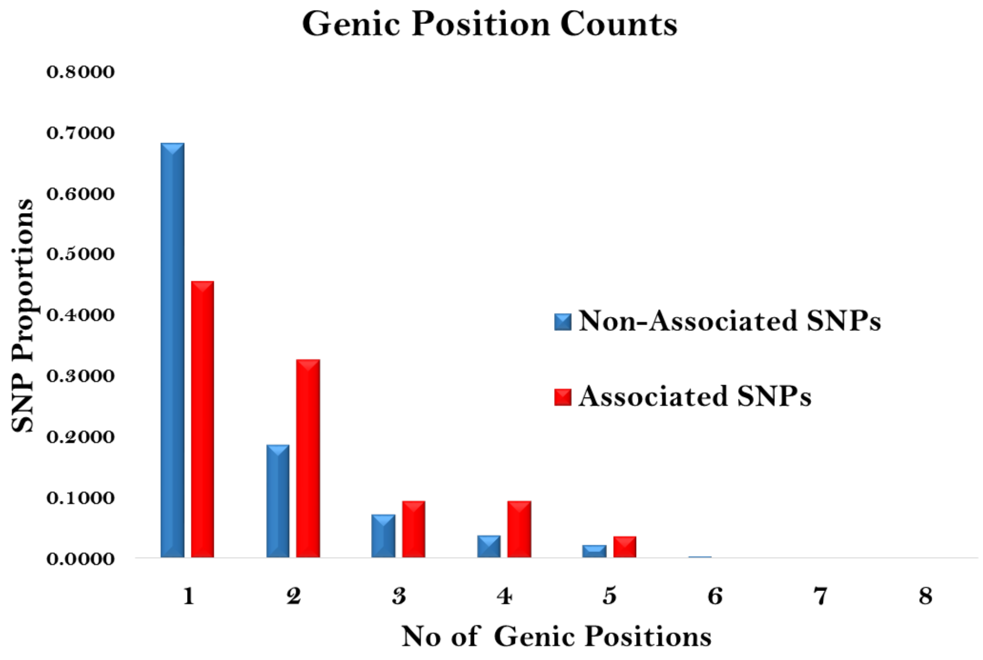


Figure 12. A histogram and a cumulative frequency plot of unique genic position counts for associated-SNPs and non-associated-SNPs. The histogram shows that higher proportions of associated-SNPs (red) affect multiple genic positions than the non-associated-SNPs (blue). The largest proportion of the non-associated-SNPs are at a genic position count of one.

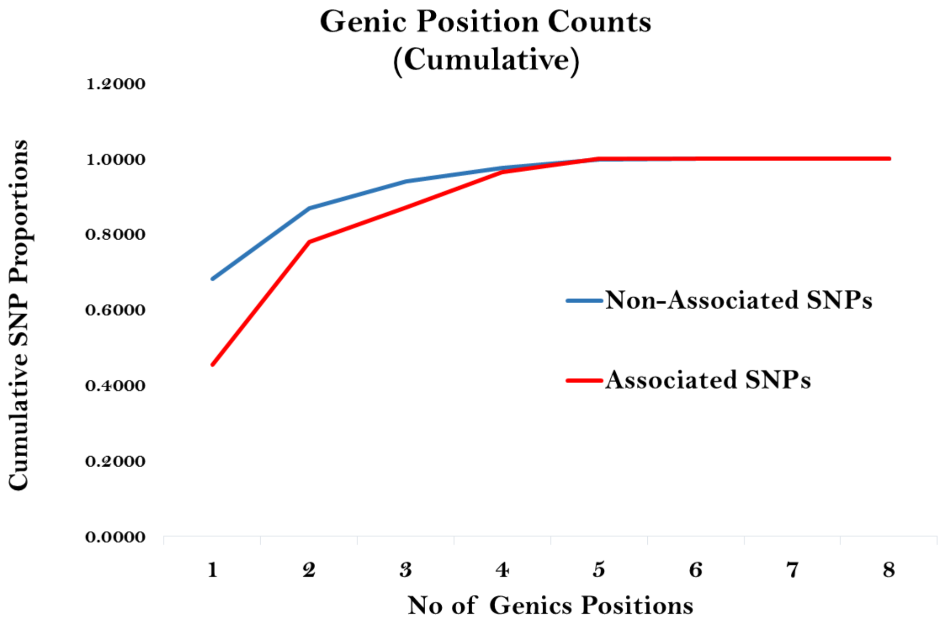


Figure 13. The cumulative frequency plot indicates that the genic position count of one has the largest difference in proportions between both SNP sets. . The computed D value (0.23) is higher than the critical value (0.15) indicating a significant difference at **α**= 0.05.

3.5 Type of Transcript overlap at SNP Locus

To identify the type of overlap (of transcripts) of T1D SNPs, four types of overlaps were considered:

Gene & Transcript overlap: in this type of overlap, the overlaying transcripts are variants of more than one gene. The overlapping genes may be all coding or all non-coding genes (described in chapter 2), or a combination of both.

Transcript overlap: in this type of overlap, the overlaying transcripts are variants of the same gene. The gene maybe a coding gene or a non-coding RNA gene. The coding gene may have both coding and non-coding transcripts (described in chapter 2), whilst the noncoding gene will have only non-coding transcripts.

Gene flanking overlap: SNPs in this type of overlap include those solely in upstream and downstream positions of multiple transcripts, which are regions flanking genes.

Single Genic Position: This refers to SNPs that are within a gene that has only one transcript or that are in inter-genic positions.

The chi-square (χ^2) statistic was applied to compare the frequencies between groups. The method tests if the distribution of observed frequencies deviate from what would be expected by chance (i.e. calculated expected frequencies). This is a non-parametric statistic, and has been chosen because the data to be analysed are nominal data. They are also discrete occurrences that are assumed to occur independently of each other. Generally, it is the preferred method for analysing nominal data (McDonald, 2015).

The test indicates a significant association between type of SNPs (associated and non-associated SNPs) and type of overlap ($\chi^2 = 13.25$, df = 3, p = 0.004, $\alpha = 0.05$) (Figure 14). The standardised residuals (Table 5) show that associated SNPs are under-represented in gene-flanking overlaps.

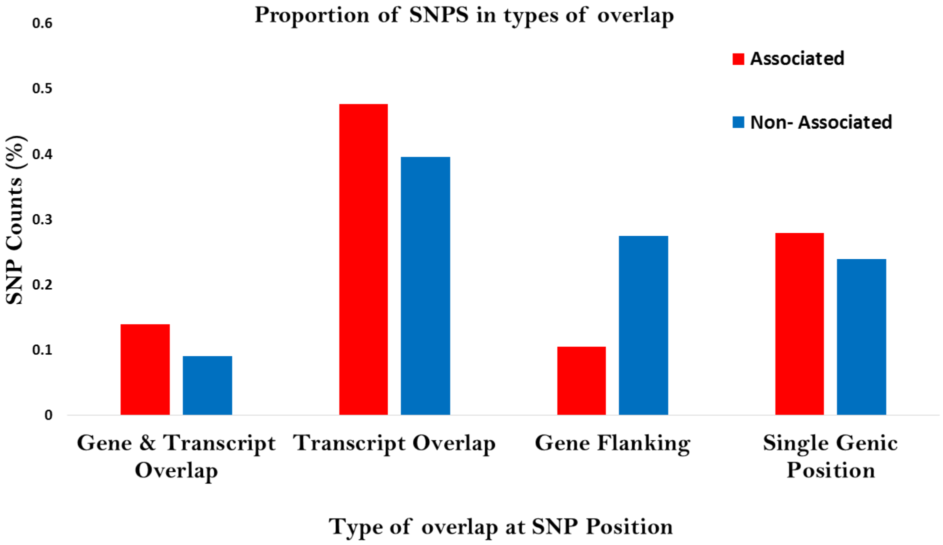


Figure 14. A Bar plot of SNP proportions in types of overlap, indicating that there is a much lower proportion of associated SNPs than non-associated SNPs in gene flanking (upstream and downstream) regions of transcripts.

Table 5. Standard residual values indicating differences between SNP counts in types of overlap. The table has been colour coded to highlight the trend in residual values. The sharp colour contrast (bright yellow) in the cell representing associated SNPs in gene flank overlap indicates that these SNPs are much less in flanking regions than would be expected by chance.

Standardized Residuals	Associated	Non- Associated
Gene & Transcript Overlap	1.50	-0.03
Transcript Overlap	1.19	-0.02
Single Genic Position	0.75	-0.01
Gene Flanking	-3.00	0.05

3.6 SNP Genic Profiles

The genic profile of a SNP is a list of the identified types of genic positions in which it occurs. The components of each SNP profile were the name/names of unique genic positions in which the SNP occurs. In other words “SNP genic profiles” refine the “genic position counts” of each SNP by giving names to numbers. For example, the genic profile of SNP rs281417 is {Exonic/5’UTR/NMD/Intronic/Non-coding} (Figure 15). 286 unique genic profiles types were identified. The profiles size ranged from one to eight components, in accordance with the unique genic position counts. A complete list of created profiles and SNP counts is presented in the Appendix A. Figure 16 shows the counts of associated-SNPs (red bars) and non-associated-

SNPs (blue bars) belonging to the 286 identified genic profiles, expressed as percentages. Due to large size and software limitation, the figure does not properly capture the names of all the 286 profiles even though they are all included in the plot. However, the figure gives a good visual overview of profile sizes. The non-associated-SNPs are largely grouped into the 'one component' genic profile category while the associated-SNPs are more spread out.

To identify genic profiles common to associated-SNPs and non-associated-SNPs, I plotted the frequency (%) of associated SNPs (red) in each genic profile against that of non-associated SNPs (blue). In the plot (Figure 17), genic profiles for which the percentages of associated SNPs and non-associated SNPs are equal, fall along the diagonal line. Profiles with a much higher occurrence in one set than in the other will show a marked deviation from the diagonal. Genic profiles typical for the associated SNPs fall above the diagonal, those below the diagonal are characteristic for the non-associated SNPs. The margin for choosing over-represented profiles was set at 3%. The over- and under-represented genic profiles are listed in Table 6.

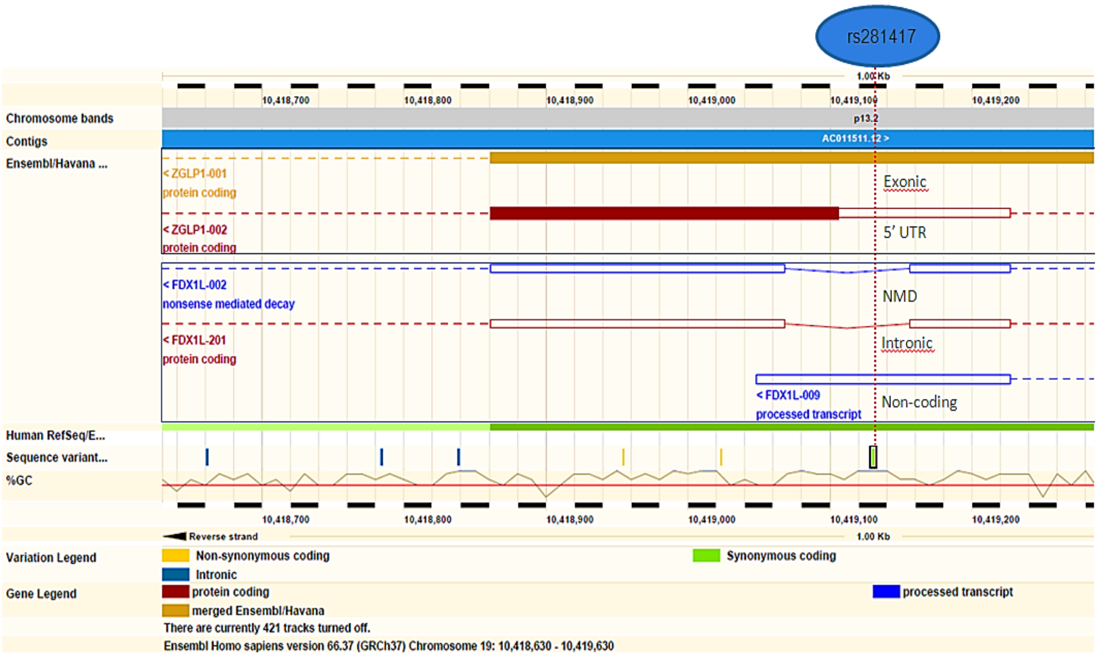


Figure 15. Genic profile {Exonic/5'UTR/NMD/Intronic/Non-coding} of SNP rs281417 which is overlapped by different genic parts of overlapping gene transcripts. (source: Ensembl genome browser)

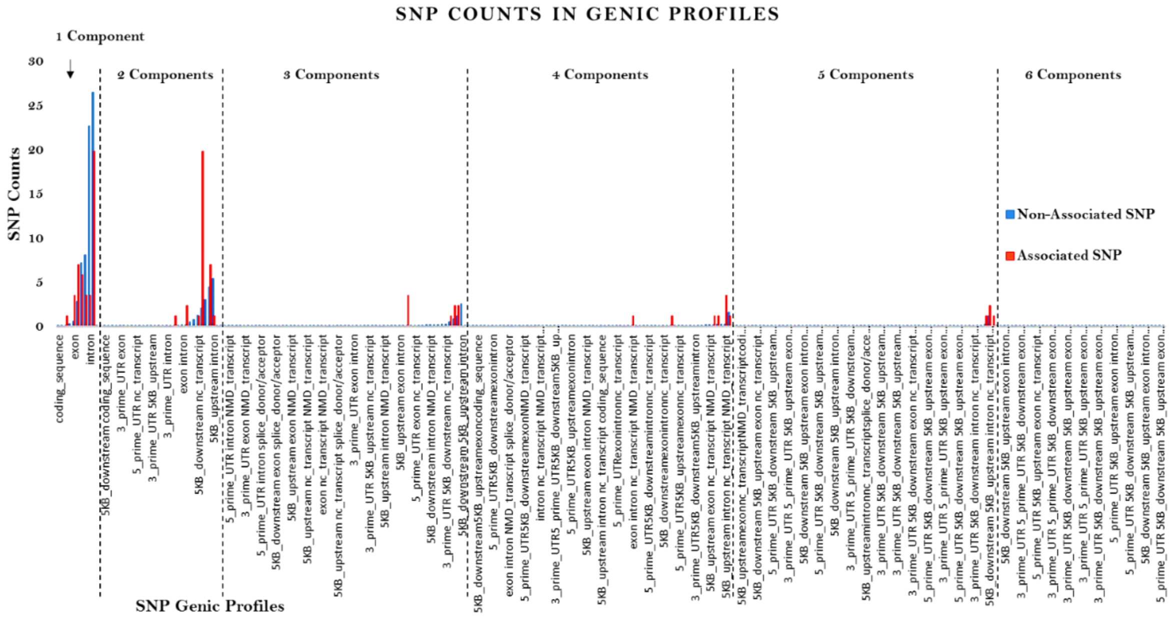


Figure 16. Counts of associated-SNPs (red) and non-associated-SNPs ((blue) belonging to the 286 identified genic profiles, expressed as percentages.

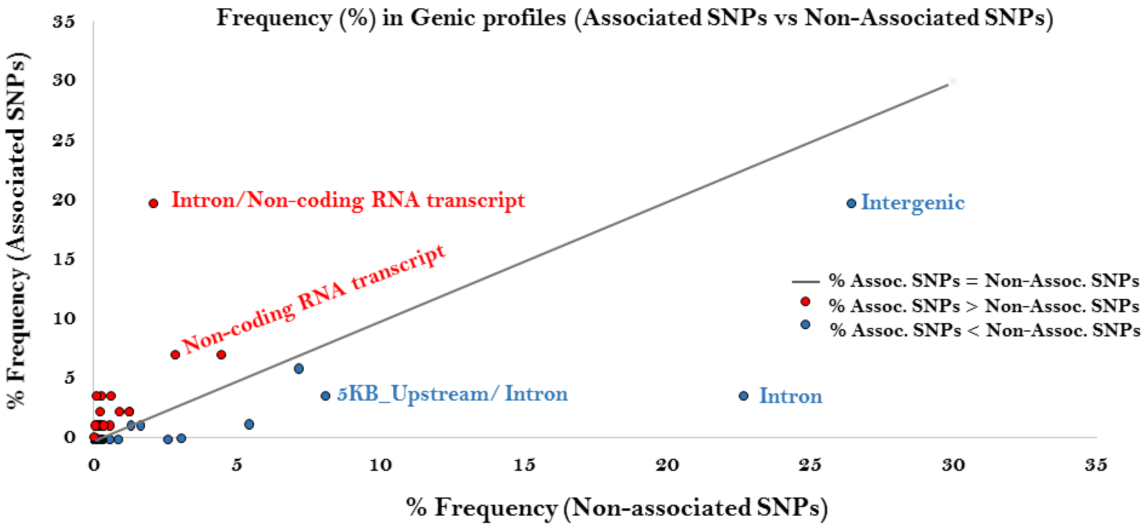


Figure 17. A plot of frequency of associated SNPs (red) in each genic profile against non-associated SNPs (blue) in order to identify genic profiles common to associated-SNPs and non-associated-SNPs.

The most common genic profile in both SNP groups is “Intergenic”. The profile over-representing for associated SNPs is “***intron nc_transcript***”, whereas non-associated SNPs typically have an “***intron***” genic profile but these are not in overlap with non-coding transcripts as is for the associated SNPs. The non-associated SNPs are also more in gene flanking parts as indicated by the third and fourth most represented profiles, “***5KB_downstream***” and “***5KB_upstream intron***.” Table 6 gives a better and more reliable representation of the genic positions of T1D SNPs than the pie charts in Figures 7 and 8, recurring genic positions are not over-counted.

Table 6. The genic profiles over-representing for associated SNPs (red) and the non-associated SNPs (blue).

SN	**Genic Profiles**	**No of Profile**	**SNP Counts (%)**	
		Components	**Assoc**	**Non_Assoc**
1	intron nc_transcript	2	19.767	2.069
2	nc_transcript	1	6.977	2.842
3	intron nc_transcript NMD_transcript	3	3.488	0.070
4	5KB_upstream intron nc_transcript NMD_transcript	4	3.488	0.236
5	exon	1	3.488	0.581
6	5KB_downstream intron	2	6.977	4.469
7	5KB_upstream	1	5.814	7.156
8	5KB_downstream 5KB_upstream	2	0.000	3.044
9	5KB_upstream intron	2	1.163	5.421
10	5KB_downstream	1	3.488	8.075
11	intergenic	1	19.767	26.479
12	intron	1	3.488	22.685

3.7 Chapter Summary

This study revealed that one and the same SNP may affect multiple processes because it occurs in more than one transcript or overlapping genes. Furthermore, these transcripts may have different functions. The analysis also showed that associated SNPs are found more often in non-coding than in coding parts the genome, (especially in introns overlapping non-coding RNA transcripts). Although not specifically tied to non-coding transcripts, non-associated SNPs are also frequently occurring in introns. This is important because introns and certain non-coding RNA transcripts (like microRNAs and lincRNAs) may be involved in regulatory activity.

It is now widely recognized that most complex-disease-associated SNPs map to non-protein coding regions (Dirk et al., 2014; Zhang et al., 2014; Djebali et al., 2012) either within genes or outside genes. Yet, discovering the effects of non-coding variants is a challenge, especially because there is a wide variety of non-coding functions. Furthermore, annotation for human regulatory elements is incomplete, so there are still potentially unknown mechanisms of regulation in the genome (Ward and Kellis, 2012). Nevertheless, post-GWA studies have recently demonstrated how disease risk variants can affect non-coding functions (Zhang, et al., 2014). For instance, microRNA can be negatively affected by a rare mutation that changes its sequence or modifies its complementary target sequence in the 3’ UTR region of its target mRNA transcript (Bartel, 2009). The latter has been demonstrated for the Crohn’s disease (an autoimmune disease that has markers in T1D susceptibility regions) associated SNP, rs10065172. This mutation attenuates binding of miRNA-196 to the mRNA transcript of the IRGM gene (Brest et al., 2011; Singh et al., 2006). The resulting fluctuation in IRGM expression leads to an increase in intracellular bacteria which can lead to a Crohn's disease associated inflammation (Brest et al., 2011; Singh et al., 2006). Also, in lincRNAs, a risk variant can be detrimental by altering the tertiary structure of the lincRNA transcript (Shen, et al., 1999). The highly conserved structure of lincRNA is important in guiding recruitment of chromatin regulators to the chromatin (Tsai et al., 2010; Rinn et al., 2007). The risk allele of SNP rs35955962 maps to the MIAT (Myocardial Infarction Associated Transcript) lncRNA (Broadbent et al., 2008). This variant has been found to affect the transcript by increasing it affinity for nuclear proteins compared to the non-risk allele (Broadbent et al., 2008; Ishii et al., 2006). Though the particular influenced protein, and its functional consequence on heart disease, is still yet to be characterised (Zhang et al., 2014).

SNPs that affect multiple processes have also been recognised. A prominent example is the impact of the rs1045642 SNP on different functional parts. The SNP maps to an exon in the multidrug resistant gene MDR1 (Hoffmeyer et al., 2000). The SNP is synonymous, which means it does not change the amino acid sequence (primary structure) of the protein (MDR1) the gene is building (Kimchi-Sarfaty et al., 2007). However, it alters the drug substrate specificity of the protein. It is suggested that the SNP slows down the rate of translation of the MDR1 mRNA, which in turn impacts protein folding (Komar, 2007). This altered MDR1 conformation decreases the drug substrate specificity of the protein (Fung and Gottesman, 2009; Kimchi-Sarfaty et al., 2007; Hoffmeyer et al., 2000). Recently, it has also been shown that a part of the coding sequence of the exon in MDR1 not only specifies an amino acid, but a transcription factor binding site (Stergachis et al., 2013). This provides an additional avenue through which the SNP may impart another functional effect. These sporadic references raise the impression that one SNP affecting multiple processes is a rather extraordinary event. This research shows that it might be a quite common but an overlooked phenomenon that is characteristic of complex diseases.

The associated and non-associated SNPs have been successfully characterised by their genic positions in the T1D regions. Yet, the layers of genetic information embedded in the DNA that forms the susceptibility regions can also throw more light on the unique characteristics of the regions themselves. Therefore, in the following chapter, the genomic composition of the susceptibility regions including structural and functional parts are analysed. This is done in order to identify unique features characterising each region.

CHAPTER 4

CLASSIFICATION OF T1D SUSCEPTIBILITY REGIONS

In the previous chapter, it was shown that associated and non-associated SNPs can be characterised and distinguished by their genomic location within T1D susceptibility regions. It was also established that disease associated SNPs occur quiet frequently in non-coding parts. Led on by this finding, an understanding of the genomic make-up of the susceptibility regions themselves became the next objective of my research. To do this, the genomic composition of the susceptibility regions and other associated genetic features were identified and analysed. In this study I found that the T1D susceptibility regions can be grouped into three clusters reflecting genomic content. The clusters are mainly separated by differences in intronic content and gene density. Furthermore, there are twenty-five T1D regions carry markers for fourteen other autoimmune diseases. The study revealed that the cluster of regions characterised by the most relative gene density and counts of non-coding transcript nucleotides than others, also had the strongest degree of susceptibility region sharing with other diseases.

4.1 INTRODUCTION

Forty-nine genomic regions that confer susceptibility to T1D have been identified by genome wide association studies (GWAS) (Burren et al., 2011; Barett et al., 2009; Burton et al., 2007). Association studies typically identify the specific locations of genetic variants (mutations such as SNPs, insertions and deletions) that correlate with the phenotype of the disease. Identification of these loci is often followed up by intricate quantitative and statistical models to define disease risk patterns (Bush and Moore, 2012). Although relevant information has been generated from these GWAS studies, the aetiology of many complex diseases still remain unknown (Dirk et al., 2014; Noble and Erlich, 2012). Understanding genomic aspects of disease, such as the discovery of relevant gene regulatory pathways and biochemical pathways for drug targets (Lander, 2013; Collins, 2010), can revolutionize medical practice (Ward, 2013). Therefore, it is important to build a clear picture of the genomic makeup of these susceptibility regions and the special features that describe them. Characterisation of a disease regions have been previously done in medical genomics to understand the genetic mechanism underlying complex diseases like Coeliac disease (Hrdlickova et al., 2011), Ovarian cancer (Permuth-Wey et al., 2013), and sex-related diseases (Handel et al., 2013). These studies reveal that GWAS findings provide good starting points towards identifying the disease-associated variants and genes, but also that bioinformatics approaches are needed help pinpoint the true causal variants. Many studies have been conducted on T1D, but to my knowledge characterising the T1D susceptibility regions on the basis of structural genomic content has not been reported.

The aim of this study is to find out if the T1D susceptibility regions differ strikingly in genomic content among each other, and also if such eventual differences are related to the presence of loci associated with other autoimmune diseases. More specifically, it would be pertinent to know if certain regions have higher proportions of non-coding (intronic DNA and non-coding RNA) material than others. Because intronic DNA and many functional non-coding RNA sequences are involved in some form of regulatory activity (Djebali et al., 2012), querying the content of non-coding DNA in the susceptibility regions is in line with the main aim of this research (the role of disturbed regulation in the occurrence of T1D). The genic profiles of the associated T1D-SNPs identified in the previous chapter also indicate that many of them are within non-coding sequences. Indeed, other studies have also shown that most genetic risk variants fall outside of coding sequences (Zhang et al., 2014; Encode, 2012; Frazer, 2009). The second aim was to establish if certain regions have a higher gene and SNP density than others. Gene density is particularly important from two regulatory perspectives. First, gene dense regions are expected to contain regulatory modules with binding sites that are involved in the activation or repression of gene transcription within the region. Secondly, gene dense regions may carry certain genes that are also involved in regulation of other genes via gene regulatory pathways. Although SNPs in binding sites are the main focus of this research, also SNPs in important non-coding regulatory sequences (but not in binding sites) or coding sequences for regulatory proteins may implicate the corresponding gene in the manifestation of disease (Zhang et al., 2014; Laurila and Lähdesmäki, 2009). Finally, the third question to be addressed by this study is: does the genomic composition of susceptibility regions relate to the presence of loci associated with other autoimmune diseases (Welter et al., 2014; Hindorff et al., 2013)? Susceptibility regions that harbour more disease associated variants could be more likely to harbour other trait-associated SNPs that are not detected by GWA studies (Lim et al., 2014; Pierce and Ashan, 2011).

4.2 METHODS

Susceptibility regions can be characterised by genomic features like total region size, the amount of coding- and non-coding DNA, and the number of genes and SNPs they carry. The number of nucleotides of a region that build up exons (coding sequences), introns, 5'UTR, 3' UTR and intergenic parts will be referred to as structural features. Functional features will include the abundance (density) of SNPs, coding genes, regulatory modules and non-coding (putative regulatory) RNA transcripts (Table 7).

Table 7. Structural and functional features used for characterising the T1D susceptibility regions.

Type	Feature
	Number of:
Structural	3'UTR nucleotides
	5' UTR nucleotides
	Exonic nucleotides
	Intronic nucleotides
	Intergenic nucleotides
Functional	
	Non-coding RNA nucleotides
	Regulatory module nucleotides
	Protein coding genes
	RNA genes
	SNPs

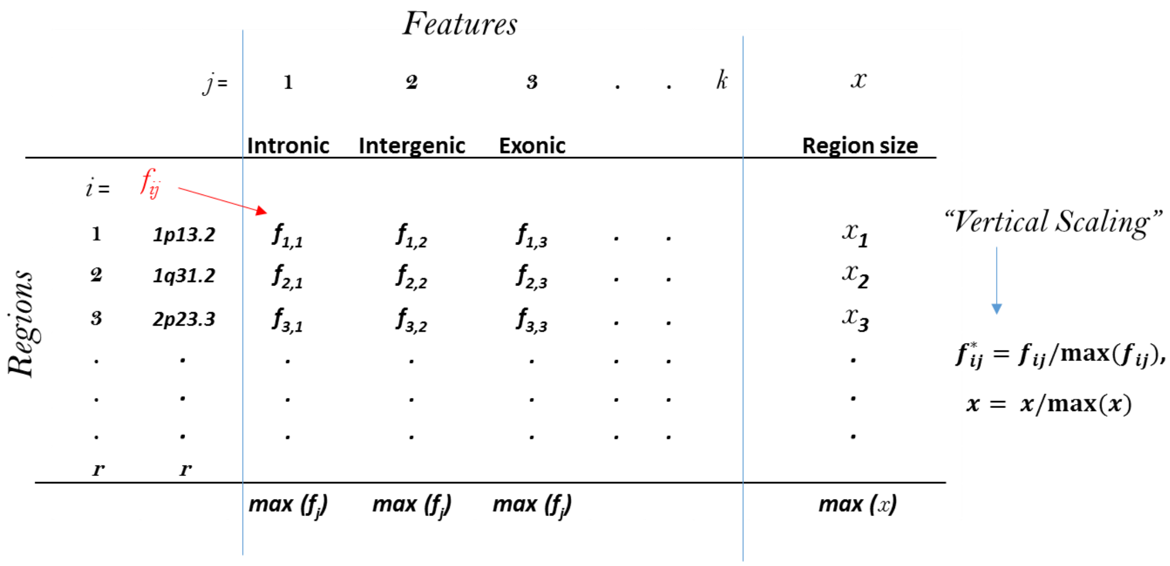
4.3 Data Normalisation

The T1D susceptibility regions vary widely in size, ranging from 45,078 bps in region 14q32.2 (Chr14:101283661-101328739) to 3,808,585 bps in the HLA region (chr6:29690000-33498585) (Burren et al., 2011). In order to eliminate bias due to size, the data should be normalised. Ideally, this can be dealt with by expressing the features as proportions of the total region size. This would scale the sizes of features per region to values between zero and one, and add them up to one. However, this cannot be applied to counts of functional features because some counts are very small in comparison to the total susceptibility region size. Therefore, for this study, data normalisation was achieved by a two-step vertical and horizontal scaling.

Feature-wise (vertical) scaling

The objects to be clustered are a data set consisting of the structural features sizes for each susceptibility region. The columns are the structural features whilst the rows are the susceptibility regions. In vertical scaling, the abundance for each structural feature ($\boldsymbol{f_{ij}}$) are normalised separately, by expressing the value of that feature j as a proportion of the maximum ($\boldsymbol{max(f_j)}$) of all susceptibility regions (Figure 18). Vertical scaling yields values ($\boldsymbol{f^*_{ij}}$) that are more comparable between regions but does not eliminate region size bias in the data.

$$\boldsymbol{f^*_{ij} = f_{ij} / max\,(f_j)}$$



		Features						
	$j =$	1	2	3	.	.	k	x
		Intronic	Intergenic	Exonic				Region size
$i =$	f_{ij}							
1	1p13.2	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	.	.		x_1
2	1q31.2	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	.	.		x_2
3	2p23.3	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	.	.		x_3
.	.	.	.	.	.	.		.
.	.	.	.	.	.	.		.
.	.	.	.	.	.	.		.
r	r							
		max (f_j)	max (f_j)	max (f_j)				max (x)

$$f^*_{ij} = f_{ij}/\max(f_{ij}),$$
$$x = x/\max(x)$$

Figure 18. Vertical scaling showing how feature sizes are expressed as proportions of the maximum feature size

Region-wise (horizontal) scaling

Horizontal scaling is applied to the data to eliminate the effect of region size. This is done because the size range of the susceptibility regions is large (3,763,507). The differences in the sizes of regions will have an inherent influence on the amount of each structural feature, and possibly the functional features, characterising each region. A cluster analysis simply done on the $\boldsymbol{f^*_{ij}}$ values would produce results reflecting the influence of region size bias as regions of similar size would simply come together when clustered. Thus, to normalise the data and correct for this problem, polynomial fitting was applied to the vertically scaled observed feature values $\boldsymbol{f^*_{ij}}$.

To do this, $\boldsymbol{f^*_{ij}}$ is measured as a residual from a polynomial regression model. This is done for each feature separately (i.e. Intron, Intergenic, Exon etc.) (Figure 19). $\boldsymbol{f^*_{ij}}$ is plotted against the vertically scaled susceptibility region size $\boldsymbol{x_j}$, then a 2nd order polynomial is used to fit data for features $\hat{\boldsymbol{f}}$ ($\boldsymbol{f_{estimated}}$) (Figure 19). Subsequently, residual values are calculated by subtraction of the expected values $\hat{\boldsymbol{f}}$ from the scaled observed values $\boldsymbol{f^*_{ij}}$. These residual values are devoid of region size bias and can used for row-wise clustering of features. Polynomial regression has been chosen for data fitting in this work because it produces the best results for the data. This method, in comparison to linear, logarithmic regression, gives the best line of fit with the highest R^2 values (Table 8). The graphical plots from the polynomial regression are presented in Figure 20 and in Appendix B.

$$\hat{\boldsymbol{f}} \; = \boldsymbol{a}\boldsymbol{x}_j^2 + \boldsymbol{b}\boldsymbol{x}_j + c; \text{ where, } \boldsymbol{x}_j = \textit{scaled region size}$$

$$Residual \ = \boldsymbol{f}^{*}_{\boldsymbol{ij}} \ - \ \hat{\boldsymbol{f}}_{\boldsymbol{ij}}$$

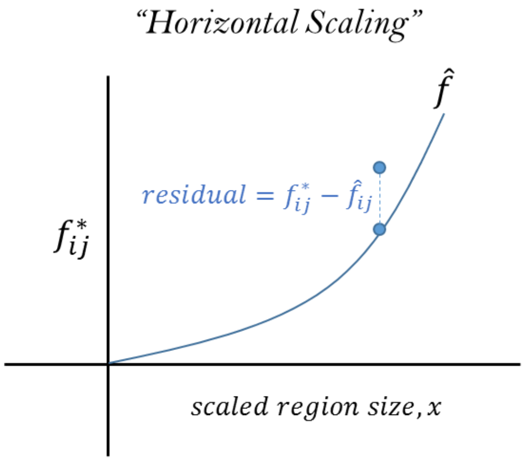

"Horizontal Scaling"
$\hat{f}$
$residual = f^{*}_{ij} - \hat{f}_{ij}$
f^{*}_{ij}
$scaled\ region\ size, x$

Figure 19. Horizontal scaling showing how vertically scaled feature sizes are expressed as residuals from a regression model

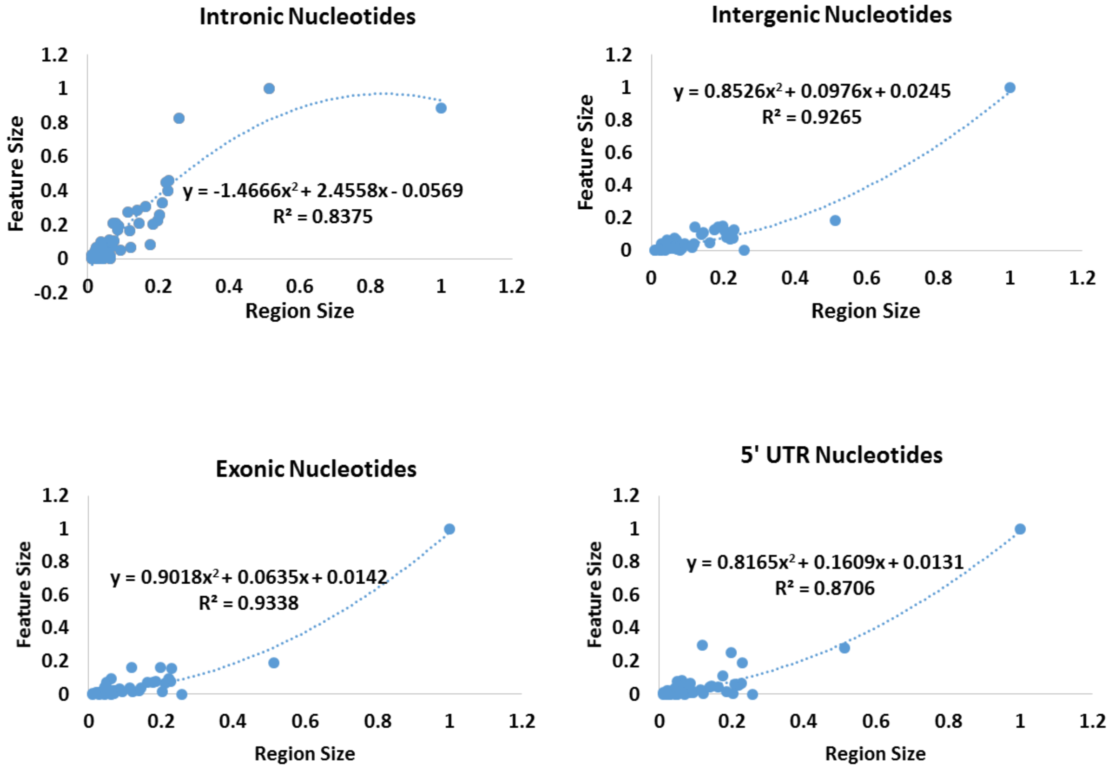


Figure 20. Scatter plots showing data–fitting using a second order polynomial regression for intronic, intergenic exonic, and 5'UTR data. The rest can be found in Appendix B.

Table 8. R2 values of structural and functional features from second order polynomial regression plots

Genomic Feature		**R values**		
Type	**Feature**	**Linear**	**Logarithmic**	**Polynomial**
Structural	Intronic Nucleotides	0.7127	0.6477	0.8375
	Intergenic Nucleotides	0.8232	0.3555	0.9265
	Exonic Nucleotides	0.8207	0.3336	0.9338
	5'UTR Nucleotides	0.7868	0.3346	0.8706
	3'UTR Nucleotides	0.8187	0.4015	0.8603
Functional	Non-Coding Transcripts	0.7800	0.5312	0.7879
	Non-Coding Gene Counts	0.8289	0.3179	0.9627
	Protein-Coding Gene Counts	0.7528	0.2740	0.9064
	SNP Counts	0.7607	0.2350	0.9776
	Regulatory Nucleotides	0.8045	0.6988	0.8572

4.4 Cluster Analysis of T1D susceptibility regions

In order to sort out similarities among susceptibility regions on the basis of their genomic structural content, the regions were subjected to a cluster analysis. There are two popular methods for data clustering, Hierarchical clustering and K-means clustering. The former is the preferred method for this analysis because it gives organization and structure within cluster sets, whereas the latter gives simple cluster sets with flat partitioning i.e. no particular organization or structure.

The cluster criterion used for the hierarchical clustering is the Ward's method, and the distance measure used is the Euclidean distance between regions. Wards method (Murtagh and Legendre, 2014) is an agglomerative cluster method that uses a bottom up approach to group small clusters into larger ones. This reduces the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged. For this analysis, Ward's method is preferred to other existing methods (i.e. single linkage and complete linkage methods). This is because the data to which the cluster algorithm will be applied is (normalised base pair counts) are not overly complicated. In addition, Ward's method produces better compactness (or clustering) and balance for this data than other methods used in this work, like the single linkage and complete linkage methods.

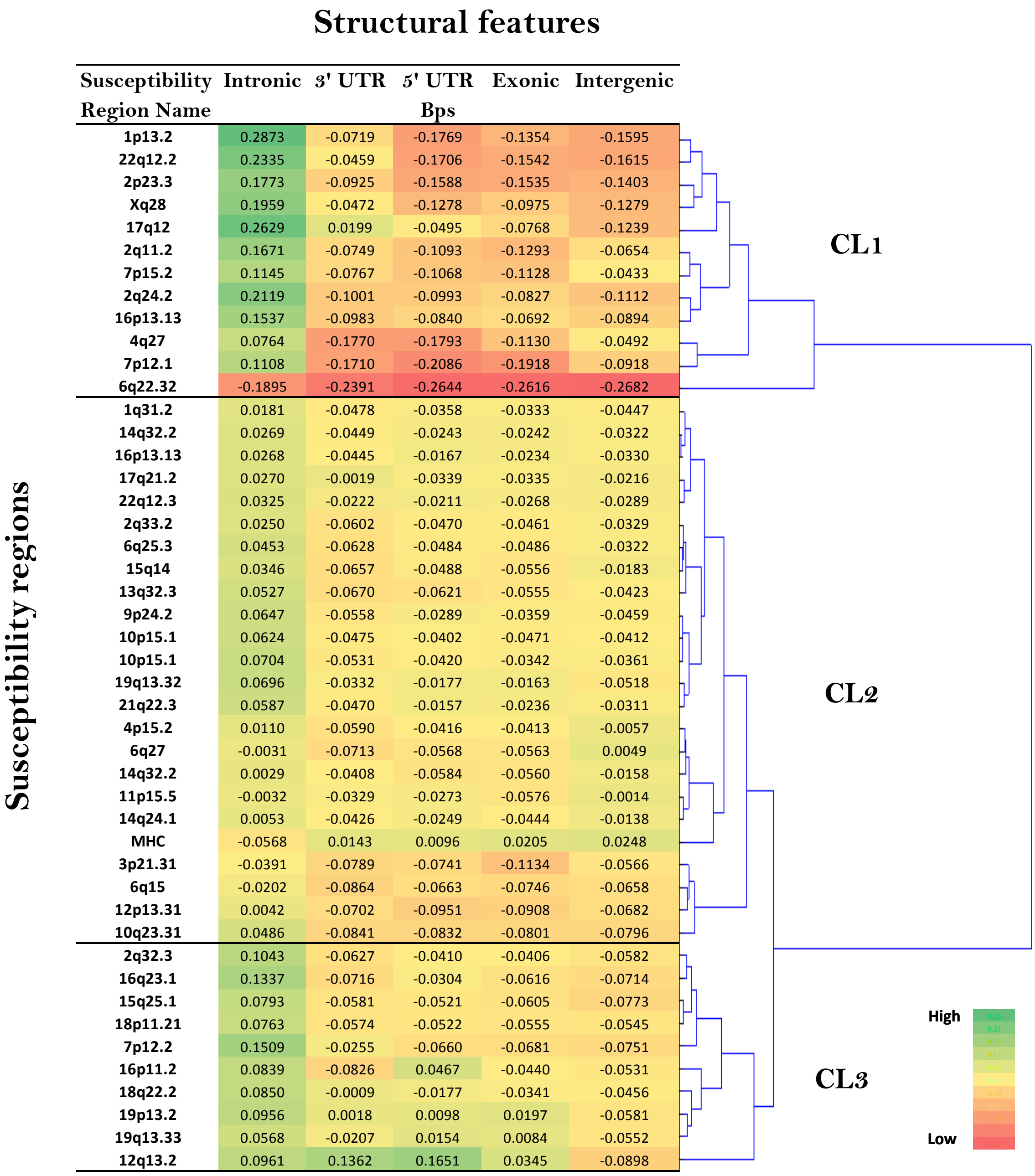
The set of normalised values computed for each of the five structural features of the T1D regions (*Appendix B* {*Tables 8 and 10*}) was subjected to clustering by the Ward's neighbour joining algorithm. This produced a dendrogram (Figure 21) that illustrates the clustering of regions into the most closely related groups (Mount, 2004) on the bases of genomic structure.

Three well generalized clusters can be distinguished from the data. The first cluster, **CL1**, is mainly separated from the other two by low exonic, UTR and intergenic content. Mean normalised value for each feature per cluster are presented in Table 9. Also shown are ***p***-values indicating the difference in means between clusters. All three clusters have higher than expected amounts of intronic nucleotides (high intron content), however, **CL1** has the highest mean content of all three clusters. **CL2** and **CL3** are mainly separated by differences in intronic content (higher for **CL3**) and intergenic nucleotides (higher for **CL2**). These differences are statistically significant, and P-values for pairwise comparisons between clusters are shown in Table 10.

For the functional features, to find out if the structural attributes of the clustered regions associate with their functional features, the set of normalised values computed for functional features (*Appendix B {Tables 9 and 11}*) are juxtaposed to the dendrogram formed on structural attributes (Figure 22). For each cluster, the mean normalised value per functional feature is calculated. Subsequently, the differences between cluster means the tested for significance (Table 11 and 12). All three clusters appear to contain considerable amounts of regulatory DNA and non-coding RNA with significant differences from **CL2** (Table 11). The highest mean content for regulatory DNA is in **CL1**. But the gene density and total SNP count of this cluster are both less than expected and also the lowest of all three clusters (Table 11). **CL3** differs from **CL1** in gene density, it has the highest gene density of all three clusters although the normalised values are also less than would be expected by chance. **CL2** has the highest total SNP count, which is especially visible in the HLA/MHC region, however the normalised value is also less than expected.

4.5 Unique attributes of T1D Clusters

The first question of this study is do certain regions have higher proportions of non-coding (intronic DNA and non-coding RNA) material than others? A closer look at the first cluster, **CL1,** reveals that it consists of regions with significantly high intronic DNA content (except the outlier, 6q22.32, see Figure 21). The regions are also characterised by low content of intergenic, exonic, and UTR nucleotides. The third cluster, **CL3,** is similar to **CL1** in containing regions with relatively high intronic DNA and low intergenic DNA content, but differs from **CL1** by having more exonic, and utr nucleotides. Both clusters are also characterised by high content of non-coding RNA, and these attributes in **CL1** and **CL3** are positive for the first question. This finding is interesting as intronic and non-coding RNA sequences are known to be involved in various regulatory processes within the genome (Djebali et al., 2012; Ward and Cooper, 2010; Khalil et al., 2009).



Susceptibility	Intronic	3' UTR	5' UTR	Exonic	Intergenic
Region Name			Bps		
1p13.2	0.2873	-0.0719	-0.1769	-0.1354	-0.1595
22q12.2	0.2335	-0.0459	-0.1706	-0.1542	-0.1615
2p23.3	0.1773	-0.0925	-0.1588	-0.1535	-0.1403
Xq28	0.1959	-0.0472	-0.1278	-0.0975	-0.1279
17q12	0.2629	0.0199	-0.0495	-0.0768	-0.1239
2q11.2	0.1671	-0.0749	-0.1093	-0.1293	-0.0654
7p15.2	0.1145	-0.0767	-0.1068	-0.1128	-0.0433
2q24.2	0.2119	-0.1001	-0.0993	-0.0827	-0.1112
16p13.13	0.1537	-0.0983	-0.0840	-0.0692	-0.0894
4q27	0.0764	-0.1770	-0.1793	-0.1130	-0.0492
7p12.1	0.1108	-0.1710	-0.2086	-0.1918	-0.0918
6q22.32	-0.1895	-0.2391	-0.2644	-0.2616	-0.2682
1q31.2	0.0181	-0.0478	-0.0358	-0.0333	-0.0447
14q32.2	0.0269	-0.0449	-0.0243	-0.0242	-0.0322
16p13.13	0.0268	-0.0445	-0.0167	-0.0234	-0.0330
17q21.2	0.0270	-0.0019	-0.0339	-0.0335	-0.0216
22q12.3	0.0325	-0.0222	-0.0211	-0.0268	-0.0289
2q33.2	0.0250	-0.0602	-0.0470	-0.0461	-0.0329
6q25.3	0.0453	-0.0628	-0.0484	-0.0486	-0.0322
15q14	0.0346	-0.0657	-0.0488	-0.0556	-0.0183
13q32.3	0.0527	-0.0670	-0.0621	-0.0555	-0.0423
9p24.2	0.0647	-0.0558	-0.0289	-0.0359	-0.0459
10p15.1	0.0624	-0.0475	-0.0402	-0.0471	-0.0412
10p15.1	0.0704	-0.0531	-0.0420	-0.0342	-0.0361
19q13.32	0.0696	-0.0332	-0.0177	-0.0163	-0.0518
21q22.3	0.0587	-0.0470	-0.0157	-0.0236	-0.0311
4p15.2	0.0110	-0.0590	-0.0416	-0.0413	-0.0057
6q27	-0.0031	-0.0713	-0.0568	-0.0563	0.0049
14q32.2	0.0029	-0.0408	-0.0584	-0.0560	-0.0158
11p15.5	-0.0032	-0.0329	-0.0273	-0.0576	-0.0014
14q24.1	0.0053	-0.0426	-0.0249	-0.0444	-0.0138
MHC	-0.0568	0.0143	0.0096	0.0205	0.0248
3p21.31	-0.0391	-0.0789	-0.0741	-0.1134	-0.0566
6q15	-0.0202	-0.0864	-0.0663	-0.0746	-0.0658
12p13.31	0.0042	-0.0702	-0.0951	-0.0908	-0.0682
10q23.31	0.0486	-0.0841	-0.0832	-0.0801	-0.0796
2q32.3	0.1043	-0.0627	-0.0410	-0.0406	-0.0582
16q23.1	0.1337	-0.0716	-0.0304	-0.0616	-0.0714
15q25.1	0.0793	-0.0581	-0.0521	-0.0605	-0.0773
18p11.21	0.0763	-0.0574	-0.0522	-0.0555	-0.0545
7p12.2	0.1509	-0.0255	-0.0660	-0.0681	-0.0751
16p11.2	0.0839	-0.0826	0.0467	-0.0440	-0.0531
18q22.2	0.0850	-0.0009	-0.0177	-0.0341	-0.0456
19p13.2	0.0956	0.0018	0.0098	0.0197	-0.0581
19q13.33	0.0568	-0.0207	0.0154	0.0084	-0.0552
12q13.2	0.0961	0.1362	0.1651	0.0345	-0.0898

Figure 21. Dendrogram produced by hierarchical cluster analysis of T1D regions on genomic structural features. The dendrogram splits the susceptibility regions into three cluster groups CL1, CL2 and CL3. The region size corrected feature values are colour coded using the inset scale.

Table 9. Differences in genomic content between CL 1, CL 2 and CL 3. All values are normalised, and the highest mean values and significant p values are highlighted in red/bold text. (K is the test-statistic of the Kruskal Wallis test; statistical properties are derived from the normalised values).

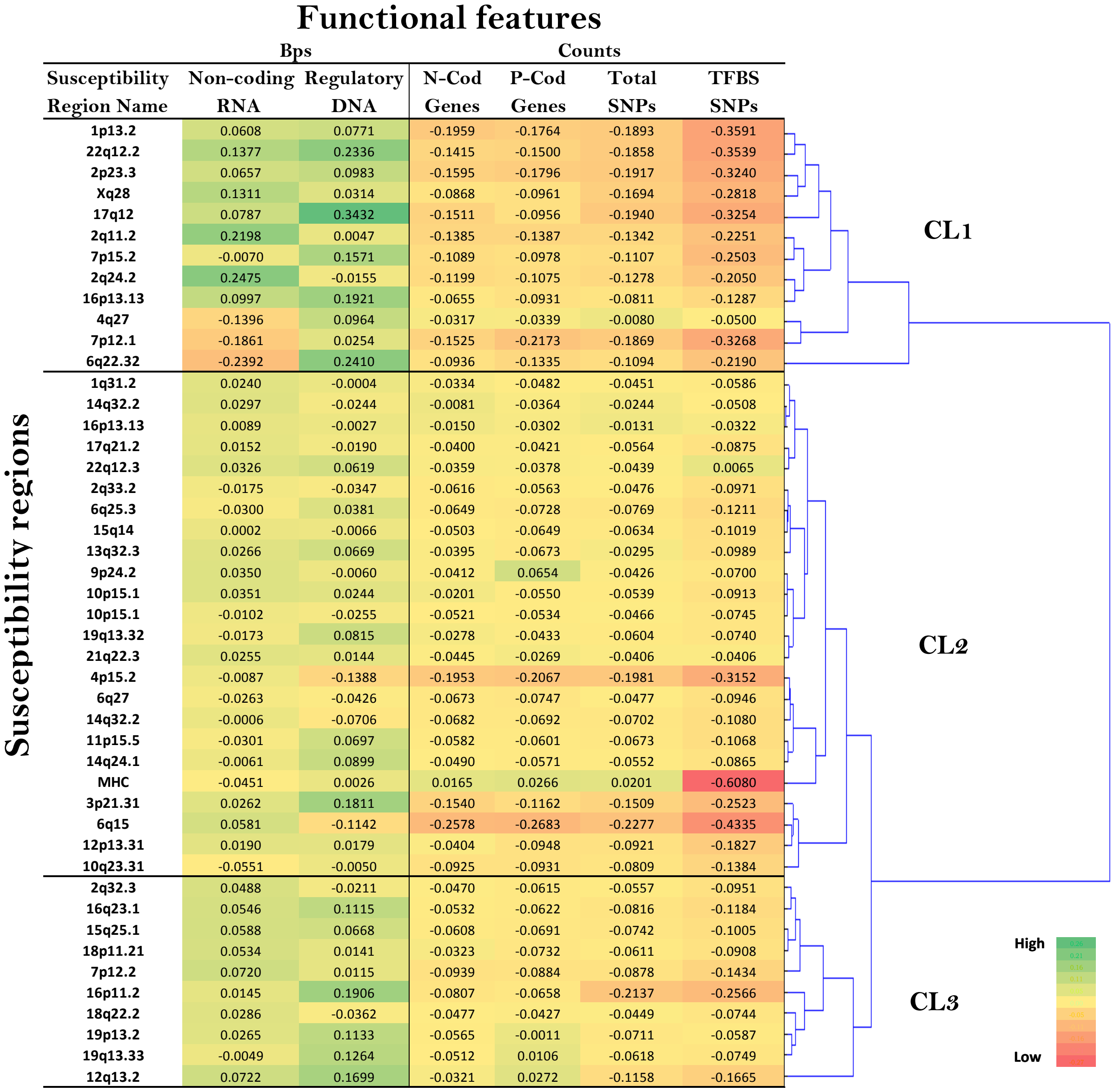
Clusters	Stats	Intronic	3'UTR	5'UTR	Exonic	Intergenic	Region Size
CL1	Mean	**0.1501**	-0.0979	-0.1446	-0.1315	-0.1193	**597513**
	Stdev	0.1240	0.0690	0.0598	0.0545	0.0613	242454
CL2	Mean	0.0235	-0.0502	-0.0417	-0.0458	**-0.0321**	388354
	Stdev	0.0336	0.0239	0.0240	0.0270	0.0241	763859
CL3	Mean	0.0962	**-0.0241**	**-0.0022**	**-0.0302**	-0.0638	298098
	Stdev	0.0278	0.0635	0.0687	0.0372	0.0138	191162
	K(obs)	27.61	9.66	24.71	23.24	25.79	11.86
	P value	<0.0001	0.0080	<0.0001	<0.0001	<0.0001	0.0030

Kruskal – Wallis, $k(crit) = 5.9$

Table 10. P-values for pairwise comparisons of features in CL 1, CL 2 and CL 3

Clusters	Intronic	3'UTR	5' UTR	Exonic	Intergenic	Region Size
CL 1 vs CL 2	**0.0001**	**0.0160**	**0.0001**	**<0.0001**	**0.0001**	**0.0060**
CL 1 vs CL 3	**0.0330**	**0.0330**	**0.0010**	**0.0001**	**0.0460**	**0.0190**
CL 2 vs CL 3	**<0.0001**	**0.5880**	**0.2510**	**0.8750**	**0.0010**	**0.3420**

Mann Whitney U test



Functional features

	Bps		Counts			
Susceptibility Region Name	Non-coding RNA	Regulatory DNA	N-Cod Genes	P-Cod Genes	Total SNPs	TFBS SNPs
1p13.2	0.0608	0.0771	-0.1959	-0.1764	-0.1893	-0.3591
22q12.2	0.1377	0.2336	-0.1415	-0.1500	-0.1858	-0.3539
2p23.3	0.0657	0.0983	-0.1595	-0.1796	-0.1917	-0.3240
Xq28	0.1311	0.0314	-0.0868	-0.0961	-0.1694	-0.2818
17q12	0.0787	0.3432	-0.1511	-0.0956	-0.1940	-0.3254
2q11.2	0.2198	0.0047	-0.1385	-0.1387	-0.1342	-0.2251
7p15.2	-0.0070	0.1571	-0.1089	-0.0978	-0.1107	-0.2503
2q24.2	0.2475	-0.0155	-0.1199	-0.1075	-0.1278	-0.2050
16p13.13	0.0997	0.1921	-0.0655	-0.0931	-0.0811	-0.1287
4q27	-0.1396	0.0964	-0.0317	-0.0339	-0.0080	-0.0500
7p12.1	-0.1861	0.0254	-0.1525	-0.2173	-0.1869	-0.3268
6q22.32	-0.2392	0.2410	-0.0936	-0.1335	-0.1094	-0.2190
1q31.2	0.0240	-0.0004	-0.0334	-0.0482	-0.0451	-0.0586
14q32.2	0.0297	-0.0244	-0.0081	-0.0364	-0.0244	-0.0508
16p13.13	0.0089	-0.0027	-0.0150	-0.0302	-0.0131	-0.0322
17q21.2	0.0152	-0.0190	-0.0400	-0.0421	-0.0564	-0.0875
22q12.3	0.0326	0.0619	-0.0359	-0.0378	-0.0439	0.0065
2q33.2	-0.0175	-0.0347	-0.0616	-0.0563	-0.0476	-0.0971
6q25.3	-0.0300	0.0381	-0.0649	-0.0728	-0.0769	-0.1211
15q14	0.0002	-0.0066	-0.0503	-0.0649	-0.0634	-0.1019
13q32.3	0.0266	0.0669	-0.0395	-0.0673	-0.0295	-0.0989
9p24.2	0.0350	-0.0060	-0.0412	0.0654	-0.0426	-0.0700
10p15.1	0.0351	0.0244	-0.0201	-0.0550	-0.0539	-0.0913
10p15.1	-0.0102	-0.0255	-0.0521	-0.0534	-0.0466	-0.0745
19q13.32	-0.0173	0.0815	-0.0278	-0.0433	-0.0604	-0.0740
21q22.3	0.0255	0.0144	-0.0445	-0.0269	-0.0406	-0.0406
4p15.2	-0.0087	-0.1388	-0.1953	-0.2067	-0.1981	-0.3152
6q27	-0.0263	-0.0426	-0.0673	-0.0747	-0.0477	-0.0946
14q32.2	-0.0006	-0.0706	-0.0682	-0.0692	-0.0702	-0.1080
11p15.5	-0.0301	0.0697	-0.0582	-0.0601	-0.0673	-0.1068
14q24.1	-0.0061	0.0899	-0.0490	-0.0571	-0.0552	-0.0865
MHC	-0.0451	0.0026	0.0165	0.0266	0.0201	-0.6080
3p21.31	0.0262	0.1811	-0.1540	-0.1162	-0.1509	-0.2523
6q15	0.0581	-0.1142	-0.2578	-0.2683	-0.2277	-0.4335
12p13.31	0.0190	0.0179	-0.0404	-0.0948	-0.0921	-0.1827
10q23.31	-0.0551	-0.0050	-0.0925	-0.0931	-0.0809	-0.1384
2q32.3	0.0488	-0.0211	-0.0470	-0.0615	-0.0557	-0.0951
16q23.1	0.0546	0.1115	-0.0532	-0.0622	-0.0816	-0.1184
15q25.1	0.0588	0.0668	-0.0608	-0.0691	-0.0742	-0.1005
18p11.21	0.0534	0.0141	-0.0323	-0.0732	-0.0611	-0.0908
7p12.2	0.0720	0.0115	-0.0939	-0.0884	-0.0878	-0.1434
16p11.2	0.0145	0.1906	-0.0807	-0.0658	-0.2137	-0.2566
18q22.2	0.0286	-0.0362	-0.0477	-0.0427	-0.0449	-0.0744
19p13.2	0.0265	0.1133	-0.0565	-0.0011	-0.0711	-0.0587
19q13.33	-0.0049	0.1264	-0.0512	0.0106	-0.0618	-0.0749
12q13.2	0.0722	0.1699	-0.0321	0.0272	-0.1158	-0.1665

Figure 22. Functional feature values (normalised) are juxtaposed on the dendrogram produced by hierarchical cluster analysis of T1D regions on structural features. The feature values have been corrected for the effect of region size, and are colour coded using the inset scale.

Table 11. Differences in functional features between CL 1, CL 2 and CL 3. All values are normalised, and the highest mean values and significant *p* values are highlighted in red/bold text. (K is the test-statistic of the Kruskal Wallis test; statistical properties are derived from the normalised values).

		Non-Coding	Regulatory	Non-Coding	P-Coding	Total
Clusters	**Stats**	**RNA**	**DNA**	**Genes**	**Genes**	**SNPs**
CL 1	**Mean**	**0.0391**	**0.1237**	-0.1205	-0.1266	-0.1407
	Stdev	0.1546	0.1108	0.0496	0.0496	0.0572
CL 1	**Mean**	0.0037	0.0066	-0.0625	-0.0660	**-0.0673**
	Stdev	0.0288	0.0673	0.0604	0.0650	0.0546
CL 1	**Mean**	**0.0424**	**0.0747**	**-0.0555**	**-0.0426**	-0.0868
	Stdev	0.0254	0.0797	0.0194	0.0401	0.0487
	k(obs)	9.21	12.09	11.81	27.82	20.72
	P value	0.0100	0.0020	0.0030	<0.000	<0.000

Kruskal-Wallis, *k(crit)* = 5.9

Table 12. P-values for pairwise comparisons of functional features in CL 1, CL 2 and CL 3

Clusters	**Non-Coding RNA**	**Regulatory DNA**	**Non-Coding Genes**	**P-Coding Genes**	**Total SNPs**
CL 1 vs CL 2	**0.1350**	**0.0030**	**0.0060**	**0.0001**	**0.0001**
CL 1 vs CL 3	**0.5420**	**0.5840**	**0.0070**	**0.7450**	**0.2940**
CL 2 vs CL 3	**0.0060**	**0.1030**	**0.8380**	**0.0010**	**0.0160**

Mann Whitney U test

In the past, intronic DNA was mostly recognised as sequences that are spliced out during mature mRNA production. But recent studies indicate that certain intronic DNA sequences are further processed after splicing to give rise to functional non-coding RNA transcripts (Djebali et al., 2012; Rearick et al., 2011). In addition, the high intron content in **CL1** may be attributed to the encoding of large genes within the T1D regions in **CL1**. Intronic nucleotides make up a large part of coding genes, and about 26-40% of the human genome is reported to be comprised of intronic regions (Palazzo and Gregory, 2014, Gregory, 2005). Hence, large protein coding genes will most likely contain larger intronic sequences than smaller genes. The T1D susceptibility regions in **CL1** are some of the largest with an average region size of 597,513bps, twice the average susceptibility region size in **CL3** (Table 9). **CL1** also has the highest average count of gene associated nucleotides, approximately 4 times that of **CL2** and 2.5 times that of **CL3**, yet it has the lowest gene density. To support this, an independent study of genomic makeup of the human chromosomes showed that genomic regions with lower gene density tend to contain genes with increased lengths as well as more introns per gene (Atambaeva et al., 2006). The average content of gene associated nucleotides, approximately 4.5 and 2.5 times that of **CL2** and **CL3**, respectively.

The abundance of non-coding sequences (with possibly important regulatory functions) found in the susceptibility regions of **CL1** and **CL3** is significant. It suggests further research to regulatory activity, specifically to RNA regulation in these regions, is needed. SNPs that occur in the regulatory RNA sequences can be studied for deleterious effects such as distortion of binding motifs and formation of aberrant molecules. These are factors that can negatively influence regulatory activity of non-coding RNA and gene regulatory networks (Weinberg and Morris, 2013; Knowling and Morris, 2011; Morris, 2011; Herranz and Cohen, 2010). Recent work by Wan et al., (2014) indicates that over 1,900 transcribed single nucleotide variants (approximately 15% of all transcribed single nucleotide variants) actually alter local RNA structure.

The second question of the study is: are T1D regions are more gene dense than others? This indeed is the case, **CL3** had the highest average gene density relative to region size. The high gene density can be linked with the higher content of exonic and UTR nucleotides in **CL3** which are closer to the expected than for the other two clusters. The difference between clusters, particularly with **CL1** was significant (Table 9). Lastly, **CL2** is the largest cluster of twenty-two susceptibility regions. It includes the HLA region, which is the largest susceptibility region and the region most associated with susceptibility to T1D (rs6916742/C>T, p=4E-307) (Bradfield, 2011). Altogether, the regions in **CL2** are characterised by average counts of intergenic nucleotides (Table 9) and total SNP counts (Table 11) that are less than, but closest to the expected than for the other two clusters.

4.6 Clusters associated with other autoimmune diseases

The third question is, are certain regions are more associated with other diseases than others? This is associated with pleiotropy. This term was coined by a German scientist, Ludwig Herman Plate (Levit and Hoßfeld, 2006), in 1910 and describes the genetic effect of a single gene on multiple phenotypic traits. Certain susceptibility regions can be described as pleiotropic in the sense that they contain markers and genes associated with more than one distinct phenotype (diseases). Twenty-five of the susceptibility regions for T1D are pleiotropic, in that they are also susceptibility regions for fourteen other autoimmune diseases, which share at least one susceptibility region with T1D (Burren et al., 2011) (Table 2 and Table 13).

Pleiotropic regions were found to be dispersed across all three clusters. The relative occurrence (%) calculated for each cluster showed higher occurrences of pleiotropic regions in **CL1** and **CL3**, with the highest being in **CL3**. This is another interesting finding, especially because **CL1** and **CL3** are already outstanding with high content of intronic material, functional non-coding nucleotides, and regulatory module sequences. The pleiotropic regions in **CL1** were most common with Rheumatoid arthritis, Ulcerative colitis and Crohn's disease. Whereas in **CL3,** they were more associated with Multiple sclerosis, Irritable bowel disease as well as Crohn's disease (Table 13). The relative occurrence (%) of these special regions was least in **CL2**, indicating that these regions are mostly only associated with T1D which I also found interesting. Even so, there was some shared susceptibility in **CL2** that was most associated with Coeliac disease followed closely by Rheumatoid arthritis.

Table 13. Diseases associated with the clusters of T1D susceptibility regions

		No of Loci	**Relative occurrence (%)**		
Disease Name	**Alias**	**shared with T1D**	**CL1**	**CL2**	**CL3**
Rheumatoid Arthritis	RA	14	***33.33***	29.17	30
Coeliac Disease	Coeliac	13	16.67	33.33	30
Crohn's Disease	Crohn	12	***33.33***	16.67	**40**
Multiple Sclerosis	MS	11	16.67	20.83	**40**
Irritable Bowel Disease	IBD	11	25.00	16.67	**40**
Ulcerative Colitis	UC	7	***33.33***	4.17	20
Primary Biliary Cirrhosis	PBC	6	16.67	4.17	30
Juvenile Idiopathic Arthritis	JIA	6	16.67	8.33	20
Aqueous Tear Disease	ATD	6	25.00	12.50	0
Systemic Lupus Erythematosus	SLE	5	25.00	4.17	10
Vitiligo	Vitiligo	4	8.33	12.50	0
Primary Sclerosing Cholangitis	PSC	3	8.33	4.17	10
Psoriasis	Psoriasis	2	8.33	0	10
Juvenile Rheumatoid Arthritis	JRA	1	0	4.17	0
Sjögren's Syndrome	Sjögren's	1	0	0	10
Sweet's Syndrome	SS	1	0	0	10
Alopecia	Alopecia	1	0	0	10
Mean			15.69	10.05	18.24
Median			16.67	4.17	10

Friedman; $\chi^2 = 7.03$, $\boldsymbol{df} = 2$, $\boldsymbol{p} = 0.03$

4.7 Chapter summary

The foregoing analysis is based on the premise that characterisation of susceptibility regions in order to highlight prominent attributes can help to focus further research into the disease of interest. Distinct groups of T1D susceptibility regions have been identified by this work. These include a cluster of regions that is rich in non-coding DNA including intronic, non-coding transcript and regulatory nucleotides. A second cluster of regions that contain relatively more genes than others and an abundance of non-coding transcript and regulatory nucleotides was also identified. Furthermore, these regions are associated with more diseases than others. These findings are positive for approaching the study of T1D from a regulatory perspective. This is especially important as the genetic determinants of T1D, being a complex disease, is thought to be better sought in problems associated with gene regulation rather than gene coding (Djebali et al., 2012; Ward and Kellis, 2012; Burton et al., 2007).

What is unique to this part of my study is that measurements have been taken relative to region size which was not done in the previous chapter. By doing this two main features have been identified as characterising the susceptibility regions for T1D which are intron richness and an abundance of non-coding nucleotides. These unique features may explain why the associated T1D-SNPs occur frequently in a combination of introns and non-coding RNA transcripts, and non-associated T1D-SNPs similarly occur frequently in introns. These interesting results still serve to highlight the intricacy of the human genome. Problems in intronic DNA have been linked to genetic problems especially caused by disruption in gene splicing (Flanagan et al., 2013; Wang and Cooper, 2007; Hastings et al., 2005, Lopez-Bigas et al., 2005). Also, mutations in the products of non-coding RNA expression, like linc-RNA and micro-RNA, have also been linked with diseases including cancers (Chen et al., 2013; Shi et al., 2013; Wahlested, 2013; Salta, 2012). The abundant pleiotropy that is characteristic of human complex traits has also been taken up in recent studies in order to dissect and understand genetic relationships between SNPs, genes and clusters of complex diseases (Zhang et al., 2014; Park and Kim, 2012; Sivakumaran et al., 2011; Stranger et al, 2011).

However, the T1D susceptibility regions that are abundant in regulatory DNA are more interesting for the next part of this work which is focused on SNPs in binding sites. The regions contain more regulatory nucleotides than expected, especially in **CL1** and **CL3**, and regulatory regions are known to contain clusters of transcription factor binding sites. **CL2**, is equally important because it contains the most SNPs. Mutations in binding sites as well as active non-coding RNA molecules can cause problems in regulation that lead faulty gene expression, formation of faulty proteins and obstruction of important biological networks, thereby causing disease.

CHAPTER 5

SNP SENSITIVITY

5.1 INTRODUCTION

A mutation in a regulatory sequence can affect transcription factor binding and, as a consequence, the rate of gene transcription. It may lead to an up-mutation, resulting in increased gene expression, or a down-mutation that does the reverse. Clearly, any study devoted to the genomic aspects of a complex disease should take these mutations seriously. In the case of my research, it forms the core of this thesis. Variation in regulatory sequences is common (Garfield et al., 2012) and ever more of these mutations have been detected in binding sites over the years (1,969 in 2005 (Guo and Jamison, 2005), 47,832 in 2008 (Kim et al., 2008)) (Zheng et al., 2012). According to statistics compiled by the Human Gene Mutation Database (HGMD, 2014), 1909 regulatory mutations have been identified in more than 700 genes that cause human-inherited disorders.

Although some publications mention a possible association of regulatory SNPs with increased risk of T1D (Gillespie and Owen, 2014; Asad et al., 2007; Nielsen et al., 2003), until now no study has been done that takes into account the regulatory T1D SNPs. An important objective of this research is therefore to provide an analysis of all regulatory SNPs and, more specifically, to investigate how they might affect the structure of transcription factor binding motifs.

For this, a "SNP sensitivity test" has been developed based on a previous method by Abnizova et al., (unpublished, 2007). The method, which is outlined in detail in section 5.2, assesses the extent to which a mutant allele in a binding site (from now on referred to as a "TFBS-SNP"), compared to its matching reference allele, distorts the representation of the binding motif in which it occurs. Unlike related methods (Chen et al, 2014; Schuab et al, 2012; Laurilla and Lahdesmaki, 2009; Andersen, et al., 2008; Laurilla and Lahdesmaki, 2008; Xu and Taylor, 2009; Abnizova et al., 2007) that rely on the correctness of computationally identified functional regulatory sequences (Chen et al., 2014), my work will only use those SNPs that occur in experimentally confirmed TFBSs (as given by Ensembl's Genome Browser (Cunningham et al., 2014)). This is done to eliminate the problem of false positives associated with the use of computationally predicted binding sites (Struckmann et al., 2011).

5.2 SNP Sensitivity

The SNP sensitivity method, as initially proposed by Abnizova et al. (2007), is a computational approach with two main functions. The first is computational identification of regulatory elements in DNA. However, since this method was proposed, different computational methods for identification of DNA regulatory elements have been developed (Laurilla and Lahdesmaki, 2009; Xu and Taylor, 2009; Andersen et al., 2008; Abnizova et al., 2007). But these algorithms tend to yield a significant amount of false positives. In recent times, high-through-put experimental methods are now applied to computational predictions for ultimate identification of true regulatory elements. These data can be found in dedicated online databases.

The second function of the SNP sensitivity method is to identify variants in the regulatory elements that may affect gene expression, particularly those that affect transcription factor binding. This pertains particularly to change in the motif representation of a binding site, caused by the presence of the mutant allele of a SNP. The purpose of the SNP sensitivity test therefore, is to measure in how far a change in the identity of a SNP alters the underlying motif of a binding site in which it occurs. The motif of the binding site functions as a binding signal for specific families of transcription factor proteins, and change in a single nucleotide identity within a binding sequence could have either a slight or considerable effect on the representation of the underlying motif or binding signal. If the mutant allele of a SNP causes considerable change, it can influence transcription factor binding thus impacting the regulation of gene transcription. The SNP may cause an up-mutation where the mutated nucleotide causes a sub-sequence in the binding region to look more like the consensus sequence of a binding site. This makes the motif of the binding site more conspicuous and triggers transcription by. It increases binding intensity of transcription factors, and a tighter bind leads to an up-regulation of gene transcription. Alternatively, a down-mutation may occur, which destroys a conserved nucleotide in the binding sequence causing it to look less like a binding motif. This reduces binding at the core sequence leading to a down-regulation of transcription.

My research involves developing and implementing a computational method (i.e. the SNP sensitivity test) to measure the change in signal representation caused by the presence of the alternate (mutant) allele of the SNP. The local neighbourhood (adjacent sequence of nucleotides) of each SNP will be analysed for change in sequential properties that occurs when the reference allele of the SNP is substituted with its alternate allele, this is referred to as **SNP sensitivity**. The alternate allele may alter the signal strength of the binding site in which the SNP occurs by causing it to become significantly over-represented (more pronounced) or under-represented (less pronounced). The SNP sensitivity test measures this change. The outcome of this test will be the identification of SNPs with mutant alleles that **significantly** change the representation of their surrounding local neighbourhood. As previously mentioned, the biologically effect on transcription factor binding is either an increase in binding affinity of transcription factors to the binding site, leading to an up-mutation in gene transcription; or a decrease in binding affinity causing a down-mutation in gene transcription. Thus, the overall intention of this study, is to produce a list of regulatory T1D-SNPs that have either of both effects on the binding site in which they occur. They will be suggested as candidate functional regulatory mutations with the potential to influence gene expression by alteration of TF binding signals.

Testing for SNP sensitivity involves three steps: (1) Identification of T1D-SNPs that occur in regulatory regions and transcription factor binding sites, (2) The SNP sensitivity test, (3) Significance testing. These are outlined in the following sub-sections.

5.2.1. Identification of TFBS-SNPs

For this project, SNPs that occur experimentally verified[14] regulatory elements will be identified and accepted as given in the Ensembl genome browser (Cunningham et al., 2014) (see section 2.11). SNPs that occur in regulatory regions (REG-SNPs) and particularly in transcription factor binding sites (TFBS-SNPs) will be selected using Ensembl's Variant Effect Predictor (VEP) tool (version6.3) (McLaren et al., 2010). VEP searches the Ensembl genome browser to locate SNPs that occur in one or more, possibly overlapping, experimentally verified regulatory sequences and binding motifs. In addition, the names of the identified binding motifs, in which the TFBS-SNPs occur, are taken from the Jasper database (Mathelier et al., 2014). This is the largest and freely accessible online resource for Transcription factor binding motifs in genomes of different organisms.

VEP found 10,085 T1D-SNPs to be in regulatory regions (i.e. REG-SNPs). From the REG-SNPs, the tool also identified 92 to be in transcription factor binding sites (i.e. TFBS-SNPs). All other SNPs will be designated as NON-REG-SNPs. The results of the VEP search revealed that none of the disease-associated T1D-SNPs statistically associated with susceptibility to T1D is in a binding site (i.e. they are not TFBS-SNPs)(Table 14). As a result of this finding, the rest of the analysis is focused on the non-associated T1D-SNPs. So, though the 92 TFBS-SNPs identified using the VEP tool are not statistically linked to T1D susceptibility, they are in genetic linkage with the disease-associated T1D-SNPs and are therefore relevant for further study.

[14] Experimental verification by methods such as DNase-seq, ChiP-seq, Histone modification techniques

Table 14. Numbers of associated and non- associated T1D SNPs in three genomic regions

SNP counts	Associated SNPs	Non-Associated SNPs
TFBS	0	92
REG	22	10085
NON-REG	57	250125
Total	79	260302

However, it is interesting to note that 22 of the disease-associated T1D-SNPs do occur in regulatory regions (see Table 14), and this occurrence is significantly more than would be expected by chance (χ^2 = 137.64, df = 2, p << 0.0001, $SR_{REG/ASSOC}$ = 11.50). Also, the disease-associated T1D-SNPs are significantly under-represented in non-regulatory regions ($SR_{NON-REG/ASSOC}$ = -2.29). The expected frequency (0.03) of disease-associated T1D-SNPs in TFBS motifs does not differ much from the observed (0), indicating that a chance occurrence in binding sites within the susceptibility regions for T1D is rather low. The VEP search also reveals that the disease-associated T1D-SNPs that are REG-SNPs are mostly in promoters and promoter flanking regions.

Finally, for the purpose of comparison, a random selection of 400 REG-SNPs and 400 NON-REG-SNPs was made. These numbers are larger than the number of TFBS-SNPs (N=92). The reason for this was simply to have a better representation for these two large SNP categories.

5.2.2. Markov models for local environments

The aim of testing for SNP sensitivity is to identify T1D-SNPs in binding sites that cause significant change in the binding signal of their local environment (i.e. the binding motif in which they occur). Testing for SNP sensitivity starts with fitting a Markov model [15] (Fink, 2007) for the local environment of a SNP. This is a selected part of the regulatory region surrounding a regulatory SNP. The local environment is made up of 601 base pairs, with 300 bps flanking the SNP on both sides. For each TFBS-SNP, this should typically encompass the binding site that overlaps the SNP (Figure 23).

[15] An algorithm (Markov algorithm 1) implemented in python 2.7, establishes the Markov order of nucleotide dependency of each 601bp local environment, and for both alleles of the 92 TFBS-SNPs. A full description of the method can found in (Appendix C).

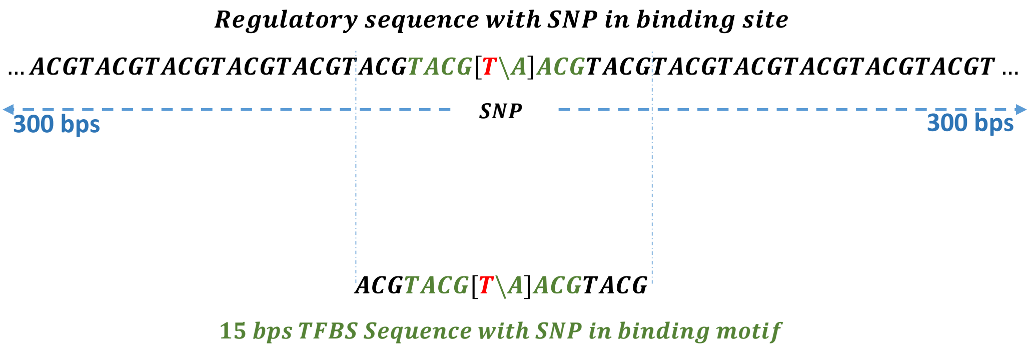

Regulatory sequence with SNP in binding site
... ACGTACGTACGTACGTACGTACGTACG[T\A]ACGTACGTACGTACGTACGTACGTACGT ...
300 bps
SNP
300 bps
ACGTACG[T\A]ACGTACG
15 bps TFBS Sequence with SNP in binding motif

Figure 23. Local environment and neighbourhood of TFBS-SNPs

The fitted Markov model predicts the sequential characteristics of the local environment of the SNP. The order of the fitted model $\boldsymbol{m}$ is an estimate of the degree of sequential dependency of DNA nucleotides in the region (Edwards et al., 2009). Model fitting is done twice per regulatory sequence; first, with the sequence containing the reference allele of the SNP and again with the sequence containing the mutant allele of the SNP. Establishing the Markov models separately ensures for proper computation of expected probabilities for signal representation that will later be done. Ideally, the calculation of expectancies should be on the basis of the established Markov order of the sequence, and it is possible that the order of a given sequence may differ between both alleles of the SNP. A detailed explanation of how the regulatory sequences are fitted with the Markov model and the algorithm design is given in Appendix C. The Ensembl Biomart tool was used to select local environments of each SNP such that the 300 flanking nucleotides remain within the regulatory module in which the SNP occurs. The reason for this is that the Markov order could also differ within a sequence between the regulatory and non-regulatory parts.

Findings: Markov models for local environments

Markov models could only be established for orders $\boldsymbol{m} = 0$, 1 or 2; and only for less than half of the sequences (Table 15). This is either due to the length of the sequence (601 bps), or due to strong non-stationarity. In the first case, it has been shown that the number of nucleotides used to construct a Markov model limits the order to be fitted (Thijs et al., 2001). Exponentially larger sequence lengths are needed to build appropriate transition matrices needed to fit a model for sequences with higher Markov orders. In the second case, it is unlikely that DNA sequences are simple, stationary and low-order Markov chains. A stationary series is one with statistical properties that are constant overtime. Such properties would include the mean, variance, auto-correlation and so on. In a stationary series, there is no change or relationship between adjacent time periods, and the series may be referred to as time homogenous or memoryless. Conversely, in a non-stationary process, there is a difference or relationship in properties between adjacent periods over time (Chatfield, 2003; Priestly, 1981).

Table 15. Number of established Markov models for three types of SNPs. The sequences for which Markov models could not be established are assumed to have Markov orders › = 3.

Markov model	TFBS-SNPs		REG-SNPs		NON-REG-SNPs	
	N	%	N	%	N	%
0	13	14.13	49	12.53	41	10.28
1	5	5.43	68	17.39	86	21.55
2	3	3.26	42	10.74	18	4.51
Not Established	71	77.17	232	59.34	254	63.66
	92		391		399	

In DNA, the genomic signals are likely non-stationary because there is a statistical difference between adjacent coding and non-coding sequences. For example, a three-base periodicity (second order dependency) of nucleotides has been established for coding regions (Howe et al., 2013). Regulatory regions in non-coding DNA also contain distinct motifs that deviate from zero and first-order dependency (Howe et al., 2013; Abnizova and Gilks, 2006; Thijs et al., 2001). This makes them typically non-stationary and their sequence of a fractal nature (Abnizova et al., 2007). These notions are supported by the data presented in Table 15, which indicates a difficulty in establishing models for many regulatory and non-regulatory sequences. Interesting though, a chi square test indicates a significant association between Markov order and genic region ($\chi^2 = 29$, df = 6, $p < 0.001$). Chi-square values are also significant if the test is restricted to two categories of genic regions: i) NON-REG and ALL-REG ({TFBS + REG}, i.e. all regulatory sequences, including those with SNPs in binding sites), ($\chi^2 = 13.35$, df = 3, $p < 0.004$); (ii) TFBS and REG, ($\chi^2 = 15.44$, df = 3, $p < 0.0025$). Standardized residuals point to an over-representation of $\boldsymbol{m} = 0$ models for NON-REG regions, $\boldsymbol{m} = 1$ models for ALL REG regions and more $\boldsymbol{m} = 2$ models than expected by independence for TFBS regions (Table 16). In addition, a regulatory region may overlap with another type of genic region, for instance an exon, which may lead to complex dependencies.

Table 16. Standardized residuals after chi-square tests for associations between genic regions and Markov orders fitted to the data of Table 15. m = Markov order; G-R = Genic Region.

Genic-Region	TFBS	REG	NON-REG
m			
0	0.688	0.494	-0.820
1	-2.845	-0.296	1.659
2	-1.393	2.663	-1.967
higher	1.692	-0.950	0.128

G-R	ALL- REG	NON-REG
m		
0	0.745	-0.820
1	-1.508	1.659
2	1.788	-1.967
higher	-0.116	0.128

G-R	TFBS	REG
m		
0	0.346	-0.17
1	-2.39	1.158
2	-1.9	0.923
higher	1.749	-0.85

It is also important to point out that the total number of the SNPs analysed in the REG- and NON-REG-SNP categories have each been revised (from 400 to 391 and 399 respectively) because some of the selected SNPs have been described as "failed SNPs" in the recently updated version of the Ensembl database (v 73, 80). These are SNPs that have not passed a quality control pipeline[16] set by Ensembl for SNPs.

5.2.3 Change in Signal Representation

A transcription factor binding site is characterised by a special sequence motif which serves as a binding signal for a specific family of transcription factor proteins. In this work, the binding signal will be taken as the direct neighbourhood of each TFBS-SNP. The direct neighbourhood of the SNP is extracted as a 15bps sequence[17], which includes the SNP at the centre of the sequence and 7bps sequences flanking the SNP on both sides (Figure 23).

The representation of the direct neighbourhood (which is assumed to include the binding signal surrounding the SNP) is calculated as a standard residual value $\boldsymbol{SR}$. The value of $\boldsymbol{SR}$ should be calculated on the basis of the established Markov order of the local environment of the SNP. In order to do this a sequence of steps are taken.

The direct neighbourhood is decomposed into sub-strings of "k-mer words" (i.e. a sequence of $\boldsymbol{k}$ nucleotides) using a single step sliding window method (Figure 24). In this study, $\boldsymbol{k} = 3$. This value has been chosen because the Markov order of dependency for regulatory sequences could only be established up to $\boldsymbol{m} = 2$ (i.e. trinucleotide dependency) (see section 5.2.2). The sliding window process generates thirteen trimers per direct neighbourhood. Subsequently, the expected frequency of "trimers" is derived from the best fitting Markov model of the local environment (Thijs et al., 2001, Thijs et al., 2002).

The expected frequencies of each *i*-th trimer ($\boldsymbol{E_i}$) are compared to the corresponding observed frequencies ($\boldsymbol{O_i}$) by converting them to standard residuals $\left(\boldsymbol{SR_i} = (\boldsymbol{O_i} - \boldsymbol{E_i})/\sqrt{\boldsymbol{E_i}}\right)$. The $\boldsymbol{SR}$s indicate if a word is significantly over-represented or under-represented (for $\boldsymbol{SR} > 2.00$ or $\boldsymbol{SR} < -2.00$, respectively). $\boldsymbol{SR}$s are obtained for both alleles of the SNP, i.e. for the same sequence containing the mutant allele and the reference allele. Finally, the difference between scores for both allelic sequences are tested for statistical significance.

[16] The ensembl quality control pipeline flags SNPs with ambiguous information. For example, a SNP that maps to more than one chromosomal position.

[17] The Binding motifs have an average length of around 15 bp (Stewart et al., 2012).

$$\boldsymbol{ACGTACGTACGTACGTACGTACGTACGTACG[T/A]ACGTACGTACGTACGTACGTACGT}$$
$$\boldsymbol{Regulatory\ sequence\ with\ SNP\ in\ binding\ motif}\ (\mathbf{601}\boldsymbol{bps})$$
$$\downarrow$$
$$\boldsymbol{CGTACG[T/A]ACGTACGT}$$
$$\boldsymbol{TFBS\ sequence}\ (\mathbf{15}\ \boldsymbol{bps})$$

$$\begin{gathered}
\boldsymbol{[CGT]ACGTACGTACGT} \\
\boldsymbol{C[GTA]CGTACGTACGT} \\
\boldsymbol{CG[TAC]GTACGTACGT} \\
\boldsymbol{CGT[ACG]TACGTACGT} \\
\dots\dots \\
\boldsymbol{CGTACGTACGT[ACG]T} \\
\mathbf{CG}\boldsymbol{TACGTACGTA[CGT]}
\end{gathered}$$

Figure 24. Sliding window method to decompose SNP neighbourhood into tri-mers

The mutant allele of the TFBS-SNPs may lead to two types of change: (i) the word becomes over-represented - the presence of the SNP enhances signal effect, making the motif a more conspicuous signal for transcription factor binding; or (ii) the word becomes under-represented - the SNP reduces signal effect, making the motif a less conspicuous signal.

5.2.4 Statistical Significance of $\boldsymbol{D_{max}}$

The change in over/under-representation of trimers in SNP neighbourhoods is captured by the difference between the standard residuals of the *i*-th word in the neighbourhood of the reference allele and that of the matching mutant allele ($\boldsymbol{\Delta SR_i}$). The biological interpretation of this is that such a change may lead to increased or decreased binding affinity of a transcription factor. Thirteen $\boldsymbol{\Delta SR_i}$ scores that are generated for each neighbourhood and are subsequently converted to absolute values. The location of their maximum ($\boldsymbol{D_{max}}$) indicates the region in the neighbourhood where the highest change in over- or under-representation between the reference and the mutant allele sequence occurs (Figure 25). $\boldsymbol{D_{max}}$ values are obtained for all the TFBS-SNPs, as well as the REG-SNPs and NON-REG-SNPs.

A large $\boldsymbol{D_{max}}$ suggests that $\boldsymbol{SR_r}$ is much greater than $\boldsymbol{SR_m}$ or vice versa (where $\boldsymbol{r}$ = reference allele and $\boldsymbol{m}$ = mutant allele. This implies that by switching to the mutant allele of the SNP, a core nucleotide in the motif has been affected, consequently causing substantial change in motif representation. Figure 26 illustrates an example of the change in motif representation caused by switching the alleles of SNP rs200372524, $\boldsymbol{SR_m}$ is less than $\boldsymbol{SR_r}$ indicating a decrease in representation. The opposite is shown in Figure 27 for SNP rs3130456, $\boldsymbol{SR_r}$ is less than $\boldsymbol{SR_m}$ indicating that the mutant allele of the SNP causes an increase in motif representation.

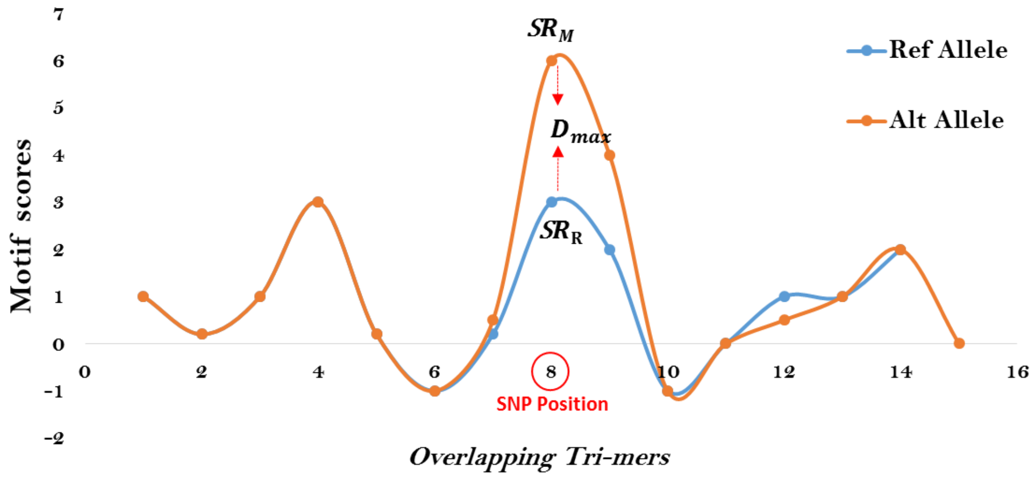


Figure 25. Location of $\boldsymbol{D_{max}}$, the largest change in over- or under-representation between the reference and the mutant allele sequence

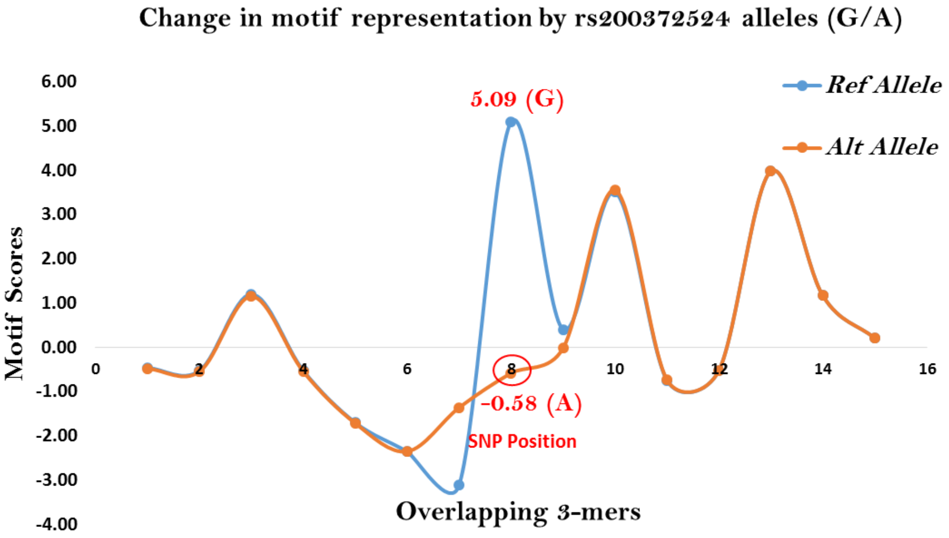


Figure 26. A decrease in motif representation caused by the substitution of alleles of SNP rs200372524 in its local environment, SR_R › SR_M

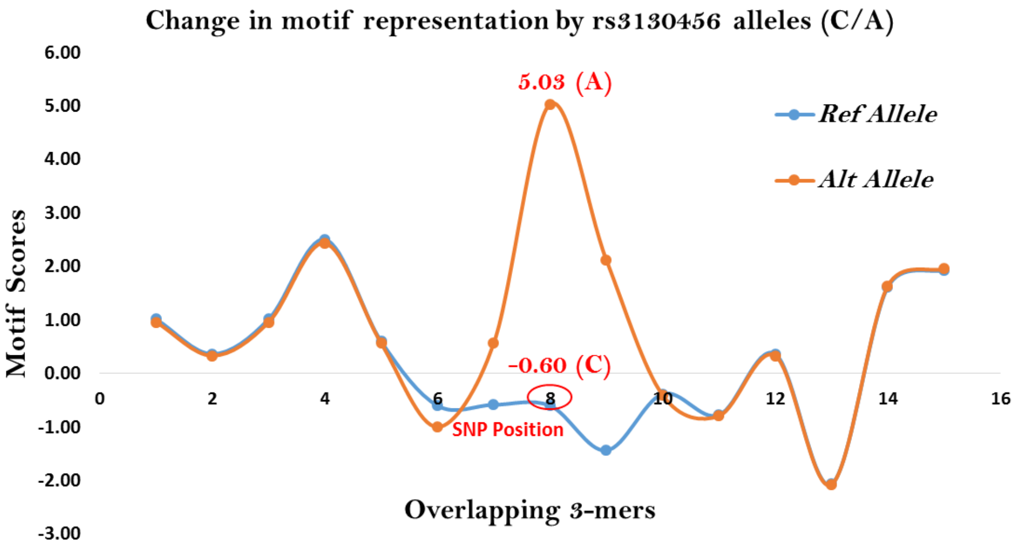


Figure 27. A depiction of increase in motif representation caused by the substitution of alleles of SNP rs3130456 in its local environment, $SR_R < SR_M$

The statistical significance of each $\boldsymbol{D_{max}}$ score was determined so as to assess in how far the change in representation is due to chance. To do this, each DNA sequence was reshuffled 5000 times to yield random permutations of the same sequence. The $\boldsymbol{D_{max}}$ score was obtained for each permuted sequence following the same procedure used to obtain the original $\boldsymbol{D_{max}}$ score. The original $\boldsymbol{D_{max}}$ is considered to be significant when it is larger than 4750 (95%) of the $\boldsymbol{D_{max}}$ scores of the permuted sequences. This corresponds to an empirical $\boldsymbol{p}$-value cut off of 0.05. Those SNPs resulting in a $\boldsymbol{p} < 0.05$, test positive for SNP sensitivity and are associated with substantial change in the trimers making up their direct neighbourhood when alleles are substituted. For those cases in which the Markov order of the sequence could not be established, a SNP was considered significant if at least any of the three $\boldsymbol{D_{max}}$ values (computed for $\boldsymbol{m} = 0$, 1, and 2) are significant. SNPs that test positive for SNP sensitivity (from now on referred to as significant TFBS-SNPs and distinguished as such from non-significant TFBS-SNPs) are those with mutant alleles that have the potential to distort the recognition of the binding motif. They will be selected as candidate functional mutations with the propensity to disturb transcriptional regulation.

An investigation of the possible disturbance in gene activity due to regulatory SNPs should involve a comparison with non-regulatory SNPs. This is taken up in the last section of this chapter which addresses to the following questions: (i) Do TFBS-, REG- and NON-REG SNPs differ significantly in the over- or under-representation of trimers that surround them?, (ii) do the TFBS-SNPs occur in binding motifs that differ from those harbouring non-significant TFBS-SNPs. If so, are these motifs associated with families of transcription factor proteins that relate to particular features or processes typical for T1D?

5.2.5 Differences in D_{max} values between SNP categories

A one way analysis of variance to test for possible differences in D_{max} between the three locations of SNPs was done. The ANOVA Table (Table 17) shows that the D_{max} indeed differ significantly between the locations, with TFBS-SNPs having the lowest values. The overlapping confidence intervals (Figure 28) indicate no differences between TFBS-SNPs and those in regulatory regions (but not in binding sites). This is probably due to the similar sequential properties of the local environments of both types of SNPs. In a comparable study (Andersen et al., 2008), the difference in motif change caused by mutant alleles between SNPs in (computationally predicted) TFBS and other SNPs in the surrounding regulatory regions was compared. As in my study, the difference was not statistically significant.

Table 17. Analysis of variance indicating a significant difference between the Dmax scores of TFBS-SNPs, REG- and Non-REG-SNPs

Anova: Single Factor (Summary)

Groups	*Count*	*Sum*	*Average*	*Variance*
TFBS-SNPs	**92**	**254.450**	**2.766**	**1.928**
REG-SNPs	**391**	**1189.016**	**2.965**	**3.397**
NON-REG-SNPs	**399**	**1300.239**	**3.242**	**2.378**

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**24.856**	**2**	**12.428**	**4.456**	**0.012**	**3.006**
Within Groups	**2485.174**	**891**	**2.789**			
Total	**2510.030**	**893**				

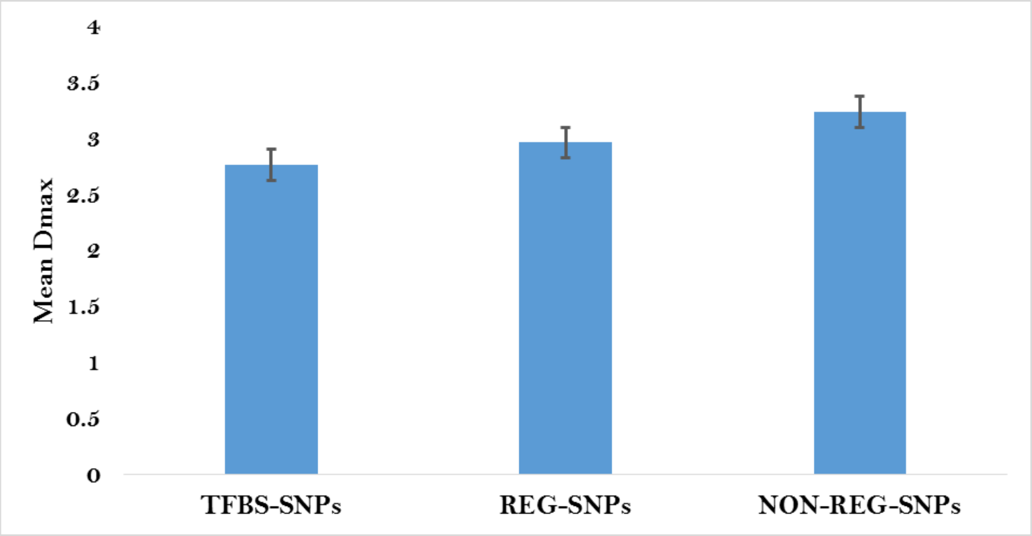


Figure 28. Bar chart depicting a significant difference between the average D_{max} scores of TFBS-SNPs (N = 92) and NON-REG-SNPs (N = 399), but not the REG-SNPs (N = 391).

5.3 Features of Significant TFBS-SNPs

To find out more about the nature of the significant TFBS-SNPs, the values of a number of features (relating to possible regulatory activities) were compared with those from the (other) regulatory SNPs and non-regulatory SNPs. These features include: (i) the type of nucleotide substitution that characterises the SNP, (ii) the other genic positions in which the SNP occurs, (iii) distance to nearby disease-associated T1D-SNPs, and (iv) the type of binding motif in which the SNP is localised. I will first deal with features that relate to what the significant TFBS-SNPs are and where they are found. The last section explores how these SNPs may affect regulation by identifying the motifs they affect.

5.3.1. Identity and Location of significant TFBS SNPs

In this study, 37 out of 92 TFBS-SNPs were found to test positive for SNP-sensitivity. The names and alleles of these significant SNPs, the susceptibility regions in which they occur, the degree of sensitivity ($\boldsymbol{D_{max}}$), and the values of features i) and iii) are shown in Table 18.

Identity of significant TFBS-SNPs: With respect to the identity of single nucleotide substitution, two types of mutations are generally distinguished. Transitions (TI) are SNPs of which the reference and mutant allele are of the same nucleotide class, i.e. both are either a pyrimidine (C, T) or a purine (G, A). Hence, transitions are C-T and G-A SNPs (and the reverses T-C and A-G). Transversions (TV) are SNPs in which a purine is substituted by a pyrimidine (i.e. C-G, G-C, A-T and T-A).

Table 18. The Names and features of significant TFBS-SNPs, including p-values and details of nearby disease associated SNPs. D_{max} values are absolute.

SNP ID	Alleles	T1D Region	Nearby associated SNP	Distance to associated SNP (Bps)	Region Size	$\lvert D_{Max} \rvert$
rs138680304	C/T	2p23.3	rs478222	91600	468897	4.6429
rs114096282	C/T	2p23.3	rs478222	13274	468897	6.9658
rs117640654	G/A	2p23.3	rs478222	75075	468897	3.0526
rs377664089	G/T	3p21.31	rs333	124049	599694	6.0142
rs34638008	C/T	3p21.31	rs333	131238	599694	3.6256
rs140935015	T/C	MHC	rs9268645	1350064	3808585	5.5389
rs140000554	T/A	MHC	rs9268645	1996519	3808585	3.4372
rs151190212	C/G	MHC	rs9268645	621897	3808585	7.0912
rs2267646	G/T	MHC	rs9268645	561683	3808585	4.7346
rs3134944	C/T	MHC	rs9268645	229660	3808585	4.1205
rs35131721	C/T	MHC	rs9268645	831989	3808585	4.1995
rs7741418	C/T	MHC	rs9268645	2312571	3808585	4.1217
rs3130288	C/A	MHC	rs9268645	280303	3808585	3.9943
rs116431137	A/G	MHC	rs9268645	2594487	3808585	3.7497
rs56245106	T/C	MHC	rs9268645	201542	3808585	3.9326
rs201033718	G/C	MHC	rs9268645	449305	3808585	3.6274
rs6921948	A/C	MHC	rs9268645	1205047	3808585	3.4668
rs9262142	G/A	MHC	rs9268645	1726278	3808585	0.0221
rs8192582	C/T	MHC	rs9268645	212692	3808585	3.4722
rs8192581	C/T	MHC	rs9268645	212640	3808585	2.8728
rs13206219	G/T	MHC	rs9268645	201543	3808585	3.3090
rs78180266	C/T	7p12.2	rs10272724	58315	299719	3.8173
rs188548927	C/T	7p12.2	rs10272724	58315	299719	4.5298
rs182785851	G/A	7p15.2	rs7804356	245105	544327	4.1407
rs184649955	C/T	12q13.2	rs2292239	39710	446498	4.0808
rs141305257	C/G	16p11.2	rs4788084	47776	730672	4.3615
rs7203793	C/G	16p13.3	rs12708716	91596	449453	3.7016
rs371243647	C/T	16p13.3	rs12927355	61513	449453	3.4342
rs139221703	G/A	16p13.3	rs12927356	60111	449453	3.4101
rs187731105	G/A	16p13.3	rs12927357	59878	449453	3.4749
rs191450302	C/A	16q23.1	rs8056814	265264	304790	3.4001
rs200372524	G/A	19p13.2	rs2304256	14159	237839	5.6772
rs201991101	C/T	19p13.2	rs2304256	13749	237839	3.2587
rs371391397	C/A	19p13.2	rs2304256	90052	237839	4.0387
rs372996186	G/C	19p13.2	rs2304256	70838	237839	4.9207
rs201432982	C/T	19p13.2	rs2304256	54719	237839	2.6915
rs141193051	G/C	19p13.2	rs2304256	27736	237839	3.3260

The first thing one may notice in Table 18 is that the majority of significant TFBS-SNPs are C-T mutations, which reflects the dominance of transitions (see Table 19). This is not surprising, because transitions in general are more common than transversions. Furthermore, Laurilla and Lahdesmaki (2009) report that C-T transitions are among the most effective SNPs in terms of weakening transcription factor binding. In Figure 29, the nucleotide substitution types are inventoried for significant (S) and non-significant (NS) TFBS-SNPs. Although NS-TFBS-SNPs appear to have more G-A, A-G mutations and less C-T transitions than S-TFBS-SNPs, there is no significant difference between the two categories (NS, S) concerning mutant composition ($\chi^2 = 5.16$, df = 5, p = 0.40).

Table 19. Counts of TFBS-SNP nucleotide substitution types

		TI			TI Total				TV			TV Total	Grand Total
T1D Susc Region	**A/G**	**G/A**	**C/T**	**T/C**		**A/C**	**C/A**	**C/G**	**G/C**	**G/T**	**T/A**		
12q13.2			1		1								1
16p11.2								1				1	1
16p13.3		2	1		3			1				1	4
16q23.1							1					1	1
19p13.2		1	2		3		1		2			3	6
2p23.3		1	2		3								3
3p21.31			1		1					1		1	2
7p12.2			2		2								2
7p15.2		1			1								1
MHC	1	1	5	2	9	1	1	1	1	2	1	7	16
Grand Total	1	6		2	23	1	3	3	3	3	1	14	37

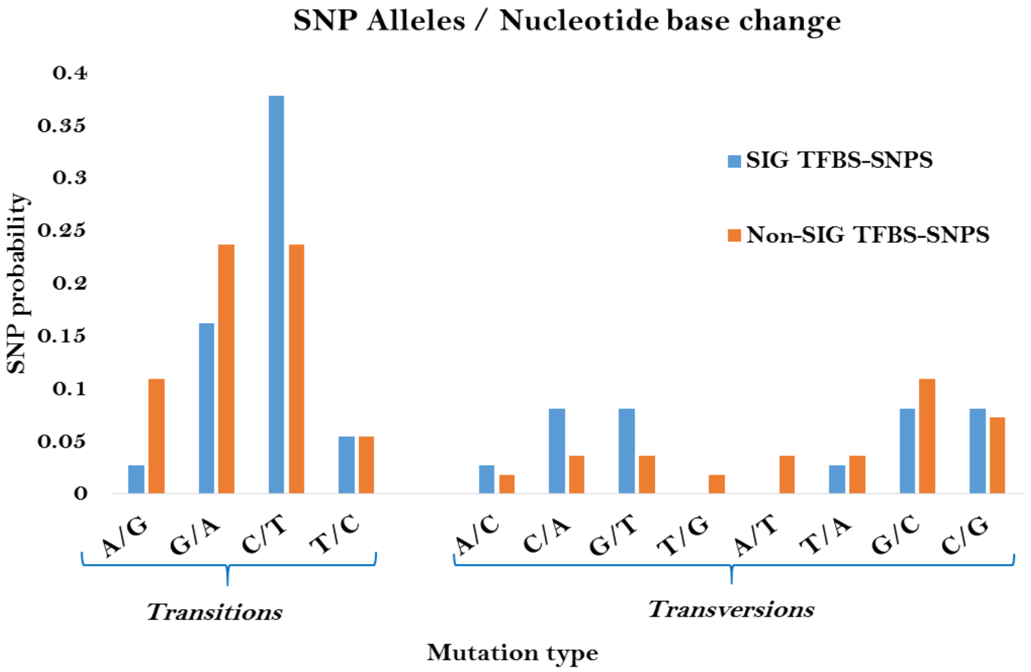


Figure 29. Percentage transition and transversions of significant and non-significant TFBS-SNPs.

Location of significant TFBS-SNPs in susceptibility regions: Another striking feature seen in Table 18 is that most of the significant TFBS-SNPs are found in the MHC (the HLA region), which has been shown to have the associated strongest with T1D (Todd et al., 2010). However, this is simply a consequence of its large size; the MHC is made up of about ten times more nucleotides than the other regions.

As I have shown in the previous chapter, on the basis of its genomic components the MHC is placed in the second cluster, which indeed includes susceptibility regions with the highest SNP density. However, those regions do not have the highest density of significant TFBS -NPs. These happens to be in cluster 3 (CL 3), the one characterised by an abundance of non-coding nucleotides including regulatory DNA (Table 20). This cluster is also composed of regions with on average the highest number of both disease-associated T1D-SNPs and markers for other autoimmune diseases (i.e. "pleiotropic" regions).

Table 20. Counts of Disease-associated T1D-SNPs and TFBS-SNPs in clusters of T1D susceptibility regions.

Cluster Name	**CL1**	**CL2**	**CL3**
Number of Regions (N)	**12**	**24**	**10**
<u>Counts</u>			
Associated SNPs	**19**	**41**	**19**
TFBS-SNPs	**12**	**43**	**32**
Significant TFBS-SNPs	**4**	**18**	**12**
<u>Normalised values</u>			
Associated SNPs	**1.58**	**1.70**	**1.90**
TFBS-SNPs	**1.00**	**1.79**	**3.20**
Significant TFBS-SNPS	**0.33**	**0.75**	**1.20**

Genic positions and –profiles of significant TFBS SNPs: As pointed out in chapter two, a single SNP can affect more than one gene and intersect multiple transcripts. To see in how far this holds for SNPs that significantly change the motif structure of binding sites, the number of genes and transcripts intersected by each TFBS-SNP was counted. Non-significant TFBS-SNPs (N = 55) appear to affect more transcripts than significant TFBS-SNPs (N = 37) (***p*** = 3.74E-06) (Figure 30). This might be due to a possible relationship between gene size and the number of transcripts that gene can produce (large genes contain more SNPs and more transcripts).

Next, I characterised the TFBS-SNPs by their genic-profiles (see chapter 2). Recall that a genic profile comprises the name of each unique type of the genic position in which a SNP occurs. The identification of genic-profiles typical for significant TFBS-SNPs is illustrated in Figure 31.

Profiles that are more typical for either significant- or non-significant TFBS -NPs differ more strongly in their proportions, and are therefore farther away from the diagonal line in Figure 31 (i.e. those profiles that constitute identical proportions of significant- and non-significant SNPs fall along the diagonal).

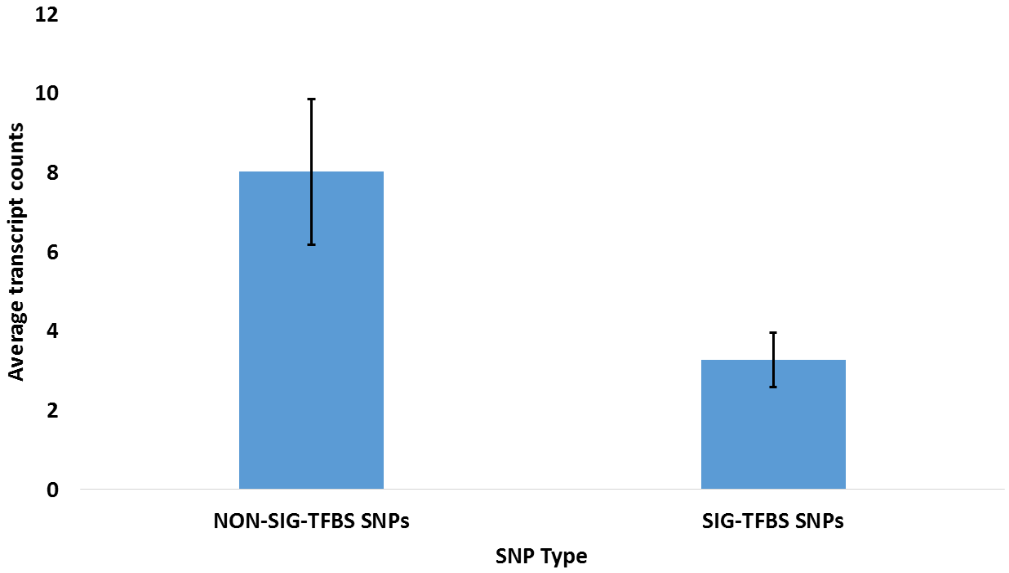


Figure 30. Average number of gene transcripts affected by significant TFBS-SNPs (N= 37) and non-significant TFBS-SNPs N= 55).

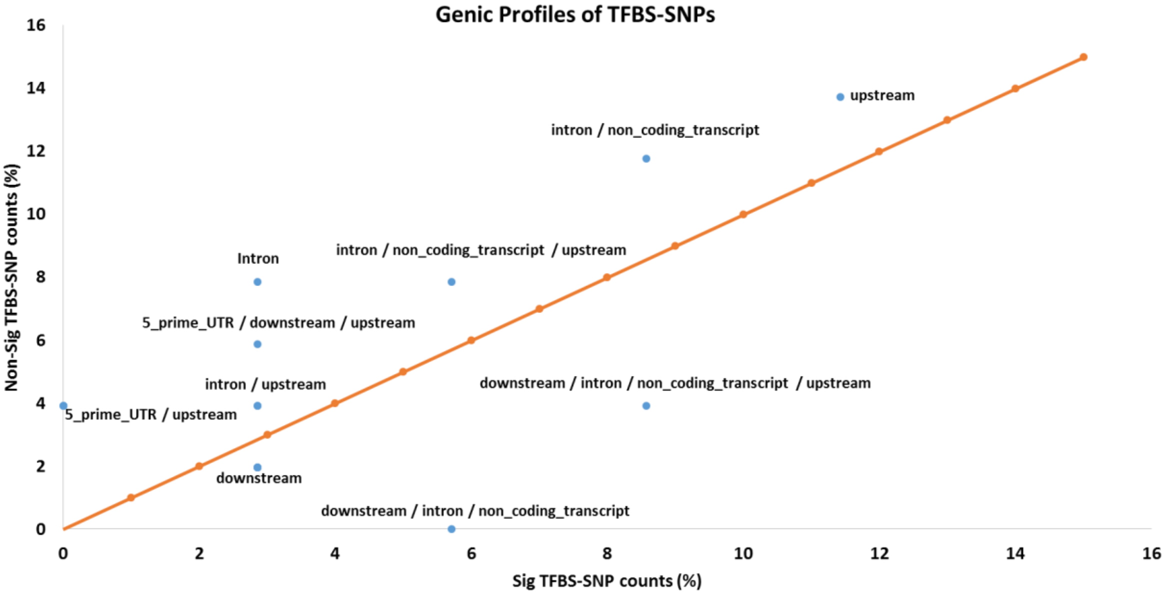


Figure 31. Isolation of genic-profiles common to the significant and non-significant TFBS-SNPs using a scatter plot

Most of the SNPs are in upstream regions of genes, this is where TFBSs are most likely to be found (Table 21). But significant TFBS-SNPs do affect other genomic parts as well, including introns and non-coding transcripts. Note that the same was found for the disease-associated SNPs (chapter 3). This may suggest that the disease-associated SNPs and TFBS-SNPs are close to each other; this is taken up in the next section. Also, all the components of the genic profiles typical for significant TFBS-SNPs (i.e. upstream, intronic, non-coding transcripts and downstream positions) are parts that are exclusively associated with regulatory activity. In other words, apart from affecting the binding motif in which they occur, some of the significant TFBS-SNPs may have an additional effect on overlaying transcripts.

Table 21. . The most frequent genic-profiles of the significant and non-significant TFBS-SNPs

Genic Profile	SIG TFBS-SNPS		Non-SIG TFBS-SNPs	
	Counts	*(%)*	*Counts*	*(%)*
upstream	4	11.43	7	13.73
intron / non_coding_transcript	3	8.57	6	11.76
intron / non_coding_transcript / upstream	2	5.71	4	7.84
downstream / intron / non_coding_transcript / upstream	3	8.57	2	3.92
intron	1	2.86	4	7.84
5_prime_UTR / downstream / upstream	1	2.86	3	5.88
intron / upstream	1	2.86	2	3.92
missense / non_coding_transcript_exon / non_coding_transcript / upstream	1	2.86	2	3.92
downstream / intron / non_coding_transcript	2	5.71	0	0.00
downstream	1	2.86	1	1.96
downstream / intron / non_coding_transcript_exon / non_coding_transcript	1	2.86	1	1.96
intron / NMD_transcript	1	2.86	1	1.96
intron / NMD_transcript / non_coding_transcript	1	2.86	1	1.96
5_prime_UTR / intron / non_coding_transcript / upstream	0	0.00	2	3.92
5_prime_UTR / upstream	0	0.00	2	3.92
downstream / intron / upstream	0	0.00	2	3.92
Total	35		51	

Localisation of significant TFBS-SNPs relative to disease-associated SNPs: Although none of the TFBS-SNPs show up as being statistically associated with T1D in GWAS, this should not be taken as proof for a lack of causality. Current research supports rather the opposite view: disease-associated SNPs, instead of being causative, might be no more than just markers for a region of disease association. As such, any other mutation within that region is a putative causal variant. Therefore, many genomic studies nowadays aim to identify other (potentially causal) SNPs that occur in close proximity and linkage with disease-associated SNPs (Schuab et al., 2012). With this in mind, the distance (bps) between the disease-associated SNPs and both the significant and non-significant TFBS-SNPs were compared. The hypothesis is that significant TFBS variants are in closer in proximity to disease-associated SNPs than non-significant TFBS-SNPs.

Indeed, the average distance between significant TFBS-SNPs and nearby disease-associated SNPs (172185 bps) turns out to be less than the average distance between the non-significant TFBS-SNPs and nearby disease-associated SNPs (355876 bps) (Figure 32). The relationship was tested by means of a two way ANOVA in which the possible effect of susceptibility region was controlled for (Table 22). The strong effect of susceptibility region is due to a large difference in the average SNP distance in the MHC/HLA (Figure 33). If the HLA is taken out, the effect of region disappears but the difference between groups still remains significant (***p*** = 0.040).

Table 22. Analysis of variance indicating a significant difference in average distance between the significant and non-significant TFBS-SNPs and nearby disease-associated SNPs.

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
2p23.3	2	263269	131634.5	10267874905
2q11.2	2	672952	336476	2.26432E+11
3p21.31	2	348150	174075	4311768385
7p12.2	2	103571	51785.5	85268740.5
7p15.2	2	549744	274872	1772148578
12q13.2	2	608582	304291	1.40006E+11
16p11.2	2	405771.5	202885.75	48118069090
16p13.3	2	143655.5	71827.75	25251171.13
16q23.1	2	508318.67	254159.3333	246627243.6
19p13.2	2	142978.83	71489.41667	1381338121
MHC	2	2061692	1030845.975	17702930122
SIG	11	1894043.6	172185.7803	71569654055
NON_SIG	11	3914640.9	355876.4424	1.03609E+11

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1.48702E+12	10	1.48702E+11	5.616340127	0.0058	2.9782
Columns	1.85582E+11	1	1.85582E+11	7.009266466	0.0244	4.9646
Error	2.64767E+11	10	26476725820			
Total	1.93737E+12	21				

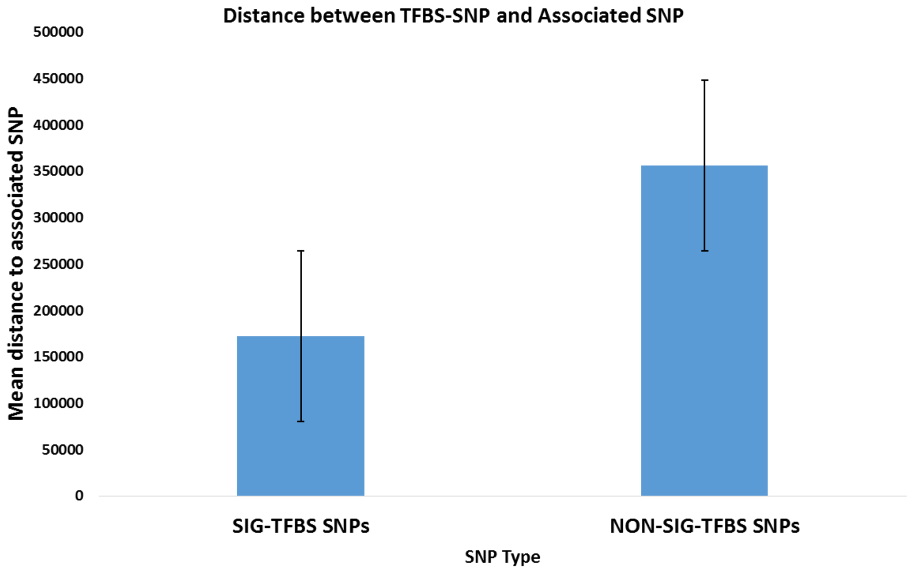


Figure 32. Bar chart depicting a significant difference in average distances of significant TFBS-SNPs (N= 37) and non-significant TFBS-SNPs (N= 55) to nearby disease-associated SNPs

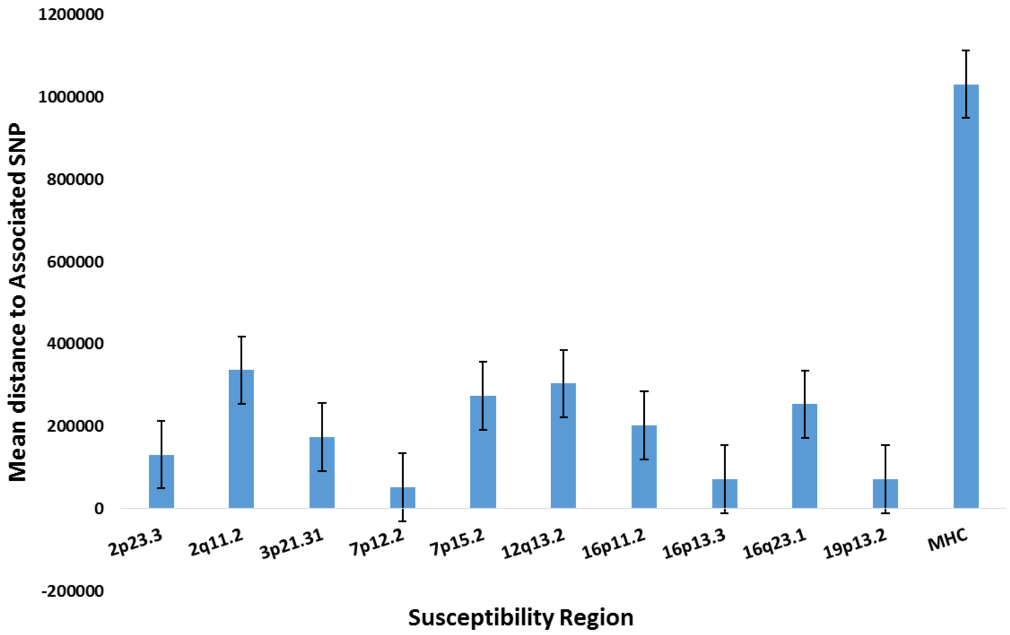


Figure 33. Plot showing differences in average distances to disease-associated SNPs between susceptibility regions.

5.2.3. Binding motifs in which the significant TFBS-SNP occur

For each TFBS-SNP, the names of binding motifs in which it occurs, and the family of transcription factors that recognise and bind to those motifs, were obtained from the Jasper database (Mathelier et al., 2014) (via the Ensembl genome browser). This was done to identify families of transcription factor proteins that might distinguish significant TFBS-SNPs from non-significant TFBS-SNPs. These proteins will then be briefly described in relation to T1D.

From my study, the TFBS-SNPs occur in a total of 31 different binding motifs. Eighteen different transcription factor protein families bind to these motifs. Some of the transcription factors, like JunD and USF, bind to more than one type of motif. These are transcription factors that display diverse target specificity and so have more than a single motif model (Mathelier et al., 2014). A scatter plot of the proportion of significant TFBS-SNPs in each type of binding motif against the proportion of non-significant TFBS-SNPs is depicted in Figure 34. The plot highlights the binding motifs more specific to either of both TFBS-SNP categories. The significant TFBS-SNPs occur more frequently, and twice as much as the non-significant TFBS-SNPs, in binding motifs for the Upstream transcription factor 1 (USF1). The significant TFBS-SNPs also have a high occurrence in binding motifs for the E2F4 transcription factor.

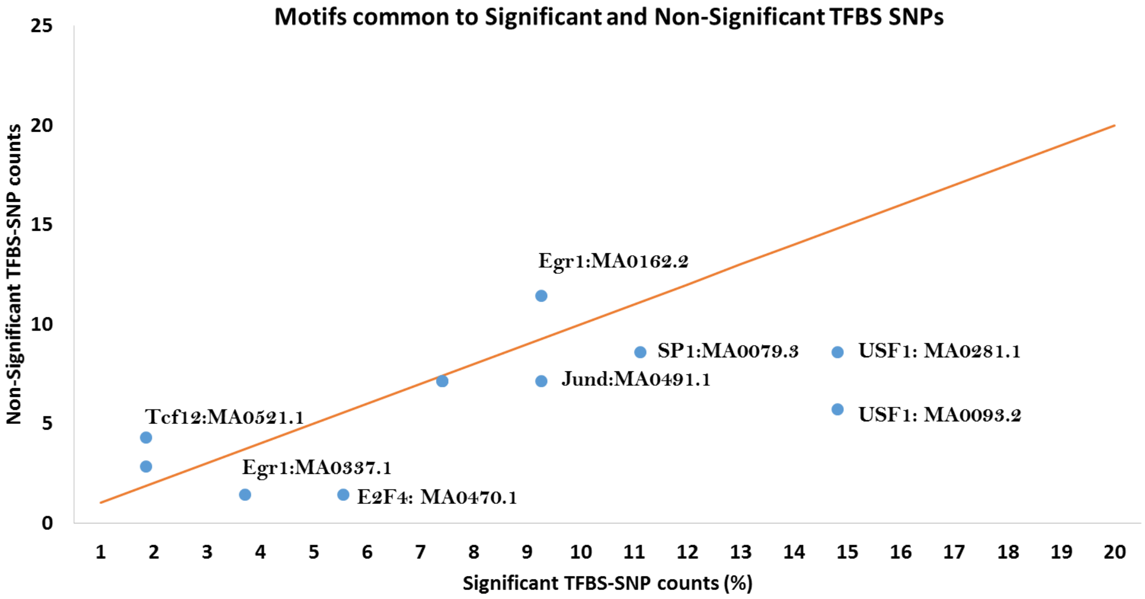


Figure 34. Scatter plot depicting counts of significant and non-significant SNPs in different binding motif structures

The USF1 protein is a cellular transcription factor (Shieh et al., 1993; Corre and Gallibert, 2006) that is thought to activate transcription through binding enhancer (E)-box motifs. (Corre and

Gallibert, 2006; Ewing et al., 2007). Proteins that bind E-box motifs are said to play a major role in regulating gene transcriptional activity (Ewing et al., 2007).

The target genes of USF1 include genes that contribute to the regulation of glucose and lipid metabolism. (Fan et al., 2014; Auer et al., 2012,). Along with another transcription factor, USF2, the USF1 protein has been found to be important for the regulation of different pancreatic islet genes involved in the control of glucose metabolism (Boonsaen et al., 2007; Mirasierra et al., 2006; Martin et al., 2003). Already, this protein has been linked to other forms of diabetes. The locus of USF1 in humans is associated with increased risk of developing Type 2 diabetes (Meex et al., 2008). It is also associated with maturity onset diabetes of the young (MODY) (Bernardo et al, 2008; Qian et al., 1999). More recent studies also link USF1 and USF2 to activation of the promoter for the Alx3 gene (Mirasierra et al., 2011). Expression of Alx3 is required for maintaining adequate levels of expression of pancreatic islet genes including insulin; Alx3 loss-of-function in mice models have shown a progressive decrease in pancreatic islet cell mass and alterations in glucose homoeostasis[18] leading to diabetes (García‑ Sanz et al., 2013). This study is still ongoing, with the intention of providing a more in-depth characterization of the regulation of Alx3 by USF1 and USF2 in pancreatic islets. Table 23 presents a brief description of the protein families associated with the binding motifs in which the significant TFBS-SNPs occur. The USF and EGR protein families are of interest because they are thought to be associated with diabetes. Four motifs associated with these proteins are affected by the TFBS-SNPs. Of these four, three are below the diagonal (Figure 34), hence specific for the significant TFBS-SNPs. Also the USF motifs are most outstanding in terms of distance to the diagonal, they are the most specific protein family for the significant TFBS-SNPs.

[18] Homoeostasis is the property of a system in which variables are regulated so that internal conditions remain stable and relatively constant. Such as the regulation of temperature or the balance between acidity and alkalinity (i.e. pH). It maintains the stability of the human body's internal environment in response to changes in external conditions

Table 23. Brief descriptions of the protein families associated with the binding motifs in which the significant TFBS-SNPs occur

Protein Binding Motif	Protein Family	Brief Description
JunD	Jun	Functional component of the AP1 transcription factor complex, most likely protect cells from p53-dependent senescence and apoptosis i.e. death.
E2F4	E2F	Play a crucial role in controlling the cell cycle and action of tumor suppressor proteins.
EGR1	EGR	Target gene proteins are required for differentiation and mitogenesis. Found in fibroblasts (serum response), neuronal cells (NGF response), lymphoid cells (T cell receptor activation); is thought to enhance insulin resistance in Type 2 Diabetes mouse models, and to be a tumor suppressor gene.
Sp1	Sp/KLF	Involved in cellular processes, including cell differentiation, cell growth, apoptosis, immune responses, response to DNA damage. A regulator of developmental processes that functions in hematopoietic stem cells (cells that differentiate into blood cells).
Tcf12	Helix-Loop-Helix-E	Expressed in many tissues, including skeletal muscle, thymus, B- and T-cells, Thought to play an important role in the development of the nervous system.
USF1	USF	USF1 and USF2 have been found to be important for the regulation of different pancreatic islet genes involved in the control of glucose metabolism Also important in regulation of Pdx1 (pancreatic and duodenal homeobox, a key regulator of pancreatic development and function whose defect is associated with maturity onset diabetes of the young The USF1 locus in humans has been found to be associated with increased risk to develop Type 2 diabetes, The gene has also been linked to familial combined hyperlipidemia (FCHL).

5.4 Chapter summary

Variants that affect regulatory functions have been recognised in the aetiology of certain diseases. Some examples include the blood related diseases β-thalassemia and haemophilia, atherosclerosis (de Vooght et al., 2009), as well as Gilberts syndrome in humans (Bosma et al., 1995). But for T1D, no regulatory SNPs have yet been implicated in the disease mechanism. A SNP that occurs in an experimentally detected binding site, and is closely linked with a disease associated SNP, is more likely to play a biological role in the genome than other SNPs that occur in parts for which there is no particular known function (Schuab et al., 2012). Through this work, I have found that though the associated T1D-SNPs are not regulatory SNPs that may influence transcription factor binding, there are other nearby non-associated SNPs that can influence this process. Thirty-seven of these rare regulatory TFBS-SNPs have been identified by their testing positive for SNP sensitivity. In addition to significantly changing the representation of their local environment, they are significantly closer to disease-associated SNPs than the other TFBS-SNPs. The significant TFBS-SNPs are mostly characterised by C-T transitions, which have previously been shown to cause weaker affinity for transcription factor (TF) binding.

Significant and non-significant TFBS SNPs influence 31 different binding sites for 18 transcription factor families. The binding sites for the USF family of transcription factors are the most affected; these proteins, USF1 and USF2, have been linked to genetic disorders involving the regulation of insulin genes and of the metabolism of glucose. These are typical features of T1D, where insulin is primary auto-antigen[19]. Despite these important findings, further research is needed to determine whether these SNPs do affect function in vivo. Experimentation can reveal if the recognition and binding of TFs to the binding sites in which the significant TFBS-SNPs occur is altered, and how this in turn disturbs the transcription of target genes.

[19] An antigen that despite being a normal tissue constituent of the body is the target of a humoral or cell-mediated immune response, it stimulates the production of autoantibodies and an autoimmune attack as in autoimmune diseases

CHAPTER 6

CONCLUSION AND DISCUSSIONS

This chapter concludes this dissertation. 6.1 highlights the results of my research, and 6.2 discusses the results found for the research questions that made up basis of this thesis. Suggestions for future research are highlighted in 6.3.

6.1 Conclusions

The results of this research suggest an involvement of non-coding SNPs in the aetiology of T1D, but they are not the disease-associated SNPs. Through this work, it is shown that the associated T1D-SNPs are mostly (93%) non-coding SNPs and about a quarter of them are regulatory SNPs. The latter are situated mostly within promoter flanking regions but, none of them occur in a TFBS. Most of the remaining non-coding associated T1D-SNPs are found in introns of protein coding transcripts and in non-coding RNA transcripts. Note, that although these SNPs are not in conventional regulatory modules, they can still affect regulation. Another important finding is that especially the disease-associated SNPs affect multiple processes because many of them (50%) occur in overlapping genes and multiple overlapping gene transcripts. Furthermore, these overlapping transcripts, have different functions. This means that the associated SNPs are capable of affecting multiple processes associated with the biological activity of overlapping parts.

This research has also been able to characterise the T1D susceptibility regions by their genomic content. The T1D regions can be split into three clusters. These characteristics are attributed to the presence of large sized protein coding genes that contain large introns and also have multiple alternative coding and non-coding transcripts. The second cluster includes the HLA (largest) region and is outstanding in the sense that its susceptibility regions contain a lot of intergenic material. It is also the cluster with regions with the largest number of SNPs (most of which are in the HLA region). With respect to its features, the third cluster is quite similar to the first one, but differs from it by having high contents of exonic and UTR nucleotides. This is due to the high gene density of the susceptibility regions that make up the cluster. The susceptibility regions of the first and third clusters are outstanding because they contain significantly more markers for other autoimmune diseases than the regions in the second cluster.

This is particularly interesting because these regions are also enriched in intronic and non-coding transcript nucleotides.

Finally, this study has been able to identify putative regulatory SNPs (mutations influencing regulatory function) that may affect transcription factor binding within the T1D susceptibility regions. A SNP sensitivity test, designed for this study, was used to identify SNPs with alternate alleles that change the representation of the binding region in which they occur. Counter-intuitively, all SNPs that tested positively are non-associated SNPs (i.e. SNPs that did not appear to be significantly associated with T1D in GWAS). However, they occur at loci that are significantly closer in distance to disease-associated SNPs than SNPs that tested negative for sensitivity. About 60% of the SNPs that significantly change the structure of the motif in which they occur are from the HLA region. This is also the susceptibility region that has the strongest association with T1D. Furthermore, they occur predominantly in binding motifs for the USF family of regulatory proteins. These proteins have been shown to be associated with other diseases including Type 2 Diabetes.

6.2 DISCUSSION

Although the aetiology of T1D is not fully understood, aberrations in the regulation of certain susceptibility genes, like CTLA-4, PTPN22 and IFIH19 (see Table 2 for susceptibility genes), are suspected to contribute to the cause of disease (Gillespie, 2014). The main aim of this study was to find out in how far T1D can be considered to be a disease caused by disruption in regulation rather than in gene coding. This thesis was addressed by the following four main objectives: (i) establishing the distribution of the T1D-SNPs (including disease-associated and non-associated SNPs) in various genomic parts, (ii) establishing the proportion of SNPs is located in regulatory regions (iii) finding out how many SNPs in regulatory regions are located in transcription factor binding sites (TFBS-SNPs), and then (1V) testing the identified TFBS-SNPs for SNP sensitivity.

Initially, I set out to out describe the T1D-SNPs by establishing their distribution in coding and non-coding parts of the genome. The aim was to find out whether the disease-associated SNPs occur more in non-coding parts of the genome compared to the non-associated SNPs. This is because it is now widely recognized that SNPs associated with complex diseases map to or are found in non-coding sequences (Dirk et al., 2014; Zhang et al., 2014; Djebali et al., 2012). It would be interesting to see if this trend also holds for the T1D-SNPs. If the disease-associated T1D-SNPs occur frequently in non-coding regions, they may also be within a regulatory module, as a large part of non-coding genome is now believed to be associated with regulatory activity (Mercer and Mattick, 2013; Thurman et al., 2012). The first part of the study yielded positive results, revealing that the disease-associated T1D-SNPs are frequently non-coding nucleotides in overlapping introns and non-coding RNA transcripts. They were also found to be less often in gene-flanking regions than would be expected by chance. Although non-associated SNPs also occur often in intronic parts of protein-coding transcripts, they are found less often in non-coding transcripts and more often in upstream regions of genes.

The subsequent study involved a classification of the susceptibility regions for T1D on the basis of their genomic content. It revealed that the regions are characterised by a high abundance of intronic nucleotides as well as high amounts of non-coding RNA nucleotides. This may relate to the high occurrence of SNPs in these particular regions.

In the final part of my study it was established that the disease associated SNPs do not occur more often in regulatory DNA and in TFBSs than the non-associated SNPS. This outcome was unexpected. Only twenty-two of the seventy-nine disease-associated SNPs occur in regulatory DNA. They are mostly within promoters and in promoter-flanking regions, but they are not in binding sites. Because the disease-associated T1D-SNPs are significantly less frequent in gene flanking regions, this outcome could have been imminent. Regulatory regions, are thought to be symmetrically distributed around genes without bias (Birney et al., 2007; Dineen et al., 2007), but TFBS are quite commonly found in upstream gene flanking regions. The non-associated T1D-SNPs are more frequent in upstream regions. Consequently, many of them (over 10,000) occur in regulatory regions, and a total of 93 are in experimentally verified TFBSs.

The last objective of the study was to identify the regulatory SNPs in TFBS (TFBS-SNPs) that change the underlying sequential structure of their surrounding binding region. Biologically, significant distortion can lead to changes in the binding affinity of a transcription factor due an up-mutation or a down-mutation. The SNP sensitivity test developed for this project allowed for the identification of SNPs that could have this type of effect. The test was designed to analyse the local region surrounding the TFBS-SNPs by computing trimer probabilities based on the established order of nucleotide dependency. Despite a challenge in establishing the Markov order of nucleotide dependency for some of the regulatory sequences, 37 out of the 93 TFBS-SNPs tested positive for sensitivity. Especially strong results were found for rs140000554 and rs3134944 in the MHC region, and rs201991101 in region 19p13.2. All three SNPs occur in binding motifs for the USF family of transcription factor proteins which have been linked with other diseases.

For T1D to be viewed as a disease that is caused by problems in gene regulation, the classical expectation would be for the disease-associated SNPs to be frequently located in regulatory regions and binding sites. Although this is not the case, this does not rule out the hypothesis that T1D is for a large part due to gene regulatory defects. The disease-associated SNPs occur the least often (7%) in protein coding regions; therefore, T1D cannot be described as a disease that is caused by disruption in protein coding alone. The findings of this research can be related to two current trends in the study of complex diseases. The first is centred on the function of disease-associated SNPs in susceptibility regions. It is now widely suggested that the disease-associated SNPs may be no more than markers that capture the variation present at a locus associated with disease risk (i.e. the disease susceptibility region) (Zhang, et al., 2014). They are unlikely to be the mutations that underlie disease association, but rather are in linkage with other genetic variants, all of which are putatively causal (Chen et al., 2014; Zhang et al., 2014; Marian, 2012; Schuab et al., 2012). This recent turn in complex disease genomics has come about because despite a number of post GWA-studies, many of the disease associated SNPs are still yet to be implicated as the underlying causal variant in associated complex diseases (Knight, 2014). Recent studies now seek to identify other rare SNPs that are close by and in strong linkage with the disease-associated SNPs, which could account for the difference in phenotype that is associated with the region (Chen et al., 2014; Zhang et al., 2014; Marian, 2012; Schuab et al., 2012). The concept presents an important challenge in the sense that there are usually numerous variants linked with the associated SNPs. For instance, the susceptibility regions for T1D contain well over 200,000 non-associated SNPs. Therefore, if the notion that disease-associated SNPs are markers is true, then is has been quite significant to be able to reduce the number of non-associated T1D-SNPs to 37 putatively causal regulatory candidates that test positive for SNP sensitivity. Furthermore, it has been important to find that these 37 regulatory SNPs have a significantly shorter mean distance to close-by disease associated SNPs in the region than the SNPs that tested negative for SNP sensitivity. One can then theorise that the associated SNPs are markers for these rare SNPs that affect regulation. Further studies could reveal that the target genes activated by TF binding to the affected TFBSs, could be implicated in the aetiology of T1D.

Secondly, the associated T1D-SNPs could still influence regulation because of the type of genic position in which they occur. Apart from the classic regulatory elements (promoters, enhancers, silencers, insulators and locus control regions) (Felsenfeld and Groudine, 2003; Gross and Garrard, 1988; Stalder et al., 1980), other non-coding sequences including introns and non-coding transcripts have been shown to be involved in regulation within the genome (Mercer and Mattick, 2013; Djebali et al., 2012; Thurman et al., 2012). The "intronic/non-coding transcript" genic profile found to characterise many associated T1D-SNPs in this study, indicates that gene regulatory activity involving introns and non-coding transcripts can possibly be disrupted by these SNPs. Particularly, regulation associated with gene splicing, nonsense-mediated decay (NMD) and RNA transcripts. In addition, the associated SNPs appear to be able to affect more than one process because they tend to occur in more than one gene and more than one alternative transcript of the same gene. A simultaneously occurrence in two genomic structures is interesting especially if they both stand for different functions. Furthermore, mutations in introns and non-coding transcripts have been found to cause disease. The following is a summary of a few examples.

The introns have long been thought to be extraneous nucleotides that are usually spliced out from the primary mRNA transcript of a gene to produce the mature mRNA transcript that guides the production of proteins. However, introns are now known to be crucial in the regulation of gene expression as well as influencing molecular evolution (Vreeswijk et al., 2008). They contain a splice site consensus motifs (a sequence of DNA that has a similar structure and function in different organisms) at the intron-exon junctions, which are bound by a protein complex known as the spliceosome during the alternative splicing process. The enzymes involved induce cleavage of the intron from the flanking exons and then joins the two flanking exons by what is called a phosphodiester bond[20]. A SNP that occurs in these motifs can disturb alternative splicing by altering recognition of the splice site or by altering splicing patterns (Heyd, 2014; Ward and, Cooper, 2010). Mis-splicing can result in exon skipping, intron retention or the activation of a cryptic splice site (Flanagan et al., 2013; Wang and Cooper, 2007; Hastings et al., 2005; Lopez-Bigas et al., 2005). Rs698 is an associated T1D-SNP that occurs at intronic splice sites in two overlapping genes, INS and INS-IGF2 on chromosome 11p15.5. The INS gene encodes insulin, a hormone which decreases blood glucose concentration and increases cell permeability to monosaccharides, amino acids and fatty acids. Research has shown that the protective allele of this SNP (homozygous "T") is associated with increase in age of disease onset (Howson et al., 2010). This mutation is associated with production of anti-insulin autoantibodies (Howson et al., 2010), but how the SNP affects the INS gene locus has not been fully described (Raha et al., 2011). Medical research has demonstrated that mutation in splicing can play important roles in disease including hereditary diseases, neuro-degenerative disorders and cancers (Romano, et al., 2013; Ward and Cooper, 2010; Vreeswijk et al., 2008). Some diseases found to be caused by problems in splicing include Familial Dysautonomia (FD) (Rubin and Anderson, 2008; Close et al., 2006), atypical cystic fibrosis (Schram, 2012), Menkes disease (de Bie et al., 2007) and Frasier Syndrome (Klamt et al., 1998; Frasier et al., 1964). Yet, intronic mutations are quite often ignored as possible causes of human disease (Shiraishi et al., 2014; Homolova et al., 2010). Also, an intronic SNP can affect regulation if it occurs in a regulatory region that overlaps the intron (gene-associated regulatory region). If it is within a binding site it could distort the binding signal as demonstrated with SNP sensitivity, thus affect binding of regulatory proteins. However, no disease-associated TFBS-SNP occurs in a gene-associated binding site.

The afore-mentioned SNP, just like 23 other associated SNPs, occurs in non-coding transcripts of the same susceptibility gene or in transcripts of a second overlapping gene. These non-coding transcripts include NMD transcripts, retained introns as well as RNA gene transcripts. NMD decay of a transcript, functions to detect and degrade transcripts harbouring premature signals for the termination of translation (Frischmeyer, 1999). An estimated one-third of inherited

[20] Phosphodiester bonds make up the backbone of the strands of DNA. It is the linkage between the 3' carbon atom of one sugar molecule (nucleotide) and the 5' carbon atom of another.

genetic disorders and many forms of cancer are thought to be caused by a nonsense or frameshift mutation which results in the generation of premature termination codons (Baker and Parker, 2004; Frischmeyer, 1999). The majority of these nonsense transcripts are recognised and degrade by the cell by the NMD pathway. However, a mutation can lead to a gain of function mutation that decreases the efficiency of NMD and causes the stability of nonsense transcripts. NMD mutations are implicated in diseases including the blood disorder Beta thalassemia (Galanello et al., 2010), Marfan syndrome (Kainulainen et al, 1994; Marfan, 1896), and Addison's disease (Nieman and Chanco Turner, 2006; Ten et al, 2001).

Regarding RNA genes, four associated T1D-SNPs occur in non-coding RNA transcripts that include lincRNAs, microRNAs and siRNAs. These gene transcripts (see chapter 2) are associated with a wide range of biological processes most of which are regulatory (Kumar et al, 2013). A lot can be said about the disease-associated T1D-SNPs in non-coding RNA transcripts, however the lincRNAs are most interesting, because they make up a large percentage of the non-coding human transcriptome (Cabili et al., 2011; Birney et al., 2007). They are associated with regulation of gene expression (Khalil, et al., 2009), differentiation of embryonic stem cells (Guttman et al., 2011), and maintaining cellular physiology (Di Gesualdo et al., 2014). LincRNAs have an exon-intron structure but they do not have open reading frames that encode proteins (Orom et al., 2010). A problem in a lincRNA transcript can lead to dysregulation of its target gene thus contributing to disease (Shi et al, 2013; Modarresi, 2012); this occurs in the aetiology of Alzheimer's disease (Tan et al., 2013; Faghihi et al., 2008). LincRNAs are also implicated in the aetiology of other diseases including Huntington's disease (Johnson, 2012), type 2 diabetes (Pasmant et al., 2011), neuro-generative disorders (Salta, 2012), and are thought to play a role in the onset and development of Cancers (Chen et al., 2013; Deng and Sui, 2013; Pasmant et al., 2011). The associated T1D-SNP, rs941576, occurs in an intron of MEG2, a maternally imprinted[21] linc-RNA gene on chromosome 14q32.3. MEG2 is expressed in many cells in the human body and is thought to be a tumor suppressor (Balik et al., 2013; Zhou et al., 2012). It has been shown to inhibit proliferation of tumor cells as well as induce apoptosis (death) of the cells in humans (Zhou et al., 2012).

However, there is robust evidence that rs941576 in the maternally expressed MEG2 gene, is associated with paternally inherited risk of T1D (Wallace, et al., 2010). It is hypothesized that this SNP may affect regulation of three other paternally expressed genes (DLK1, RTL1 and DIO3) that lie within the region (Arney, 2003). DLK1 has the strongest functional candidacy, as it is highly expressed in the pancreatic islets (*T1D is characterised by attack on pancreatic islet beta cells*) (Wallace, et al., 2010). In this study, rs941576 was also found to be in an enhancer[22]

[21] For most genes, two working copies are inherited each from mother and father. With imprinted genes, only one working copy is inherited. For maternally imprinted genes the copy from the father is epigenetically silenced (by addition of methyl groups during egg or sperm formation.) and vice versa.

[22] Enhancer is a short (50-1500 bp) region of DNA that is bound by activator proteins to initiate transcription of a gene or genes.

region within a regulatory module, but not in a binding site. From the foregoing, it is clear that mutations occurring in the non-coding parts of the genome are just as important as coding mutations in the causation of diseases.

Another significant outcome of this research is that the analyses done (classification of SNPs and susceptibility regions) and methods developed (the SNP-sensitivity test) can be applied for other diseases, especially complex diseases for which the contributing genetic factors have already been discovered. These methods can be applied to genomic data that has been made available for these diseases by dedicated research consortiums. This way, regulatory mutations that may contribute to the disease mechanism can be identified for further testing. In fact, comparing results from different diseases, especially mutations coming from pleiotropic susceptibility regions, may lead to isolation of particularly interesting mutations. These may include SNPs that have a common occurrence for more than one disease but yet affect a different process in the aetiology of each diseases because of the mere fact that SNPs intersect multiple transcripts and genes.

Sometimes discovering a mutation or a mutated gene that causes a disease is simply the first step down the long road to using information to produce a cure. Hopefully, with time, this research will contribute positively to the life of people living with T1D. Being able pinpoint mutations, and then discover how they contribute to the genetic cause of a condition, can help to open up paths for pharmaceutical treatments. Currently, most treatment strategies for genetic disorders do not alter the underlying genetic mutation; but are designed to improve particular signs and symptoms associated with the disorder. For instance, T1D is managed by administration of insulin injections and dietary management, including keeping track of the carbohydrate content of food, and careful monitoring of blood glucose levels using glucose meters (Diabetes UK, 2015). Though invaluable to people living with T1D, insulin treatment may sometimes lead to hypoglycemia (low blood glucose levels) due to imbalance between insulin, food and physical activity. Mild cases can be self-treated by intake of a high sugar content food or drink, but severe cases can lead to unconsciousness requiring treatment with intravenous glucose or glucagon injections (Diabetes UK, 2015; ADA, 2015; Chiang et al., 2014). Discovering the underlying genetic cause of T1D would be remarkable; and for possible treatments, two emerging approaches that come to mind. The first, gene therapy (Geddes, 2013; GeneEd, 2012; Sheridan, 2011), though still largely an experimental technique, has yielded promising results for treatment of some genetic diseases like haemophilia (NIH, 2014; Ponder, 2011; Ponder, 2006) and Parkinson's disease (Palfi et al., 2014; Lewitt et al., 2011). It is a process that involves delivering normal DNA to correct (repair or replace) the defect in the genome that may be causing the disease. The second, personalized/precision medicine (NIH, 2015; Dudley and Karczewski, 2014; Lu et al., 2014; PMC, 2014), is an approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. These are the three factors that constitute the definition of a complex disease.

Because more than one genetic mutation contributes to T1D, the differences that occur between individuals of different backgrounds (for instance, race and locality) may need to be considered in the design of treatments. Personalized medicine is about the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or in their response to a specific treatment (Blau and Liakopoulou, 2013; Timmeman, 2013). This will allow for a more accurate diagnosis per individual, and design of specific treatment plans including gene therapy.

To conclude, in complex diseases studies, discovering a contributing factor and then characterizing its contribution to the disease is not quite an easy undertaking. This research initially set out to describe the SNPs associated with susceptibility to T1D as variants that cause disease by influencing transcription factor binding. This was not found to be. Instead, nearby non-associated T1D-SNPs were identified as the putative causal regulatory SNPs that could impact binding. The associated SNPs are either likely to influence regulation through other alternative processes or they are markers that have led to the identification of potential causal SNPs.

6.3 Future Research

There are a number of potential avenues that can be explored for future research from this work. Some of these include in-silico methods to identify target genes and their functions.

1) **Identification of Target genes**: What genes are targeted by the transcriptions factors that bind the sites in which the significant TFBS-SNPs occur? How can they be identified?

There are online bioinformatics resources from which this type of information may be sourced. HaploReg (Ward and Kellis, 2012) is an online tool that can be used for exploring annotations of the non-coding genome. It contains data from the 1000 Genomes Project (*www. 1000genomes.org*) about disease-associated SNPs and other non-associated SNPs in linkage with the associated SNPs. It also contains a map of chromatin states, which contains information about regulatory regions including promoters, enhancers, insulators and heterochromatin in nine human cell lines (Ernst et al., 2011). Protein binding information is also incorporated into this database form protein-binding microarray (PBM) experiments (Berger et al., 2006, Berger et al., 2008, Badis et al., 2009). All of this information can be mined using simple queries.

Another tool is regulomeDB (Boyle et al., 2012), which is designed to align variants with regulatory information from a variety of sources. It includes information from high-throughput, experimental data sets from the ENCODE project and other sources. It also contains computational predictions and manual annotations to identify putative regulatory potential and to identify functional variants. The available information is expected to help and guide interpretation of regulatory variants in the human genome (Boyle, et al., 2012).

2) **Candidate causal genes**: Have any of the identified target genes already been recognised as candidate causal gene of T1D? Once target genes have been identified, an interesting line of enquiry would be to find out the function of these genes and to see how their functions may be linked to the biology of T1D. Mainly, one can ask if they are candidate genes already thought to be associated with T1D. If not, are they genes involved in the regulation or activation of T-cells or genes involved in innate immunity? To check if an identified target gene is also a candidate gene, one can look up information about T1D from T1Dbase, OMIM or Ensembl. The functional annotations of genes can also be searched from other online biological databases including UniProt, NCBI, DAVID, NONCODE, miRBase and fRNAdb.

3) **Not a candidate causal gene**: If the genes are not found to be directly linked with T1D, then one can ask whether they are part of a gene-gene regulatory network that links them to a T1D susceptibility gene. This can reveal if and how their function relates to disease.

4) **Associated T1D-SNPs in regulatory regions**: In a different context, the associated T1D-SNPs that occur in regulatory regions, particularly promoters, can also be further analysed for influence on their surrounding region. Computationally, one can test whether these SNPs may lead to an up-mutation by creating a new/false binding site within the regulatory module.

Results obtained from these computation-based questions can further streamline information and help guide the going about of experimental validation to identify the SNPs that impact binding in-vivo.

6.4 Publications and Conferences

Characteristics of T1D susceptibility regions
Sylvia Beka, Rene te Boekhorst and Irina Abnizova
Bioinformatics Italian Society (BITS) meeting,
University of Catania, Italy.
May 1-4, 2012.

Participant
Lipari Summer School for Bioinformatics 2013. Lipari, Italy

The Genomics of T1D susceptibility regions
Sylvia Beka, Irina Abnizova and Rene te Boekhorst
Mathematical and Statistical Aspects of Molecular Biology (MASAMB) meeting 2014,
Sheffield Institute for Neuro Science, Sheffield, UK.

Poster presentation
Excellence in Research Conference
University of Hertfordshire, Herts, UK.
September 26, 2014

Other events attended

Cambridge Next Generation Sequencing Day IV
Cambridge Computational Biology Institute
Centre for Mathematical Sciences, Wilberforce Road
University of Cambridge, Cambridge.
March 26, 2012

Bio-Linux Course
European Bioinformatics Institute (EBI)
Wellcome Trust Sanger Institute, Cambridge.
April 1-3, 2012

Annual Symposium 2012
Cambridge Computational Biology Institute
Centre for Mathematical Sciences, Wilberforce Road
University of Cambridge, Cambridge.
May 24, 2012

Ensembl Course
European Bioinformatics Institute (EBI)
Wellcome Trust Sanger Institute, Cambridge.
April 1-3, 2013

Cambridge Next Generation Sequencing Day V
Cambridge Computational Biology Institute
Centre for Mathematical Sciences, Wilberforce Road
University of Cambridge, Cambridge.
March 9, 2013

Women in Science Day
University of Hertfordshire, Herts, UK.
August, 2014

REFERENCES

Abnizova, I., Foco, L., te Boekhorst, R., Bernardinelli, L. (2007). Sequence-oriented epidemiology: Regulatory SNPs associated with disease can be inferred by DNA sequence information directly. [Unpublished Work].

Abnizova, I. and Gilks, W.R. (2007). Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in vertebrate genomes. *Brief Bioinformatics, 7*(1), 48–54.

Abnizova, I., Subhankulova, T., Gilks, W. (2007). Recent Computational Approaches to Understand Gene Regulation: Mining Gene Regulation In Silico. *Current Genomics, 8*(2), 79–91.

ADA (American Diabetes Association). (2015). Type 1 Diabetes. [Online]. Available from:
http://www.diabetes.org/diabetes-basics/type-1/?referrer=https://www.google.co.uk/.
[Accessed 13 June, 2015].

Aguado, B. and Campbell, R.D. (1998). Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class III region of the human major histocompatibility complex. *The Journal of Biological Chemistry, 273*(7), 4096–4105. doi:10.1074/jbc.273.7.4096.

Altshuler, D., Daly, M.J., Lander, E.S. (2008). Genetic mapping in human disease. *Science, 322*(5903), 881 – 888.

Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., … Odeberg, J. (2008). In Silico Detection of Sequence Variations Modifying Transcriptional Regulation. PLoS *Computational Biology, 4*(1), e5. doi:10.1371/journal.pcbi.0040005.

Apanius, V., Penn, D. Slev, P.R., Ruff, L.R., Potts, W.K. (1997). The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology, 17*(2), 179–224.

Arney, K.L. (2003). H19 and Igf2 – enhancing the confusion? *Trends in genetics, 19*(1), 17–23.

Asad, S., Nikamo, P., Törn, C., Landin-Olsson, M., Lernmark, A., Alarcón-Riquelme, M., Kockum, I.; Diabetes Incidence in Sweden Study Group. (2007). No evidence of association of the PDCD1 gene with Type 1 diabetes. *Diabetes Medication, 24*(12), 1473-1477.

Atambaeva S.A., Ivashchenko A.T., Khailenko V., Boldina G., Turmagambetova A. (2006). Length of exons and introns in genes of some human chromosomes. *Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure*, Volume 1. Novosibirsk, Russia. July 16-22, 2006.

Auer, S., Hahne, P., Soyal, S.M., Felder, T., Miller, K., Paulmich, M., Kremple, F., Oberkofler, H., Patsch, W. (2012). Potential Role of Upstream Stimulatory Factor 1 Gene Variant in Familial Combined Hyperlipidemia and Related Disorders. *Arteriosclerosis, Thrombosis, and Vascular Biology, 32*, 1535-1544.

Ayyoub, M., Hesdorffer, C. S., Montes, M., Merlo, A., Speiser, D., Rimoldi, D., … Valmori, D. (2004). An immunodominant SSX-2–derived epitope recognized by CD4+ T cells in association with HLA-DR. *Journal of Clinical Investigation, 113*(8), 1225–1233. doi:10.1172/JCI200420667.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Bulyk, M. L. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. *Science, 324*(5935), 1720–1723. doi:10.1126/science.1162327.

Bailey, T.L. and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology, 3*, 21–29.

Baker, K.E. and Parker, R. (2004). Nonsense-mediated mRNA decay: Terminating erroneous gene expression. *Current. Opinions.in Cell Biology, 16*, 293–299.

Balik, V., Srovnal, J., Sulla, I., Kalita, O., Foltanova, T., Vaverka, M., Hrabalek, L., Hajduch, M. (2013). MEG3: a novel long noncoding potentially tumour-suppressing RNA in meningiomas. *Journal of Neuro-Oncology, 112*(1), 1-8.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics, 12*, pp. 745-755.

Barber, R.C. (2012). The genetics of Alzheimer's disease. *Scientifica, 2012* (246210), 1-14. doi:10.6064/2012/246210.

Barreiro, L., Laval, G., Quach, H., Patin, E., Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics, 40*, 340-345.

Barrett, J. C., Clayton, D., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., … the Type 1 Diabetes Genetics Consortium. (2009). Genome-wide association study and meta-analysis finds over 40 loci affect risk of type 1 diabetes. *Nature Genetics, 41*(6), 703–707. doi:10.1038/ng.381.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell, 136*, 215-233.

Beavis W.D. (1998) "*QTL analyses: power, precision, and accuracy*". In: Paterson, A.H. (ed), Molecular analysis of complex traits. CRC Press, Boca Raton, pp 145–161.

Becker, P.B. (2011).The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology, 9*(4), e1001046. doi:10.1371/journal.pbio.1001046.

Beka, N. (2012). *Web Interface Development for a Biological Database.* Unpublished BSc. Project report. University of Hertfordshire, UK.

Beijersbergen, R.L., Kerkhoven, R.M., Zhu, L., Carlée, L., Voorhoeve, P.M., Bernards, R. (1994). E2F-4, a new member of the E2F gene family, has oncogenic activity and associates with p107 in vivo. *Genes and Development, 8*(22), 2680–2690. doi:10.1101/gad.8.22.2680.

Bernardo, A. S., Hay, C. W. and Docherty, K. (2008) Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic β cell. *Molecular and Cellular Endocrinology, 294*, 1-9.

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Peña-Castillo, L, Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., Khalid, F., Zhang, W., Newburger, D., Jaeger, S.A., Morris, Q.D., Bulyk, M.L., Hughes, T. R. (2008). Variation in homeodomain DNA-binding revealed by high-resolution analysis of sequence preferences. *Cell, 133*(7), 1266–1276. doi:10.1016/j.cell.2008.05.024.

Berger, M. F., Philippakis, A. A, Qureshi, A.M., He, F.S., Estep, P.W., Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology, 24*(11):1429-1435.

Bertram, L. and Tanzi, R.E. (2009). Genome-wide association studies in Alzheimer's disease. Human *Molecular Genetics, 18*(R2), pp. R137-R145.

Bertram, L. and Tanzi, R.E. (2012). The genetics of Alzheimer's disease. *Molecular Biology of Neurodegenerative Diseases, 107*, 79-100. doi: 10.1016/B978-0-12-385883-2.00008-4.

Bierhaus, A., Schiekofer, S., Schwaninger, M., Andrassy, M., Humpert, P.M., Chen, J., Hong, M., Luther, T., Henle, T., Klöting, I., Morcos, M., Hofmann, M., Tritschler, H., Weigle, B., Kasper, M., Smith, M., Perry, G., Schmidt, A.M., Stern, D.M., Häring, H.U., Schleicher, E., Nawroth, P.P. (2001). Diabetes-associated sustained activation of the transcription factor nuclear factor-kappaB. *Diabetes, 50*(12), 2792–808. doi:10.2337/diabetes.50.12.2792.

Birney, E., ENCODE Project Consortium, Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder M, Dermitzakis, E.T., Thurman, R.E.,… de Jong, P.J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447*(7146):799-816.

Black, D.L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Reviews Biochemistry, 72*, 291-336.

Blau, C.A. and Liakopoulou, E. (2013). Can we deconstruct cancer, one patient at a time? *Trends in Genetics, 29*(1), 6–10.

Bluestone, J. A. and Bour-Jordan, H. (2012). Current and Future Immunomodulation Strategies to Restore Tolerance in Autoimmune Diseases. *Cold Spring Harbor Perspectives in Biology, 4*(11), a007542. doi:10.1101/cshperspect.a007542.

Bluestone, J.A., Herold, K., Eisenbarth. G. (2010).Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature, 464*(7293), 1293-300.

Bonaldo, M.F., Lennon, G., Soares, M.B. (1997). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research, 6*(9), 791–806. doi:10.1101/gr.6.9.791.

Boonsaen, T., Rojvirat, P., Surinya, K. H., Wallace, J. C., Jitrapakdee, S. (2007). Transcriptional regulation of the distal promoter of the rat pyruvate carboxylase gene by hepatocyte nuclear factor 3β/Foxa2 and upstream stimulatory factors in insulinoma cells. *Biochemical Journal, 405*(Pt 2), 359–367. doi:10.1042/BJ20070276.

Boyer, S., Brown, S. D. J., Collins, R. A., Cruickshank, R. H., Lefort, M. C., Malumbres-Olarte, J., Wratten, S.D. (2012). Sliding Window Analyses for Optimal Selection of Mini-Barcodes, and Application to 454-Pyrosequencing for Specimen Identification from Degraded DNA. *PLoS ONE, 7*(5), e38215.

Borges, E., Pendl, G., Eytner, R., Steegmaier, M., Zo¨llner, O.,Vestweber, D. (1997). The Binding of T Cell-expressed P-selectin Glycoprotein Ligand-1 to E- and P-selectin Is Differentially Regulated. *The Journal of Biological Chemistry, 272*(45), 28786–28792.

Bosma, P.J., Chowdhury, J.R., Bakker, C., Gantla, S., de Boer, A., Oostra, B.A., Lindhout, D., Tytgat, G.N., Jansen, P.L.,... Oude Elferink, R.P.(1995). The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *New England Journal of Medicine*, *333*(18), 1171-1175.

Boyle, A.P., Hong, E. L., Hariharan, M, Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., Snyder, M. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*: 1790-1797.

Bradfield, J.P., Qu, H.Q., Wang, K., Zhang, H., Sleiman, P.M., Kim, C.E., Mentch, F.D., Qiu, H., Glessner, J.T., Thomas, K.A., Frackelton, E.C., Chiavacci, R.M., Imielinski, M., Monos, D.S., Pandey, R., Bakay, M., Grant, S.F., Polychronakos, C., Hakonarson, H. (2011). A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. *PLOS Genetics*, *7*(9):e1002293.

Brandon, L.P. and Ahsan, H. (2011). Genome-wide “Pleiotropy Scan” Identifies HNF1A Region as a Novel Pancreatic Cancer Susceptibility Locus. *Cancer Research*, *71*(13), 4352–4358.

Brem, R.B., Yvert, G., Clinton, R., Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, *296*, 752–755.

Broadbent, H.M., Peden, J.F., Lorkowski, S., Goel, A., Ongen, H., Green, F.,... PROCARDIS consortium. (2008). Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Human Molecular Genetics*, *17*, 806–814.

Brown, T. A (2010). *Gene Cloning and DNA Analysis: An Introduction.* West Sussex: Wiley-Blackwell.

Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B. Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research*, *36*(Database issue), D102–D106. doi:10.1093/nar/gkm955.

Burren, O. S., Adlem, E. C., Achuthan, P., Christensen, M., Coulson, R. M., Todd, J. A. (2011). T1DBase: update 2011, organization and presentation of large-scale data sets for type 1 diabetes research. *Nucleic Acids Research*, 39(Database Isssue), D997-1001.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson A, Kwiatkowski Dp, Mccarthy Mi, Ouwehand Wh, Samani Nj, Todd, J. A., Donnelly, P., Barrett, J. C., Stratton, M. R., Worthington, J., Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661-78.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. PLOS *Computational Biology*, *8*, e1002822.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, *25*(18), 1915–1927. doi:10.1101/gad.17446611.

Cano, P; Klitz W, Mack SJ, Maiers M, Marsh SG, Noreen H, Reed EF, Senitzer D, Setterholm M, Smith A, Fernández-Viña M. (2007). Common and well-documented HLA alleles: report of the Ad-Hoc committee of the American Society for Histocompatiblity and Immunogenetics. *Human Immunology*, *68* (5): 392–417.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q., Lander, E.S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, *22*(3), 231-238.

Cavallaria, N., Balestraa, D., Branchinia, A., Maestrib, I., Chuamsunritc, A., Sasanakulc, W., Marianid, G., Paganie, F., Bernardia, F., Pinottia, M. (2012). Activation of a cryptic splice site in a potentially lethal coagulation defect accounts for a functional protein variant. *Molecular Basis of Disease*, *1822*(7), 1109–1113.

Chan, H. P., Zhang, N. R., Chen, L. H. Y. (2010). Importance Sampling of Word Patterns in DNA and Protein Sequences. *Journal of Computational Biology*, *17*(12), 1697–1709. doi:10.1089/cmb.2008.0233.

Changrani, N. R., Chonkar, A., Adeghate, E., Singh, J. (2006). Effects of Streptozotocin-Induced Type 1 Diabetes Mellitus on Total Protein Concentrations and Cation Contents in the Isolated Pancreas, Parotid, Submandibular, and Lacrimal Glands of Rats. *Annals of the New York Academy of Sciences*, *1084*, 503–519. doi: 10.1196/annals.1372.019.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction (6th Ed.).* London: Chapman & Hall/CRC.

Chen, C.Y., Chang, I.S., Hsiung, C.A., Wasserman, W.W. (2014). On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Medical Genomics*, *7*(34). doi: 10.1186/1755-8794-7-34.

Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. (2013). LincRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Research*, *(Database issue*), D983-986.

Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K., Botstein, D. (1997). Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, *387*(6632), 67–73.

Chiang, J. L., Kirkman, M. S., Laffel, L. M. B., Peters, A. L. (2014). Type 1 Diabetes through the life span: A position statement of the American Diabetes Association. *Diabetes Care*, *37*(7), 2034–2054. doi:10.2337/dc14-1140.

Choi, J.K., Bae, J.B., Lyu, J., Kim, T.Y., Kim, Y.J. (2009). Nucleosome deposition and DNA methylation at coding region boundaries. *Genome Biology*, *10*(9), R89.

Cirulli, E.T., Singh, A, Shianna, K.V., Dongliang, G., Smith, J.P., Maia, J.M., Heinzen, E.L., Goedert, J.J., Goldstein, D.B., Centre for HIV/AIDS Vaccine Immunology (CHAVI). (2010). Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biology*, *11*(*5*):R57.

Close, P., Hawkes, N., Cornez, I., Creppe, C., Lambert, C.A., Rogister, B., Siebenlist, U., Merville, M.P., Slaugenhaupt, S.A., Bours, V., Svejstrup, J.Q., Chariot, A.(2006). Transcription impairment and cell migration defects in elongator-depleted cells: implication for familial dysautonomia. *Molecular Cell*, *22*(4), 521-31.

Corre, S. and Galibert, M.D. (2006). USF as a key regulatory element of gene expression. *Médecine Sciences Paris*, *22* (1), 62–67. doi:10.1051/medsci/200622162.

Craig, J. (2008). Complex diseases: Research and applications. *Nature Education, 1*(1):184.

Cockerill, P.N. (2011). Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS Journal, 278*(13), 2182-2210.

Collins, F. (2010). Has the revolution arrived? *Nature, 464*(7289), 674–675.

Collins, F.S., Morgan, M., Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science, 300*, 286-290.

Concannon, P., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Pociot, F., Todd, J.A., Rich, S.S.; Type 1 Diabetes Genetics Consortium. (2005). Evidence for Susceptibility Loci from Four Genome-Wide Linkage Scans in 1,435 Multiplex Families. *Diabetes, 54*(10), 2995-3001.

Cooper, G.M. (2000). *The Cell: A Molecular Approach.* 2nd edition. Chapter 4.2 Chromosomes and Chromatin. Sunderland, Massachusetts: Sinauer Associates. [Online]. Available at: http://www.ncbi.nlm.nih.gov/books/NBK9863/.

Cooper, G.M. and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Review Genetics, 12*(9), 628-640.

Corre, S. and Galibert, M.D. (2006).USF as a key regulatory element of gene expression. *Medical Science (Paris), 22*(1), 62–67.

Cowper-Sal·lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., Moore, J.H., Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics, 44*(11), 1191–1198. doi:10.1038/ng.2416.

Craig, J. (2008). Complex diseases: Research and applications. *Nature Education, 1*(1), 184.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., … Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research, 16*(1), 123–131. doi:10.1101/gr.4074106.

Cudworth, A. and Woodrow, J. (1974). Letter: HL-A antigens and diabetes mellitus. *Lancet, 2*(7889), 1153.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.… Flicek, P. (2014), Ensembl 2015. *Nucleic Acids Research, 42*(22). doi: 10.1093/nar/gku1010.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.… Flicek, P. (2015), Ensembl 2015. *Nucleic Acids Research, 43*(D1): D662-D669. doi: 10.1093/nar/gku1010.

Dame, R.T. (2005). The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular Microbiology, 56*(4): 858–870. doi:10.1111/j.1365-2958.2005.04598.x.

Davey, S.G. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology, 33*(1), pp. 30-42.

David, S. J., Mortazavi, A., Myers, R. M., Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science, 316* (5830), 1497-1502. doi:10.1126/science.1141319.

De Bie, P., Muller, P., Wijmenga, C., Klomp, L. W. J. (2007). Molecular pathogenesis of Wilson and Menkes disease: correlation of mutations with molecular defects and disease phenotypes. *Journal of Medical Genetics*, *44*(11), 673–688. doi:10.1136/jmg.2007.052746.

Delves, P. J. (2004). *Encyclopedia of Immunology*. London: Elsevier Press. pp 292 - 296.

Deng, G. and Sui, G. (2013). Noncoding RNA in oncogenesis: a new era of identifying key players. *International Journal of Molecular Sciences*, *14*(9), 18319-18349.

Deschamps, I., Lestradet, H., Bonaiti, C., Schmid, M., Busson, M., Benajam, A., Marcelli-Barge, A., Hors, J. (1980). HLA genotype studies in juvenile insulin-dependent diabetes. *Diabetologia*, *19*, 189–193.

de Vooght, K.M., van Wijk, R., van Solinge, W.W. (2009). Management of gene promoter mutations in molecular diagnostics. *Clinical Chemistry*, *55*(4), 698–708.

Di Gesualdo, F., Capaccioli, S., Lulli, M. (2014). A pathophysiological view of the long non-coding RNA world. *Oncotarget*, *5*(22), 10976–10996.

Diabetes UK. (2015). What is Type 1 diabetes? [Online]. Available from: https://www.diabetes.org.uk/Guide-to-diabetes/What-is-diabetes/What-is-Type-1-diabetes/. [last accessed: 13/06/2015].

Diggs, L. W., Ahmann, C. F., Bibb, J. (1933). The incidence and significance of the sickle-cell trait. *Annals of Internal Medicine*, *7*, 769–778.

Dineen, D., Schröder, M., Higgins, D. Cunningham, P. (2010). Ensemble approach combining multiple methods improves human transcription start site prediction. *BMC Genomics*, *30*(11), 677.

Dirk, S. P., Soranzo, N., Beck, S. (2014). Functional interpretation of non-coding sequence variation: Concepts and challenges. *Bioessays*, *36*, 191-199.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K.,…Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*, 101-108.

Dudley, J. and Karczewski, K. (2014). *Exploring Personal Genomics.* Oxford: Oxford University Press.

Duncan, E., Brown, M., Shore, E. M. (2014). The Revolution in Human Monogenic Disease Mapping. *Genes*, *5*(3), 792–803. doi:10.3390/genes5030792.

Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., Cherry, J.M. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research*, *30*(1), 69–72. doi:10.1093/nar/30.1.69.

Edwards, D., Stajich, J., Hansen, D. (Eds). (2009). *Bioinformatics: Tools and Applications.* Springer: New York. pp 145-146.

Elsik, C.G., Tellam, R.L., Worley, K.C., The Bovine Genome Sequencing and Analysis Consortium. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, *324*(5926), 522–528. doi:10.1126/science.1169588.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., Bernstein, B. E. (2011). Systematic analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43–49. doi:10.1038/nature09906.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., … Figeys, D. (2007). Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular Systems Biology, 3*(1), 89. doi:10.1038/msb4100134.

Faghihi, M.A., Modarresi, F., Khalil, A.M., Wood, D.E., Sahagan, B.G., Morgan, T.E., Finch, C.E., St. Laurent, G., Kenny, P.J., Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase expression. *Nature Medicine, 14*(7): 723–730. doi: 10.1038/nm1784.

Fairweather, D. and Rose, N.R. (2002). Type 1 diabetes: virus infection or autoimmune disease? *Nature Immunology, 3*(4), 338–340. doi:10.1038/ni0402-338.

Fan, Y.-M., Hernesniemi, J., Oksala, N., Levula, M., Raitoharju, E., Collings, A., Hutri-Kähönen, N., Juonala, M., Marniemi, J., Lyytikäinen, L.P.,… Lehtimäki, T. (2014). Upstream Transcription Factor 1 (USF1) allelic variants regulate lipoprotein metabolism in women and USF1 expression in atherosclerotic plaque. *Scientific Reports, 4*, 4650. doi:10.1038/srep04650.

Frasier, S.D., Bashore, S.D., Bashore, R.A., Mosier, H.D. (1964). Gonadoblastoma assosicated with pure gonadal dysgenesis in monozygous twins. *Journal of Paediatrics, 64*(5), 740–7455. doi:10.1016/S0022-3476(64)80622-3.

Felsenfeld G. and Groudine M. (2003). Controlling the double helix. *Nature, 421*(6921), 448-453.

Fink, G.A. (2007). *Markov Models for Pattern Recognition: From Theory to Applications.* Springer: New York.

Flanagan, S.E., Xie, W., Caswell, R., Damhuis, A., Vianey-Saban, C., Akcay, T., Darendeliler, F., Bas, F., Guven, A., Siklar, Z.,… Ellard, S. (2013) Next-generation sequencing reveals deep intronic cryptic ABCC8 and HADH splicing founder mutations causing hyperinsulinism by pseudoexon activation. *American Journal of Human Genetics, 92*:131–136. doi: 10.1016/j.ajhg.2012.11.017.

Frischmeyer, P.A. and Dietz, H.C. (1999). Nonsense-mediated mRNA decay in health and disease. *Human Molecular Genetics, 8*(10), 1893-900.

Fung, K. L. and Gottesman, M. M. (2009). A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. *Biochimica et Biophysica Acta, 1794*(5), 860–871. doi:10.1016/j.bbapap.2009.02.014.

Gabriel, A. and Przybylski, J. (2010) Sickle-cell anemia: A Look at Global Haplotype Distribution. *Nature Education, 3*(3), 2.

Gagliano, S.A., Barnes, M.R., Weale, M.E., Knight, J. (2014). A Bayesian method to incorporate hundreds of functional characteristics with association evidence to improve variant prioritization. *PLoS One, 9*(5), e98122.

Galanello, R. and Origa, R. (2010). Beta-thalassemia. *Orphanet Journal of Rare Diseases, 5*, 11. doi:10.1186/1750-1172-5-11.

Gale, E.A.M. and Gillespie, K. (2014). Genetics of type 1 diabetes. Diapedia, 21040851211 rev. no. 59. [Online]. Available at: http://dx.doi.org/10.14496/dia.21040851211.59. [Accessed: 14 December, 2014].

García-Sanz, P., Fernández-Pérez, A., Vallejo, M. (2013). Differential configurations involving binding of USF transcription factors and Twist1 regulate Alx3 promoter activity in mesenchymal and pancreatic cells. *Biochemical Journal*, *450*(1), 199-208. doi: 10.1042/BJ20120962.

Garfield, D., Haygood, R., Nielsen, W., Wray, G. (2012). Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin Strongylocentrotus purpuratus. *Evolution and Development*, *14*(2), 152-167. doi: 10.1111/j.1525-142X.2012.00532.x.

Geddes, L. (30th October 2013) 'Bubble kid' success puts gene therapy back on track' The New Scientist, magazine issue 2941. [Online]. Available from:
http://www.newscientist.com/article/mg22029413.200-bubble-kid-success-puts-gene-therapy-back-on-track.html#.VXzdcvnBzRY. [Accessed 14 June, 2015].

GeneED (Genetics, Education, Discovery). (2012). Gene therapy. [Online]. Available at:
http://geneed.nlm.nih.gov/topic_subtopic.php?tid=142&sid=144. [Accessed 13 June, 2015].

Gerstein, M., Bruce, C.,Rozowsky, J.S., Zheng, D., Du, J.,Korbel, J.,... Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, *17*, 669-681.

Gesualdo, F. D., Capaccioli, S., Lulli, M. (2014). A pathophysiological view of the long non-coding RNA world. *Oncotarget*, *5*(22), 10976–10996.

Gilliam, L.K., Palmer, J.P., Lernmark, Å. (2004). Autoantibodies and the Disease Process of Type 1 Diabetes Mellitus. In: LeRoith, D., Taylor, S.I., Olefsky, J.M., eds. (2004) *Diabetes Mellitus: A Fundamental and Clinical Text*, 3rd Edition. pp. 500-18, Philadelphia: Lippincott Williams & Wilkins.

Gillespie, K. (2014). Non-HLA genes. *Diapedia*, 2104311115(36). [Online]. Available at: http://dx.doi.org/10.14496/dia.2104311115.36. [Accessed 04 January, 2015].

Gillespie, K. and Owen, K. (2014). IL2RA. *Diapedia*, 2104135143(12). [Online]. Available at: http://dx.doi.org/10.14496/dia.2104135143.12. [Accessed 02 February, 2015].

Ginsberg, D., Vairo, G., Chittenden, T., Xiao, Z.X., Xu, G., Wydner, K.L., DeCaprio, J.A., Lawrence, J.B., Vingston, D.M. (1994). E2F-4, a new member of the E2F transcription factor family, interacts with p107. *Genes and Development*, *8*(22), 2665–2679. doi:10.1101/gad.8.22.2665.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, *17*(6), 877–885. doi:10.1101/gr.5533506.

Glazier, A.M., Nadeau, J.H., Aitman, T.J. (2002) Finding genes that underlie complex traits. *Science*, *298*, 2345–2349.

Gonzalez-Galarza, F.F.; Mack, S.J., Hollenbach, J., Fernandez-Vina, M., Setterholm, M., Kempenich, J., Marsh, S.G., Jones, A.R., Middleton, D., HLA Rare Allele Consortium. (2013). 16th International HLA and Immunogenetics Workshop (IHIW): extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles". *International Journal of Immunogenetics, 40* (1): 60–65.

Gregory, T.R. (2005). Synergy between sequence and size in large-scale genomics. Nature Review *Genetics, 6*(9), 699–708. doi:10.1038/nrg1674.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., Enright, A.J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research, 36*(Database issue), D154–D158.

Gross, D.S. and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry, 57*, 159-97.

Guhathakurta, D., 2006. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Research, 34*(12), 3585-3598.

Guo, Y. and Jamison, D. C. (2005). The distribution of SNPs in human gene regulatory regions. *BMC Genomics, 6*, 140. doi:10.1186/1471-2164-6-140.

Guttmacher, A.E. and Collins, F.S. (2002). Genomic Medicine - A Primer. *The New England Journal of Medicine, 347*(19), 1512- 1520.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R, Bruhn, L.,… Lander, E.S. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature, 477*: 295–300. doi: 10.1038/nature10398.

Handel, A. E., Sandve, G. K., Disanto, G., Handunnetthi, L., Giovannoni, G., Ramagopalan, S. V. (2013). Integrating multiple oestrogen receptor alpha ChIP studies: overlap with disease susceptibility regions, DNase I hypersensitivity peaks and gene expression. *BMC Medical Genomics, 6*, 45. doi:10.1186/1755-8794-6-45.

Hastings, M.L., Resta, N., Traum, D., Stella, A., Guanti, G., Krainer, A.R. (2005). An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nature Structural & Molecular Biology, 12*, 54-59.

Held W, Kunz B, Lowin-Kropf B, van de Wetering M, Clevers H (1999). Clonal acquisition of the Ly49A NK cell receptor is dependent on the trans-acting factor TCF-1. *Immunity, 11*, 433–442.

Heras, P., Mantziorosa, M., Mendrinosa, D., Herasa, V., Hatzopoulosa, A., Xourafasa, V., Kritikosa, K., Karagiannisa, S. (2010). Autoantibodies in Type 1 Diabetes. *Diabetes Research and Clinical Practice, 90*, 2, e40 - e42.

Herranz, H. and Cohen, S.M. (2010). MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. *Genes & Development, 24*(13), 1339–1344.

Heyd, F (2014). Alternative splicing--principles, functional consequences and therapeutic implications, [Article in German]. *Deutsche Medizinische Wochenschrift, 139*(7), 339-42.

Hindorff, L.A., MacArthur, J., Morales, J., Junkins, H.A., Hall, P.N., Klemm, A.K., Manolio, T.A.. *A Catalog of Published Genome-Wide Association Studies.* Available at: www.genome.gov/gwastudies. [Accessed Feb. 02, 2015].

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367. doi:10.1073/pnas.0903103106.

Hirschhorn, J.N, Lohmueller, K., Byrne, E., Hirschhorn, K. (2002). A comprehensive review of genetic association studies, *Genetics in Medicine*, *4*, 45–61.

Hoffmeyer, S., Burk, O., von Richter, O., Arnold, H. P., Brockmöller, J., Johne, A., … Brinkmann, U. (2000). Functional polymorphisms of the human multidrug-resistance gene: Multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(7), 3473–3478.

Hogan, G.J., Lee, C.K., Lieb, J.D. (2006). Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. *PLoS Genetics*, *2*(9), e158.

Homolova, K., Zavadakova, P., Doktor, T. K., Schroeder, L. D., Kozich, V., Andresen, B. S. (2010). The Deep Intronic c.903+469T>C Mutation in the MTRR Gene Creates an SF2/ASF Binding Exonic Splicing Enhancer, Which Leads to Pseudoexon Activation and Causes the cblE Type of Homocystinuria. *Human Mutation*, *31*(4), 437–444.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Krieger, C.J., Livstone, M.S., Miyasato, S.R., Nash, R.S., Oughtred, R., Skrzypek, M.S., Weng, S., Wong, E.D., Zhu, K.K., Dolinski, K., Botstein, D., Cherry, J. M. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Research*, *36*(Database issue), D577–D581. doi:10.1093/nar/gkm909.

Hoogendoorn, B., Coleman, S.L., Guy, C.A., Smith, S.K., O'Donovan, M.C., Buckland, P.R. (2004).Functional analysis of polymorphisms in the promoter regions of genes on 22q11. *Human Mutation*, *24*, 35–42.

Hoogendoorn, B., Coleman, S.L., Guy, C.A., Smith, S.K., Bowen, T., Buckland, P.R., O'Donovan, M.C. (2003). Functional analysis of human promoter polymorphisms. *Human Molecular Genetics*, *12*, 2249–2254.

Huang, H., Kao, M., Zhou, X., Liu, J.S., Wong, W.H. (2004). Determination of local statistical significance of patterns in Markov sequences with applications to promoter element identification. *Journal of. Computational Biology*, *11*(1), 1–14.

Houslay, M.D., Schafer, P., Zhang, K.Y. (2006). Keynote review: phosphodiesterase-4 as a therapeutic target. *Drug Discovery Today*, *10*(22), 1503–1519. doi:10.1016/S1359-6446(05)03622-6.

Howe, E. D. and Song, J. S. (2013). Categorical spectral analysis of periodicity in human and viral genomes. *Nucleic Acids Research*, *41*(3), 1395–1405. doi:10.1093/nar/gks1261.

Howson, J.M.M., Walker, N.M., Smyth, D.J. Todd, J.A., the Type I Diabetes Genetics Consortium, (2009). Analysis of 19 genes for association with type I diabetes in the Type I Diabetes Genetics Consortium families. *Genes and Immunity*, *10*(1), S74–S84.

Hrdlickova, B., Westra, H.J., Franke, L., Wijmenga, C. (2011). Celiac disease: moving from genetic associations to causal variants. *Clinical Genetics, 80*(3), 203-313. doi: 10.1111/j.1399-0004.2011.01707.x.

Hudson, B.I., Hofman, M.A., Bucciarelli, L., Wendt, T., Moser, B., Wu Qu, Y.L., Stern, D.M., D'Agati, V., Yan, S. D., Yan, S. F., Grant, P. J., Schmidt, A. M. (2002). Glycation and diabetes: The RAGE connection. *Current Science, 83*(12), 1515–1521.

Hudson, B.I., Stickland, M.H., Futers, T.S., Grant, P.J. (2001). Effects of novel polymorphisms in the RAGE gene on transcriptional regulation and their association with diabetic retinopathy. *Diabetes 50*(6), 1505–1511. doi:10.2337/diabetes.50.6.1505.

Hurley, C.K. (1997). DNA-based typing of HLA for transplantation. In: Leffell, M.S., Donnenberg, A.D. and Rose, N.R.., eds. (1997) *Handbook of Human Immunology*. pp. 521-55, Boca Raton: CRC Press.

Imkampe, A.K. and Gulliford, M.C. (2011). Trends in Type 1 diabetes incidence in the UK in 0- to 14-year-olds and in 15- to 34-year-olds, 1991-2008. *Diabetic Medicine, 28*(7), 811-814.

International Human Genome Sequencing Consortium (IHGSC). (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931–945.

Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M., Saito, S., Nakamura, Y., Tanaka, T. (2006). Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *Journal of Human Genetics, 51*, 1087–1099.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology, 3*, 318–356. doi:10.1016/S0022-2836(61)80072-7.

Janeway, C.A. Jr, Travers, P., Walport, M., Shlomchik, M.J. (2001). *Immunobiology: The Immune System in Health and Disease*. The production of armed effector T cells. 5th edition. New York: Garland Science.

Jensen, L.J. and Knudsen, S. (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics, 16*, 326–333.

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B. (2007). Genome-wide mapping of in vivo protein–DNA interactions. *Science, 316*, (5830), 1497–1502.

Johnson, R. (2012). Long non-coding RNAs in Huntington's disease neurodegeneration. *Neurobiology Disease, 46*(2):245-54.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research, 36*(16), 5221-5231. doi: 10.1093/nar/gkn488.

Kainulainen, K., Karttunen, L., Puhakka, L., Sakai, L., Peltonen, L.(1994). Mutations in the fibrillin gene responsible for dominant ectopia lentis and neonatal Marfan syndrome. *Nature Genetics*, (1),64-69.

Kapustin, Y., Chan, E., Sarkar, R., Wong, F., Vorechovsky, I., Winston, R.M., Tatusova, T.,Dibb, N. J. (2011). Cryptic splice sites and split genes. *Nucleic Acids Research, 39*(14), 5837-5844.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., … Snyder, M. (2010). *Variation in Transcription Factor Binding Among Humans. Science (New York, N.Y.), 328*(5975), 232–235. doi:10.1126/science.1183621.

Kendall, E., Sargent, C.A., Campbell, R.D. (1991). Human major histocompatibility complex contains a new cluster of genes between the HLA-D and complement C4 loci. *Nucleic Acids Research, 18*(24), 7251–7257. doi:10.1093/nar/18.24.7251.

Kersey, P. J., Allen, J. E., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Hughes, D. S., Humphrey, J., Kerhornou, A., Khobova, J., Langridge, N., Mcdowall, M. D., Maheswari, U., Maslen, G., Nuhn, M., Ong, C. K., Paulini, M., Pedro, H., Toneva, I., Tuli, M. A., Walts, B., Williams, G., Wilson, D., Youens-Clark, K., Monaco, M. K., Stein, J., Wei, X., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D., Staines, D. M. (2014). Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Research, 42*, D546-D552.

Kersey, P. J., Staines, D. M., Lawson, D., Kulesha, E., Derwent, P., Humphrey, J. C., Hughes, D. S., Keenan, S., Kerhornou, A., Koscielny, G., Langridge, N., Mcdowall, M. D., Megy, K., Maheswari, U., Nuhn, M., Paulini, M., Pedro, H., Toneva, I., Wilson, D., Yates, A., Birney, E. (2012). Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Reearch, 40*, D91-D97.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., … Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America, 106*(28), 11667–11672. doi:10.1073/pnas.0904715106.

Kim, A., Song, S.H., Brand, M., Dean, A. (2007). Nucleosome and transcription activator antagonism at human beta-globin locus control region DNase I hypersensitive sites. *Nucleic Acids Research, 35*(17), 5831-5838.

Kim, B. C., Kim, W.Y., Park, D., Chung, W.-H., Shin, K., Bhak, J. (2008). SNP@Promoter: a database of human SNPs (Single Nucleotide Polymorphisms) within the putative promoter regions. *BMC Bioinformatics, 9*(1), S2. doi:10.1186/1471-2105-9-S1-S2.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M.Q., Lobanenkov, V.V, Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF binding sites in the human genome. *Cell, 128*(6), 1231–1245. doi:10.1016/j.cell.2006.12.048.

Kim, T. H. and Ren, B. (2006). Genome-Wide Analysis of Protein-DNA Interactions. *Annual Review of Genomics and Human Genetics, 7*, 81-102.

Kimchi-Sarfaty, C., Marple, A. H., Shinar, S., Kimchi, A. M., Scavo, D., Roma, M. I., … Gottesman, M. M. (2007). Ethnicity-Related Polymorphisms and Haplotypes in the Human ABCB1 *Gene. Pharmacogenomics, 8*(1), 29–39. doi:10.2217/14622416.8.1.29.

King, M.C. and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science, 188*, 107–116.

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation, 2011*, bar030. doi:10.1093/database/bar030.

Klamt, B., Koziell, A., Poulat, F., Wieacker, P., Scambler, P., Berta, P., Gessler, M. (1998). Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms. *Human Molecular Genetics*, *7*(4), 709–714. doi:10.1093/hmg/7.4.709.

Knight, J.C. (2010). Understanding human genetic variation in the era of high-throughput sequencing. *EMBO Reports*, *11*(9), 650-652.

Knight, J. C. (2014). Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Medicine*, *6*(10), 92. doi:10.1186/s13073-014-0092-4.

Knip, M. and Siljandera, H. (2008). Autoimmune mechanisms in type 1 diabetes. *Autoimmunity Reviews*, *7*, 550-557.

Knip, M., Veijola, R., Virtanen, S.M., Hyöty, H., Vaarala, O., Akerblom, H.K. (2005). Environmental Triggers and Determinants of Type 1 Diabetes. *Diabetes*, *54*, S125–S136. doi:10.2337/diabetes.54.suppl_2.S125.

Komar, A.A. (2007). Silent SNPs; impact on gene function and phenotype, *Pharmacogenomics*, *8*, 1075-1080.

Kornberg, R.D. (1974). Chromatin Structure: A Repeating Unit of Histones and DNA. *Science*, *24*, 868-871.

Kuo, H.W., Huang, C.Y., Fu, C.K., Liao, C.H., Hsieh, Y.H., Hsu, C.M., Tsai, C.W., Chang, W.S., Bau, D.T. (2014). The significant association of CCND1 genotypes with gastric cancer in Taiwan. *Anticancer Research*, *34*(9), 4963-4968.

Knowling, S. and Morris, K.V. (2011). Epigenetic Regulation of Gene Expression in Human Cells by Noncoding RNAs. *Progress in Molecular Biology and Translational Science*, *102*, 1–10.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. *Nature*, *470* (7333), 187–197.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.,… International Human Genome Sequencing Consortium. (2011). Initial impact of the sequencing of the human genome.. *Nature*, *409* (6822), 860–921.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*, 1813–1831.

Lario, P.I., Bobechko, B., Bateman, K., Kelly, J., Vrielink, A., Huang, Z. (2001). Purification and characterization of the human PDE4A catalytic domain (PDE4A330–723) expressed in Sf9 cells. *Archives of Biochemistry and Biophysics*, *394*(1), 54–60. doi:10.1006/abbi.2001.2513.

Laurila, K. and Lähdesmäki, H. (2009). Systemic Analysis of Disease-Related Regulatory Mutation Classes Reveals Distinct Effects on Transcription Factor Binding. *In silico Biology*, *9*, 209-224.

Laurila, K. and Lähdesmäki, H. (2008). Effects of Disease-Related Mutations on Transcription Factor Binding. In: *Proceedings of the Fifth TICSP Workshop on Computational Systems Biology* (*WCSB* 2008), pp. 89-92.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, *262*, 208–214.

Leung, D.W. (2001). The structure and functions of human lysophosphatidic acid acyltransferases. *Frontiers in Bioscience*, *6*(1): D944–953.

Levit, G.S. and Hoßfeld, U. (2006). The Forgotten "Old-Darwinian" Synthesis: The Evolutionary Theory of Ludwig H. Plate (1862–1937). *NTM International Journal of History & Ethics of Natural Sciences Technology & Medicine*, *14*, 9–25.

Lewin, B. (2008). *Genes IX*. Sudbury: Jones & Bartlett. pp. 213-214.

Lewitt, P. A., Rezai, A. R., Leehey, M. A., Ojemann, S. G., Flaherty, A. W., Eskandar, E. N.,… Feigin, A. (2011). AAV2-GAD gene therapy for advanced Parkinson's disease: A double-blind, sham-surgery controlled, randomised trial. *The Lancet Neurology*, *10*(4), 309–319.

Lim, U., Kocarnik, J. M., Bush, W. S., Matise, T. C., Caberto, C., Park, S. L., … Le Marchand, L. (2014). Pleiotropy of Cancer Susceptibility Variants on the Risk of Non-Hodgkin Lymphoma: The PAGE Consortium. *PLoS ONE*, *9*(3), e89791. doi:10.1371/journal.pone.0089791.

Lina, M., Diaz, G., Martin, J. (2012). PTPN22 splice forms: a new role in rheumatoid arthritis. *Genome Medicine*, *4*(13), doi: 10.1186/gm312.

Long, S. A., Cerosaletti, K., Wan, J. Y., Ho, J.-C., Tatum, M., Wei, S., Shilling, H.G., Buckner, J. H. (2011). An autoimmune-associated variant in PTPN2 reveals an impairment of IL-2R signalling in CD4+ T cells. *Genes and Immunity*, *12*(2), 116–125. doi:10.1038/gene.2010.54.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J. (2000) *Molecular Cell Biology*. 4th Edition. New York: W. H. Freeman.

Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., Guigo, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters*, *579*, 1900-1903.

Lu, Y.F., Goldstein, D.B., Angrist, M., Cavalleri, G. (2014). Personalized medicine and human genetic diversity. *Cold Spring Harbor perspectives in medicine*, *4*(9), a008581.

Maahs, D.M., West, N.A., Lawrence, J.M., Mayer-Davis, E.J. (2010). Epidemiology of Type 1 Diabetes. *Endocrinology and Metabolism Clinics of North America*, *39*(3), 481–497.

MacLeod, M.K.L., Kappler, J.W., Marrack, P. (2010). Memory CD4 T cells: generation, reactivation and re-assignment. *Immunology*, *130*(1), 10–15. doi: 10.1111/j.1365-2567.2010.03260.x.

Mahajan, N., Mahmood, S., Jain, S., Dhawan, V. (2013). Receptor for advanced glycation end products (RAGE), inflammatory ligand EN-RAGE and soluble RAGE (sRAGE) in subjects with Takayasu's arteritis. *International Journal of Cardiology*, *168*(1), 532–34. doi:10.1016/j.ijcard.2013.01.002.

Matlin, A.J., Clark, F., Smith, C.W. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, *6*(5), 386-398.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747–753. doi:10.1038/nature08494.

Marian, A.J. (2012). Molecular genetic studies of complex phenotypes. *Translational Research, 159*(2):64-79. doi: 10.1016/j.trsl.2011.08.001.

Marfan, A. (1896). A case of congenital deformation of the four limbs, more pronounced at the extremities, characterized by elongation of the bones with some degree of thinning, [Article in French]. *Bulletins et memoires de la Société medicale des hôspitaux de Paris, 13*(3), 220–226.

Martin, C. C., Svitek, C. A., Oeser, J. K., Henderson, E., Stein, R., O'Brien, R. M. (2003). Upstream stimulatory factor (USF) and neurogenic differentiation/beta-cell E box transactivator 2 (NeuroD/BETA2) contribute to islet-specific glucose-6-phosphatase catalytic-subunit-related protein (IGRP) gene expression. *Biochemical Journal, 371*(Pt 3), 675–686. doi:10.1042/BJ20021585.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research, 42*, 142-147.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., … Wingender, E. (2006). TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research, 34*(Database issue), D108–D110. doi:10.1093/nar/gkj143.

McDonald, J. (2015). *Handbook of Biological Statistics (3rd Ed).* Baltimore: Sparky House Publishing.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics, 26*(16), 2069-2070. doi:10.1093/bioinformatics/btq330.

Mercer, T. R., and Mattick, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Research, 23*(7), 1081–1088. doi:10.1101/gr.156612.113

Mikkelsen, T.S., Hillier, L.W., Zody, M.C., Eichler, E.E., Lander, E.S., Waterston, R.H. (2005). The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature, 437*(7055), 69–87.

Milatovich, A., Bolger, G., Michaeli, T., Francke, U. (1994). Chromosome localizations of genes for five cAMP-specific phosphodiesterases in man and mouse. Somatic *Cell and Molecular Genetics, 20*(2), 75–86. doi:10.1007/BF02290677.

Mirasierra, M., Fernández-Pérez, A., Díaz-Prieto, N., Vallejo, M. (2011) Alx3-deficient mice exhibit decreased insulin in beta cells, altered glucose homeostasis and increased apoptosis in pancreatic islets. *Diabetologia, 54*, 403-414.

Mirasierra, M. and Vallejo, M. (2006). The homeoprotein Alx3 expressed in pancreatic β-cells regulates insulin gene transcription by interacting with the basic helix-loop-helix protein E47. *Molecular Endocrinology, 20*, 2876-2889.

Modarresi, F., Faghihi, M. A., Lopez-Toledano, M. A., Fatemi, R. P., Magistri, M., Brothers, S. P., … Wahlestedt, C. (2012). Natural Antisense Inhibition Results in Transcriptional De-Repression and Gene Upregulation. *Nature Biotechnology*, *30*(5), 453–459. doi:10.1038/nbt.2158.

Monti, P., Heninger, A.K., Bonifacio, E. (2009). Differentiation, expansion, and homeostasis of autoreactive T cells in type 1 diabetes mellitus. *Current Diabetes Reports*, *9*(2), 113-118.

Morris, K.V. (2011). The emerging role of RNA in the regulation of gene transcription in human cells. *Seminars in Cell & Developmental Biology*, *22*(4), 351–358.

Mount, D.W. (2004). *Bioinformatics: Sequence and Genome Analysis.* 2nd Edition. Cold Spring Harbor Laboratory Press. pp263- 264.

Mungall, A.J., Palmer, S.A., Sims, S.K., Edwards, C. A., Ashurst, J. L., Wilming, L., Jones, M. C., Horton, R., Hunt, S. E., Scott, C. E.,… Beck, S (2003). The DNA sequence and analysis of human chromosome 6. *Nature*, *425*(6960), 805–811. doi:10.1038/nature02055.

Murtagh, F. and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274-295.

Neeper, M., Schmidt, A.M., Brett, J., Yan, S.D., Wang, F., Pan, Y.C., Elliston, K., Stern, D., Shaw, A. (1992). Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. *Journal of Biological Chemistry*, *267*(21), 14998–5004.

Nerup, J., Platz, P., Andersen, O.O., Christy, M., Lyngsoe, J., Poulsen, J.E., Ryder, L.P., Nielsen, L.S., Thomsen, M., Svejgaard, A. l. (1974). HL-A antigens and diabetes mellitus. *Lancet*, *2*(7885), 864-866.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., Cox, N. J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888. doi:10.1371/journal.pgen.1000888.

Nielsen, C., Hansen, D., Husby,S. Jacobsen, B.B., Lillevang, S.T. (2003). Association of a putative regulatory polymorphism in the PD-1 gene with susceptibility to type 1 diabetes. *Tissue Antigens*, *62*(6), 492-497.

Nieman, L.K. and Chanco Turner, M.L. (2006). Addison's disease. *Clinics in Dermatology*, *24*(4), 276–280. doi:10.1016/j.clindermatol.2006.04.006.

NIH (National Institutes of Health). (June 9th, 2015). Precision Medicine Initiative. [Online]. Available at: http://www.nih.gov/precisionmedicine/. [Accessed 13 June, 2015].

NIH (National Institutes of Health). (December 8th, 2014). Gene Therapy Used to Treat Hemophilia. [Online]. Available at: http://www.nih.gov/researchmatters/december2014/12082014hemophilia.htm. [Accessed 13 June, 2015].

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., Shendure, J., Bamshad, M.J. (2010). Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genetics*, *42*(1), 30–35.

Noble, J. A. and Erlich, H. A. (2012). Genetics of Type 1 Diabetes. *Cold Spring Harbor Perspectives in Medicine*, *2*(1), a007732. doi:10.1101/cshperspect.a007732.

Ohler, U. and Niemann, H. (2001). Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics, 17*, 56–60.

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, O., Guigo, R., Shiekhattar, R. (2010). Long non-coding RNAs with enhancer-like function in human. *Cell, 143*(1), 46–58. doi:10.1016/j.cell.2010.09.001.

O'Sullivan, B.P. and Freedman, S.D. (2009). Cystic fibrosis. *Lancet, 373*(9678), 1891-1904. doi: 10.1016/S0140-6736(09)60327-5.

Palazzo, A.F. and Gregory, T.R. (2014). The case for junk DNA. PLOS *Genetics, 10*(5), e1004351.

Palazzo, A.F. (2005). Synergy between sequence and size in large-scale genomics. *Nature Review Genetics, 6*(9), 699-708.

Palfi, S., Gurruchaga, J.M., Ralph, S.G., Lepetit, H., Lavisse, S., Buttery, P.C., Watts, C.,… Mitrophanous, K.A. (2014). Long-term safety and tolerability of ProSavin, a lentiviral vector-based gene therapy for Parkinson's disease: a dose escalation, open-label, phase 1/2 trial. *The Lancet, 383*(9923), 1138–1146.

Papaemmanuil, E., Hosking, F.J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A., Allan, J.M., Tomlinson, I.P., Taylor, M., Greaves, M., Houlston, R.S. (2009). Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Annual Review of Immunology, 41*(9), 1006–1010. doi:10.1038/ng.430.

Park, S.H. and Kim, S. (2012). Pattern discovery of multivariate phenotypes by association rule mining and its scheme for genome-wide association studies. *International Journal of Data Mining and Bioinformatics, 6*(5), 505-520.

Pasmant, E., Sabbagh, A., Vidaud, M., Bieche, I. (2011). ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *The Journal of the Federation of American Societies for Experimental Biology (FASEB J), 25*: 444–448. doi: 10.1096/fj.10-172452.

Payankaulam, S., Li, L. M., Arnosti, D. N. (2010). Transcriptional repression: conserved and evolved features. *Current Biology, 20*(17), R764–R771. doi:10.1016/j.cub.2010.06.037.

Peltonen, L. (2006). Lessons from studying monogenic disease for common disease. *Human Molecular Genetics, 15*(1), R67-R74.

Peltonen L. and McKusick V.A. (2001). Genomics and medicine. Dissecting human disease in the postgenomic era. *Science, 291*(5507), 1224–1229. doi: 10.1126/science.291.5507.1224.

Petretto, E., Liu, E.T., Aitman, T.J. (2007). A gene harvest revealing the archeology and complexity of human disease. *Nature Genetics, 39*, 1299-1301.

Permuth-Wey, J., Lawrenson, K., Shen, H. C., Velkova, A., Tyrer, J. P., Chen, Z., … Gayther, S. A. (2013). Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. *Nature Communications, 4*, 1627. doi:10.1038/ncomms2613.

Pique-Regi, R., Degner, J., Prichard, J., 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research, 21*(3), pp. 447-455.

Pierce, A.M., Schneider-Broussard, R., Philhower, J.L., Johnson, D.G. (1998). Differential activities of E2F family members: unique functions in regulating transcription. *Molecular Carcinogenesis*, *22*(3), 190–198.

Pierce, B. L. and Ahsan, H. (2011). Genome-wide "Pleiotropy Scan" Identifies HNF1A Region as a Novel Pancreatic Cancer Susceptibility Locus. *Cancer Research*, *71*(13), 4352–4358. doi:10.1158/0008-5472.CAN-11-0124.

PMC (Personalised Medicine Coalition). (2014). The case for personalised medicine. 4th edition. [Online]. Available at: http://www.personalizedmedicinecoalition.org/Resources/Personalized_Medicine_101. [Accessed 13 June, 2015].

PMC (Personalised Medicine Coalition). (2014). The case for personalised medicine. 4th edition. [Online]. Available at: http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/pmc_case_for_personalized_medicine.pdf. [Accessed 13 June, 2015].

Pociot, F., Akolkar, B., Concannon, P., Erlich, H.A., Julier, C., Morahan, G., Nierras, C.R., Todd, J.A., Rich, S. S., Nerup, J. (2010). Genetics of Type 1 Diabetes: What's Next? *Perspectives in Diabetes*, *59*, 1561 – 1571.

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., … Freedman, M. L. (2009). The 8q24 cancer risk variant rs6983267 demonstrates long-range interaction with MYC in colorectal cancer. *Nature Genetics*, *41*(8), 882–884. doi:10.1038/ng.403.

Ponder, K.P. (2006). Gene therapy for haemophilia. *Current opinions in hematology*, *13*, 301-307.

Ponder, K.P. (2011). Hemophilia Gene Therapy: A Holy Grail Found. *Molecular Therapy*, *19*(3), 427-428.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Proutski, V. and Holmes, E.C.(1997). SWAN: sliding window analysis of nucleotide sequence variablility. *Bioinformatics Applications Note*, *14*(5), pp 467-468.

Qi, D., Blake, J.A., Kadin, J.A., Richardson, J.E., Ringwald, M., Eppig, J.T., Bult, C.J. (2005). Data integration in the mouse genome informatics (MGI) database. *Computational Systems Bioinformatics Conference*, *2005* (IEEE), 37–38. doi:10.1109/CSBW.2005.48.

Qian, J., Kaytor, E. N., Towle, H. C., Olson, L. K. (1999) Upstream stimulatory factor regulates Pdx-1 gene expression in differentiated pancreatic β-cells. *Biochemical Journal*, *341*, 315-322.

Qiao, Q., Österholm, A.-M., He, B., Pitkäniemi, J., Cordell, H. J., Sarti, C., … Tuomilehto, J. (2007). A genome-wide scan for type 1 diabetes susceptibility genes in nuclear families with multiple affected siblings in Finland. *BMC Genetics*, *8*(84), 1471-2156. doi:10.1186/1471-2156-8-84.

Qin, L.Y., Zhao, L.G., Chen, X., Li, P., Yang, Z., Mo, W.N. (2014). The CCND1 G870A gene polymorphism and brain tumor risk: a meta-analysis. *Asian Pacific Journal of Cancer Prevention*, *15*(8), 3607-3612.

Rabbani, B., Tekin, M., Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*, 5–15.

Raha, O., Sarkar, B.N., Bhaskar, L.V.K.S., Veerraju, P., Chowdhury, S.; Mukhopadhyay, S., Biswas, T. Kr., Rao, V.R. (2011). Insulin (INS) promoter VNTR polymorphisms: interactions and association with Type 1 Diabetes mellitus in Bengali speaking patients of Eastern India. *Diabetologia Croatica, 40*(4), 99.

Rearick, D., Prakash, A., Mcsweeny, A., Shepard, S. S., Fedorova, L., Fedorov, A. (2011). Critical association of ncRNA with introns. *Nucleic Acids Research, 39*, 2357–2366.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y. (2007). Functional Demarcation of active and Silent chromatin domains in human HOX loci by noncoding RNAs. *Cell, 129*, 1311–1323. doi: 10.1016/j.cell.2007.05.022.

Rogan, P. K. (2010). Deeper understanding of unclassified intronic variants and ESEs. *Human Mutation, 31*(4), V. doi:10.1002/humu.21247.

Romano, M., Buratti, E., Baralle, D. (2013). Role of Pseudoexons and Pseudointrons in Human Cancer. *International Journal of Cell Biology, 2013*(810572), 1-16. doi:10.1155/2013/810572.

Rubin, B.Y. and Anderson, S.L. (2008). The Molecular Basis of Familial Dysautonomia: Overview, New Discoveries and Implications for Directed Therapies. *NeuroMolecular Medicine, 10*(3), 148-56.

Sacchetti, A., El Sewedy, T. Nasr, A F, Alberti, S. (2001). Efficient GFP mutations profoundly affect mRNA transcription and translation rates. *FEBS Letters, 492*(1–2), 151–155.

Salta, E. and De Strooper, B. (2012). Non-coding RNAs with essential roles in neurodegenerative disorders. *Lancet Neurology, 11*(2), 189-200.

Sammeth, M., Foissac, S., Guigó, R. (2008). A General Definition and Nomenclature for Alternative Splicing Events. *PLoS Computational Biology, 4*(8), e1000147. doi:10.1371/journal.pcbi.1000147.

Sanderson, F., Kleijmeer, M.J., Kelly, A., Verwoerd, D., Tulp, A., Neefjes, J.J., Geuze, H.J., Trowsdale, J. (1994). Accumulation of HLA-DM, a regulator of antigen presentation, in MHC class II compartments. *Science, 266*(5190), 1566-1569.

Sanghera, D.K. and Blackett, P.R. (2012). Type 2 Diabetes Genetics: Beyond GWAS. *Journal of Diabetes & Metabolism, 3*(198), 6948. doi:10.4172/2155-6156.1000198.

Sanna, C., Li, W., Zhang, L. (2008). Overlapping genes in the human and mouse genomes. *BMC Genomics, 9*(169). doi:10.1186/1471-2164-9-169.

Santin, I. and Eizirik, D. L. (2013). Candidate genes for type 1 diabetes modulate pancreatic islet inflammation and β-cell apoptosis. *Diabetes, Obesity and Metabolism, 15*(s3), 71–81.

Sardet, C., Vidal, M., Cobrinik, D., Geng, Y., Onufryk, C., Chen, A., Weinberg, R.A. (1995). E2F-4 and E2F-5, two members of the E2F family, are expressed in the early phases of the cell cycle. *Proceedings of the National Academy of Sciences USA, 92*(6), 2403–2407. doi:10.1073/pnas.92.6.2403.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B., Friend ,S.H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature, 422*, 297–302.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome *Research*, *22*(9), 1748–1759. doi:10.1101/gr.136127.111.

Schram, C.A. (2012). Atypical cystic fibrosis: Identification in the primary care setting. *Canada Family Physician*, *58*(12), 1341–1345.

Schwartz, D.C. and Parker, R. (1999). Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in Saccharomyces cerevisiae. *Molecular and Cell Biology*, *19*(8), 5247-56.

Shaw, D. (2004). Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Current Protocols in Bioinformatics*, Chapter 1: Unit 1.7. doi:10.1002/0471250953.bi0107s05.

Sheridan, C. (2011). Gene therapy finds its niche. *Nature Biotechnology*, *29*, 121–128.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.(2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*, 308–311.

Shi, X., Sun, M., Liu, H., Yao, Y., Song, Y. (2013). Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer Letters*, *339*(2), 159-166.

Shieh, B.H., Sparkes, R.S., Gaynor, R.B., Lusis, A.J. (1993). Localization of the gene-encoding upstream stimulatory factor (USF) to human chromosome 1q22-q23. *Genomics*, *16*(1), 266–268. doi:10.1006/geno.1993.1174.

Shiraishi Y, Fujimoto A, Furuta M, Tanaka H, Chiba K-I, Boroevich, K.A., Abe, T., Kawakami, Y., Ueno, M., Gotoh, K., Ariizumi, S.i., Shibuya, T., Nakano, K. Sasaki, A., Maejima, K., Kitada, R., Hayami, S., Shigekawa, Y., Marubashi, S., Yamada, T., Kubo, M., Ishikawa, O., Aikata, H., Arihiro, K., Ohdan, H., Yamamoto, M., Yamaue, H., Chayama, K., Tsunoda, T., Miyano, S., Nakagawa, H. (2014) Integrated Analysis of Whole Genome and Transcriptome Sequencing Reveals Diverse Transcriptomic Aberrations Driven by Somatic Genomic Changes in Liver Cancers. *PLoS ONE 9*(12), e114263. doi:10.1371/journal.pone.0114263.

Silverman, E. K. and Loscalzo, J. (2013). Developing New Drug Treatments in the Era of Network Medicine. *Clinical Pharmacology and Therapeutics*, *93*(1), 26–28. doi:10.1038/clpt.2012.207.

Singal, D. and Blajchman, M. (1973). Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes*, *22*(6), 429-32.

Singh, S.B., Davis, A.S., Taylor, G.A., Deretic, V. (2006). Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science*, *313* (5792), 1438–144.

Singh, V. P., Bali, A., Singh, N., Jaggi, A. S. (2014). Advanced Glycation End Products and Diabetic Complications. The Korean Journal of Physiology & Pharmacology : Official Journal of the Korean Physiological *Society and the Korean Society of Pharmacology*, *18*(1), 1–14. doi:10.4196/kjpp.2014.18.1.1.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan,I., McKeigue, P., Wilson, J.F., Campbell, H. (2011). Abundant Pleiotropy in Human Complex Diseases and Traits. *American Journal of Human Genetics*, *89*(5), 607–618. doi:10.1016/j.ajhg.2011.10.004.

Smyth, D. J., Plagnol, V., Walker, N. M., Cooper, J.D., Downes, K., Yang, J.H., Howson, J. M., Stevens, H., McManus, R., Wijmenga, C., Heap, G.A., Dubois, P.C., Clayton, D.G., Hunt, K.A., van Heel, D.A., Todd, J.A. (2008). Shared and Distinct Genetic Variants in Type 1 Diabetes and Celiac Disease. *The New England Journal Medicine*, *359*(26), 2767-2777. doi:10.1056/NEJMoa0807917.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., … Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, *21*(10), 1757–1767. doi:10.1101/gr.121541.111.

Stalder, J., Larsen, A., Engel, J.D., Dolan, M., Groudine, M., Weintraub, H. (1980). Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. *Cell*, *20*(2), 451-460.

Steck, A. and Rewers, M. (2011). Genetics of type 1 diabetes. Clinical Chemistry, 57(2), 176-185.

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., … Stamatoyannopoulos, J. A. (2013). Exonic transcription factor binding directs codon choice and impacts protein evolution. *Science*, *342*(6164), 1367–1372. doi:10.1126/science.1243490.

Stormo, G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, *16*(1), 16–23.

Stewart, A. J., Hannenhalli, S., Plotkin, J. B. (2012). Why Transcription Factor Binding Sites Are Ten Nucleotides Long. *Genetics*, *192*(3), 973–985. doi:10.1534/genetics.112.143370.

Stormo, G.D. (1990) Consensus patterns in DNA. Methods in Enzymology, 183, 211–221.

Stranger, B.E. and Dermitzakis, E.T. (2005). The genetics of regulatory variation in the human genome. *Human Genomics*, *2*, 126–131.

Stranger, B.E., Stahl, E.A., Raj, T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics*, *187*(2), 367–383.

Struckmann, S., Esch, D., Schöler, H., Fuellen, G. (2011). Visualization and Exploration of Conserved Regulatory Modules Using ReXSpecies 2. *BMC Evolutionary Biology*, *11*, 267. doi:10.1186/1471-2148-11-267

Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A.S., Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, *10*(6), 591-597.

Swafford, A.D., Howson, J.M., Davison, L.J., Wallace, C., Smyth, D.J., Schuilenburg, H., Maisuria-Armer, M., Mistry, T., Lenardo, M.J., Todd, J.A. (2011). An allele of IKZF1 (Ikaros) conferring susceptibility to childhood acute lymphoblastic leukemia protects against type 1 diabetes. *Diabetes*, *60*(3), 1041–1044. doi:10.2337/db10-0446.

Tait, B.D. and Boyle, A.J. (1986). DR4 and susceptibility to type I diabetes mellitus: Discrimination of high risk and low risk DR4 haplotypes on the basis of TA10 typing. *Tissue Antigens*, *28*, 65–71.

Tan, M.S., Yu, J.T., Jiang, T., Zhu, X.C., Wang, H.F., Zhang, W., Wang, Y.L., Jiang, W., Tan, L. (2013). NLRP3 polymorphisms are associated with late-onset Alzheimer's disease in Han Chinese. *Journal of Neuroimmunology*, *265*(1-2), 91–95.doi: 10.1016/j.jneuroim.2013.10.002.

Tanzi, R.E. and Bertram, L. (2005). 20 years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell, 120*(4), 545-555.

Ten, S., New, M., Maclaren, N. (2001). Clinical review 130: Addison's disease 2001. *The Journal of Clinical Endocrinology and Metabolism 86*(7), 2909–2922. doi:10.1210/jc.86.7.2909.

The ENCODE Project Consortium. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature, 489*(7414), 57–74. doi:10.1038/nature11247.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics, 17*(12), 1113-1122.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., Moreau, Y. (2002). A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology, 9*(2), 447-464.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature, 489*(7414), 75–82. doi:10.1038/nature11232.

Timmerman, L. (4 February 2013). What's in a Name? A Lot, When It Comes to 'Precision Medicine. Xconomy. [Online]. Available at: http://www.xconomy.com/national/2013/02/04/whats-in-a-name-a-lot-when-it-comes-to-precision-medicine/. [Accessed 13 June, 2015].

Todd, J.A. (2007). Etiology of type 1 diabetes. *Immunity, 32*(4):457-67. doi: 10.1016/j.immuni.2010.04.001.

Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F.,… Clayton, D.G. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics, 39*(7), 857–864. doi:10.1038/ng2068.

Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., Pontoglio, M. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *Journal of Molecular Biology, 266*(2), 231–245.

Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science, 329*, 689–693. doi: 10.1126/science.1192002.

Tsompana, M. and Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin, 7*, 33. doi:10.1186/1756-8935-7-33.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., Mecklin, J.P., Järvinen, H., Ristimäki, A., Di-Bernardo, M., East, P., Carvajal-Carmona, L., Houlston, R.S., Tomlinson, I., Palin, K., Ukkonen, E., Karhu, A., Taipale, J., Aaltonen, L.A. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signalling. *Nature Genetics, 41*(8), 885–890.

Vanet, A., Marsan, L., Labigne, A., Sagot, M.F. (2000). Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori σ80 family of promoter signals. *Journal of Molecular Biology, 297*, 335–353.

van Helden, J., Andre, B., Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology, 281*, 827–842.

van Helden, J., del Olmo, M., Perez-Ortin, J.E. (2000a). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Research, 28*, 1000–1010.

Vestweber, D. (1997). *The Selectins: Initiators of Leukocyte Endothelial Adhesion.* Amsterdam: Harwood Academic Publishers.

Vreeswijk, M.P., Kraan, J.N., van der Klift, H.M., Vink, G.R., Cornelisse, C.J., Wijnen, J.T., Bakker, E., van Asperen, C.J., Devilee, P. (2009). Intronic variants in BRCA1 and BRCA2 that affect RNA splicing can be reliably selected by splice-site prediction programs. *Human Mutations, 30*(1),107-114. doi: 10.1002/humu.20811.

Wahlestedt, C. (2013). Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nature Reviews Drug Discovery, 12*(6), 433-446.

Wallace, C., Smyth, D. J., Maisuria-Armer, M., Walker, N. M., Todd, J. A., Clayton, D. G. (2010). The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nature Genetics, 42*(1), 68–71. doi:10.1038/ng.493.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., Chang, H.Y. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature, 505*, 706–709

Wang, G.S. and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics, 8*, 749-761

Wang, Z., Gerstein, M., Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*(1), 57–63. doi:10.1038/nrg2484

Ward, A. J. and Cooper, T. A. (2010). The Pathobiology of Splicing. *The Journal of Pathology, 220*(2), 152–163. doi:10.1002/path.2649

Ward, L. D. and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research, 40*(Database issue), D930–D934. doi:10.1093/nar/gkr917

Ward, L. D. and Kellis, M. (2012). Interpreting non-coding variation in complex disease genetics. *Nature Biotechnology, 30*(11), 1095 – 1106.

Wellcome Trust Case Control Consortium, Burton, P., Clayton, D., Cardon, L., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), pp. 661-678.

Wellcome Trust Case Control Consortium, 2007. *Importance of gene regulation for common human disease. "Science Daily".* [Online]. Available at: http://www.sciencedaily.com/releases/2007/09/070916143515.htm. [Accessed 12 June, 2012].

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001-D1006.

Wenzlau, J. M., Moua, O., Sarkar, S. A., Yu, L., Rewers, M., Eisenbarth, G. S., Davidson, H. W., Hutton, J. C. (2008). SlC30A8 is a Major Target of Humoral Autoimmunity in Type 1 Diabetes and a Predictive Marker in Prediabetes. *Immunology of Diabetes V: From Bench to Bedside*, *1150*, 256–259.

Weinberg, M. S. and Morris, K. V. (2013). Long Non-Coding RNA Targeting and Transcriptional De-Repression. *Nucleic Acid Therapeutics*, *23*(1), 9–14. doi:10.1089/nat.2012.0412

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., … Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, *13(*9), R50. doi:10.1186/gb-2012-13-9-r50

Workman, C.T. and, Stormo, G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pacific Symposium on Biocomputation*, 467–478.

World Health Organisation. (2015). Diabetes. Fact sheet 312. [Online]. Available at: http://www.who.int/mediacentre/factsheets/fs312/en/ [Accessed 09 March, 2015]

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., Romano, L.A. (2003). The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, *20*, 1377–1419.

Wright, J. B., Brown, S. J., Cole, M. D. (2010). Upregulation of c-MYC in cis through a Large Chromatin Loop Linked to a Cancer Risk-Associated Single-Nucleotide Polymorphism in Colorectal Cancer Cells. *Molecular and Cellular Biology*, *30*(6), 1411–1420. doi:10.1128/MCB.01384-09.

Wu, Z, Chen, H., Sun, F., Xu, J., Zheng, W., Li, P., Chen, S., Shen, M., Zhang, W., Li, M., You, X., Wu, Q., Zhang, F., Li, Y. (2014). PTPN2 rs1893217 single-nucleotide polymorphism is associated with risk of Behçet's disease in a Chinese Han population. *Clinical and Experimental Rheumatology*. *32*(4), S20-26.

Xu, Z. and Taylor, J. A. (2009). SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Research*, *37*(Web Server issue), W600–W605. doi:10.1093/nar/gkp290.

Yang, J.H.M., Downes, K., Howson, J.M.M., Nutland, S., Stevens, H.E., Walker, N.M., Todd, J.A. (2011). Evidence of association with type 1 diabetes in the SLC11A1 gene region. *BMC Medical Genetics*, *12*, *59*. doi:10.1186/1471-2350-12-59.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., Kruglyak, L. (2003). Transacting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, *35*, 57–64.

Zama, M. (1999). Correlation between mRNA structure of the coding region and translational pauses. *Nucleic Acids Symposium Series*, *42*(1), 81–82.

Zambelli, F., Prazzoli, G.M., Pesole, G., Pavesi, G. (2012). Cscan: Finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. *Nucleic Acids Research 40*, W510–W515.

Zhang, C.T., Wang, J., Zhang, R. (2001). A novel method to calculate the G+C content of genomic DNA sequences. *Journal of Biomolecular Structure & Dynamics, 19*(2), 333-341.

Zhang, H., Zhang, X., Ji, S., Hao, C., Mu, Y., Sun, J., Hao, J. (2014) Sohlh2 inhibits ovarian cancer cell proliferation by upregulation of p21 and downregulation of cyclin D1. *Carcinogenesis, 35*(8), 1863-1871. doi: 10.1093/carcin/bgu113.

Zhang, Q, Feitosa, M., Borecki, I.B. (2014). Estimating and testing pleiotropy of single genetic variant for two quantitative traits. *Genetic Epidemiology*, *38*(6), 523-530. doi: 10.1002/gepi.21837. Epub Jul 12.

Zhang, X., Cowper-Sal lari, R., Bailey, S.D., Moore, J.H., Lupien, M. (2012). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Research, 22*(8), 1437–1446.

Zhao, X., Xuan, Z., Zhang, M.Q. (2007). Boosting with stumps for predicting transcription start sites. *Genome Biology, 8*(2), R17.

Zheng, W., Gianoulis, T.A., Karczewski, K.J., Zhao, H., Snyder, M. (2010). Regulatory variation within and between species. *Annual Review of Genomics and Human Genetics, 12*, 327-346.

Zhou, Y., Zhang, X., Klibanski, A. (2012). MEG3 noncoding RNA: a tumor suppressor. *Journal of Molecular Endocrinology, 48*(3), R45–R53. doi:10.1530/JME-12-0008

Zhu, J. and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics. 15* (7), 607-611.

APPENDICES

Appendix A: Additional data for Associated and Non-associated SNPs in the T1D susceptibility regions

A.1 Occurrence of Associated and Non-associated SNPs in overlapping and single gene transcripts.

Table 1. In overlap vs not in overlap

Counts	Associated	Non-Associated	Total
In Overlap	62	199447	199509
Not in overlap	24	62769	62793
	86	262216	262302

$X^2 = 0.74$ p= 0.3884

Table 2. Standardised residuals: In overlap vs not in overlap

(O-E)/√E	Associated	Non- Associated
In Overlap	-0.422	0.008
Not in overlap	0.752	-0.014

Table 3. Gene vs Transcript vs Gene Flanking.

Counts	Associated	Non-Associated	Total
Gene & Transcript Overlap	12	23809	23821
Transcript Overlap	41	103762	103803
Gene Flanking	9	71876	71885
	62	199447	199509

$X^2 = 13.19$ p= 0.0013

Table 4. Standardised residuals: Gene vs Transcript vs Gene Flanking.

(O-E)/√E	Associated	Non- Associated
Gene & Transcript Overlap	1.690	-0.030
Transcript Overlap	1.539	-0.027
Gene Flanking	-2.822	0.050

Table 5. Gene vs Transcript vs Gene Flanking vs 'No' overlap.

Counts	Associated	Non-Associated	Total
Gene & Transcript Overlap	12	23809	23821
Transcript Overlap	41	103762	103803
Single Genic Position	24	62769	62793
Gene Flanking	9	71876	71885
	86	262216	262302

X^2 = 13.25; p= 0.004

Table 6. Standardised residuals: Gene vs Transcript vs Gene Flanking vs 'No' overlap.

Standardized Residuals	Associated	Non-Associated
Gene & Transcript Overlap	**1.50**	**-0.03**
Transcript Overlap	**1.19**	**-0.02**
Single Genic Position	**0.75**	**-0.01**
Gene Flanking	**-3.00**	**0.05**

Table 7. The Genic Profiles of Associated and Non-Associated SNPs

No of Profile Components	Genic Profiles	SNP Counts	
		Non_Assoc	Assoc
1	coding_sequence	4	0
1	splice_donor/acceptor	38	0
1	nc_transcript	7453	6
1	intron	59483	3
1	intergenic	69432	17
1	exon	1524	3
1	5KB_upstream	18764	5
1	5KB_downstream	21174	3
1	5_prime_UTR	144	1
1	3_prime_UTR	697	0
2	nc_transcript coding_sequence	3	0
2	nc_transcript splice_donor/acceptor	6	0
2	intron splice_donor/acceptor	15	0
2	intron NMD_transcript	163	0
2	intron nc_transcript	5424	17
2	intergenic nc_transcript	1	0
2	exon splice_donor/acceptor	4	0
2	exon NMD_transcript	8	0
2	exon nc_transcript	261	1
2	exon intron	392	0
2	5KB_upstream splice_donor/acceptor	13	0
2	5KB_upstream nc_transcript	2065	0
2	5KB_upstream intron	14215	1
2	5KB_upstream exon	304	0
2	5KB_downstream coding_sequence	3	0
2	5KB_downstream splice_donor/acceptor	19	0
2	5KB_downstream NMD_transcript	78	0
2	5KB_downstream nc_transcript	3353	1
2	5KB_downstream intron	11718	6
2	5KB_downstream exon	503	2
2	5KB_downstream 5KB_upstream	7982	0
2	5_prime_UTR nc_transcript	35	0
2	5_prime_UTR intron	129	0
2	5_prime_UTR exon	96	0
2	5_prime_UTR 5KB_upstream	258	0
2	5_prime_UTR 5KB_downstream	47	0
2	3_prime_UTR nc_transcript	68	0
2	3_prime_UTR intron	166	0
2	3_prime_UTR exon	9	0
2	3_prime_UTR 5KB_upstream	82	0
2	3_prime_UTR 5KB_downstream	1393	0
3	nc_transcript splice_donor/acceptor coding_sequence	1	0
3	intron NMD_transcript splice_donor/acceptor	1	0
3	intron nc_transcript splice_donor/acceptor	2	0
3	intron nc_transcript NMD_transcript	184	3
3	intergenic intron nc_transcript	8	0
3	exon nc_transcript NMD_transcript	10	0
3	exon intron NMD_transcript	8	0
3	exon intron nc_transcript	104	0
3	5KB_upstream nc_transcript coding_sequence	2	0
3	5KB_upstream nc_transcript splice_donor/acceptor	12	0
3	5KB_upstream nc_transcript NMD_transcript	7	0
3	5KB_upstream intron splice_donor/acceptor	5	0
3	5KB_upstream intron NMD_transcript	64	0
3	5KB_upstream intron nc_transcript	3044	2
3	5KB_upstream exon NMD_transcript	5	0

Table 7. The Genic Profiles of Associated and Non-Associated SNPs contd.

No of Profile Components	Genic Profiles	SNP Counts	
		Non_Assoc	Assoc
3	5KB_upstream exon nc_transcript	269	0
3	5KB_upstream exon intron	149	0
3	5KB_downstream NMD_transcript splice_donor/acceptor	3	0
3	5KB_downstream nc_transcript coding_sequence	10	0
3	5KB_downstream nc_transcript splice_donor/acceptor	19	0
3	5KB_downstream intron splice_donor/acceptor	11	0
3	5KB_downstream intron NMD_transcript	444	0
3	5KB_downstream intron nc_transcript	2202	2
3	5KB_downstream exon splice_donor/acceptor	3	0
3	5KB_downstream exon NMD_transcript	31	0
3	5KB_downstream exon nc_transcript	357	0
3	5KB_downstream exon intron	164	0
3	5KB_downstream 5KB_upstream coding_sequence	5	0
3	5KB_downstream 5KB_upstream splice_donor/acceptor	21	0
3	5KB_downstream 5KB_upstream NMD_transcript	1	0
3	5KB_downstream 5KB_upstream nc_transcript	1300	1
3	5KB_downstream 5KB_upstream intron	6724	0
3	5KB_downstream 5KB_upstream exon	448	0
3	5_prime_UTR intron splice_donor/acceptor	2	0
3	5_prime_UTR intron NMD_transcript	1	0
3	5_prime_UTR intron nc_transcript	29	0
3	5_prime_UTR exon NMD_transcript	1	0
3	5_prime_UTR exon nc_transcript	245	0
3	5_prime_UTR exon intron	60	0
3	5_prime_UTR 5KB_upstream splice_donor/acceptor	1	0
3	5_prime_UTR 5KB_upstream nc_transcript	128	0
3	5_prime_UTR 5KB_upstream intron	290	0
3	5_prime_UTR 5KB_upstream exon	41	0
3	5_prime_UTR 5KB_downstream nc_transcript	6	0
3	5_prime_UTR 5KB_downstream intron	47	0
3	5_prime_UTR 5KB_downstream exon	2	0
3	5_prime_UTR 5KB_downstream 5KB_upstream	243	0
3	3_prime_UTR nc_transcript NMD_transcript	3	0
3	3_prime_UTR intron splice_donor/acceptor	1	0
3	3_prime_UTR intron nc_transcript	6	0
3	3_prime_UTR exon NMD_transcript	1	0
3	3_prime_UTR exon nc_transcript	8	0
3	3_prime_UTR exon intron	24	0
3	3_prime_UTR 5KB_upstream nc_transcript	32	0
3	3_prime_UTR 5KB_upstream intron	28	0
3	3_prime_UTR 5KB_upstream exon	10	0
3	3_prime_UTR 5KB_downstream splice_donor/acceptor	1	0
3	3_prime_UTR 5KB_downstream NMD_transcript	16	0
3	3_prime_UTR 5KB_downstream nc_transcript	675	0
3	3_prime_UTR 5KB_downstream intron	515	0
3	3_prime_UTR 5KB_downstream exon	85	0
3	3_prime_UTR 5KB_downstream 5KB_upstream	524	0
3	3_prime_UTR 5_prime_UTR exon	2	0
3	3_prime_UTR 5_prime_UTR 5KB_upstream	4	0
4	intronnc_transcriptNMD_transcriptsplice_donor/acceptor	3	0
4	intergenicintronnc_transcriptNMD_transcript	1	0
4	exonintronNMD_transcriptsplice_donor/acceptor	2	0
4	exonintronnc_transcriptsplice_donor/acceptor	2	0
4	exonintronnc_transcriptNMD_transcript	33	1
4	5KB_upstreamnc_transcriptNMD_transcriptsplice_donor/acceptor	6	0
4	5KB_upstreamintronNMD_transcriptsplice_donor/acceptor	2	0
4	5KB_upstreamintronnc_transcriptcoding_sequence	14	0
4	5KB_upstreamintronnc_transcriptsplice_donor/acceptor	13	0

Table 7. The Genic Profiles of Associated and Non-Associated SNPs contd.

No of Profile Components	Genic Profiles	SNP Counts	
		Non_Assoc	Assoc
4	5KB_upstreamintronnc_transcriptNMD_transcript	618	3
4	5KB_upstreamexonnc_transcriptNMD_transcript	469	0
4	5KB_upstreamexonintronNMD_transcript	10	0
4	5KB_upstreamexonintronnc_transcript	139	0
4	5KB_downstreamnc_transcriptsplice_donor/acceptorcoding_sequence	1	0
4	5KB_downstreamnc_transcriptNMD_transcriptcoding_sequence	1	0
4	5KB_downstreamnc_transcriptNMD_transcriptsplice_donor/acceptor	6	0
4	5KB_downstreamintronnc_transcriptcoding_sequence	3	0
4	5KB_downstreamintronnc_transcriptsplice_donor/acceptor	5	0
4	5KB_downstreamintronnc_transcriptNMD_transcript	518	1
4	5KB_downstreamintergenicintronnc_transcript	2	0
4	5KB_downstreamexonnc_transcriptNMD_transcript	174	0
4	5KB_downstreamexonintronNMD_transcript	46	0
4	5KB_downstreamexonintronnc_transcript	96	0
4	5KB_downstream5KB_upstreamNMD_transcriptsplice_donor/acceptor	3	0
4	5KB_downstream5KB_upstreamnc_transcriptcoding_sequence	6	0
4	5KB_downstream5KB_upstreamnc_transcriptsplice_donor/acceptor	35	0
4	5KB_downstream5KB_upstreamintronsplice_donor/acceptor	5	0
4	5KB_downstream5KB_upstreamintronNMD_transcript	435	0
4	5KB_downstream5KB_upstreamintronnc_transcript	4207	1
4	5KB_downstream5KB_upstreamexoncoding_sequence	1	0
4	5KB_downstream5KB_upstreamexonNMD_transcript	31	0
4	5KB_downstream5KB_upstreamexonnc_transcript	558	0
4	5KB_downstream5KB_upstreamexonintron	128	1
4	5_prime_UTRintronnc_transcriptsplice_donor/acceptor	1	0
4	5_prime_UTRexonnc_transcriptNMD_transcript	1	0
4	5_prime_UTRexonintronNMD_transcript	4	0
4	5_prime_UTRexonintronnc_transcript	28	0
4	5_prime_UTR5KB_upstreamnc_transcriptsplice_donor/acceptor	1	0
4	5_prime_UTR5KB_upstreamnc_transcriptNMD_transcript	18	0
4	5_prime_UTR5KB_upstreamintronnc_transcript	157	0
4	5_prime_UTR5KB_upstreamexonnc_transcript	146	0
4	5_prime_UTR5KB_upstreamexonintron	5	0
4	5_prime_UTR5KB_downstreamintronnc_transcript	37	0
4	5_prime_UTR5KB_downstreamexonNMD_transcript	2	0
4	5_prime_UTR5KB_downstreamexonnc_transcript	10	0
4	5_prime_UTR5KB_downstreamexonintron	1	0
4	5_prime_UTR5KB_downstream5KB_upstreamsplice_donor/acceptor	2	0
4	5_prime_UTR5KB_downstream5KB_upstreamNMD_transcript	2	0
4	5_prime_UTR5KB_downstream5KB_upstreamnc_transcript	60	0
4	5_prime_UTR5KB_downstream5KB_upstreamintron	194	0
4	5_prime_UTR5KB_downstream5KB_upstreamexon	19	0
4	3_prime_UTRexonnc_transcriptNMD_transcript	1	0
4	3_prime_UTRexonintronnc_transcript	30	0
4	3_prime_UTR5KB_upstreamintronnc_transcript	34	0
4	3_prime_UTR5KB_upstreamexonnc_transcript	96	0
4	3_prime_UTR5KB_downstreamnc_transcriptsplice_donor/acceptor	3	0
4	3_prime_UTR5KB_downstreamnc_transcriptNMD_transcript	157	0
4	3_prime_UTR5KB_downstreamintronnc_transcript	200	0
4	3_prime_UTR5KB_downstreamexonNMD_transcript	26	0
4	3_prime_UTR5KB_downstreamexonnc_transcript	86	0
4	3_prime_UTR5KB_downstreamexonintron	29	0
4	3_prime_UTR5KB_downstream5KB_upstreamNMD_transcript	10	0
4	3_prime_UTR5KB_downstream5KB_upstreamnc_transcript	486	1
4	3_prime_UTR5KB_downstream5KB_upstreamintron	191	0
4	3_prime_UTR5KB_downstream5KB_upstreamexon	34	0
4	3_prime_UTR5_prime_UTRexonnc_transcript	1	0
4	3_prime_UTR5_prime_UTR5KB_upstreamintron	2	0

Table 7. The Genic Profiles of Associated and Non-Associated SNPs contd.

No of Profile	Genic Profiles	SNP Counts	
Components		Non_Assoc	Assoc
4	3_prime_UTR5_prime_UTR5KB_downstreamexon	1	0
4	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstream	3	0
5	exonintronnc_transcriptNMD_transcriptsplice_donor/acceptor	1	0
5	5KB_upstreamnc_transcriptNMD_transcriptsplice_donor/acceptorcoding_sequence	1	0
5	5KB_upstreamintronnc_transcriptsplice_donor/acceptorcoding_sequence	5	0
5	5KB_upstreamintronnc_transcriptNMD_transcriptsplice_donor/acceptor	4	0
5	5KB_upstreamexonnc_transcriptNMD_transcriptcoding_sequence	1	0
5	5KB_upstreamexonnc_transcriptNMD_transcriptsplice_donor/acceptor	3	0
5	5KB_upstreamexonintronnc_transcriptcoding_sequence	1	0
5	5KB_upstreamexonintronnc_transcriptsplice_donor/acceptor	1	0
5	5KB_upstreamexonintronnc_transcriptNMD_transcript	87	0
5	5KB_downstreamintronnc_transcriptNMD_transcriptsplice_donor/acceptor	9	0
5	5KB_downstreamexonintronnc_transcriptsplice_donor/acceptor	1	0
5	5KB_downstreamexonintronnc_transcriptNMD_transcript	34	0
5	5KB_downstream5KB_upstreamnc_transcriptsplice_donor/acceptorcoding_sequence	2	0
5	5KB_downstream5KB_upstreamnc_transcriptNMD_transcriptcoding_sequence	23	0
5	5KB_downstream5KB_upstreamnc_transcriptNMD_transcriptsplice_donor/acceptor	23	0
5	5KB_downstream5KB_upstreamintronNMD_transcriptsplice_donor/acceptor	3	0
5	5KB_downstream5KB_upstreamintronnc_transcriptcoding_sequence	4	0
5	5KB_downstream5KB_upstreamintronnc_transcriptsplice_donor/acceptor	14	0
5	5KB_downstream5KB_upstreamintronnc_transcriptNMD_transcript	3188	2
5	5KB_downstream5KB_upstreamexonnc_transcriptcoding_sequence	1	0
5	5KB_downstream5KB_upstreamexonnc_transcriptNMD_transcript	532	1
5	5KB_downstream5KB_upstreamexonintronsplice_donor/acceptor	2	0
5	5KB_downstream5KB_upstreamexonintronNMD_transcript	22	0
5	5KB_downstream5KB_upstreamexonintronnc_transcript	284	0
5	5_prime_UTRexonintronnc_transcriptNMD_transcript	1	0
5	5_prime_UTR5KB_upstreamnc_transcriptNMD_transcriptsplice_donor/acceptor	2	0
5	5_prime_UTR5KB_upstreamintronnc_transcriptsplice_donor/acceptor	2	0
5	5_prime_UTR5KB_upstreamintronnc_transcriptNMD_transcript	57	0
5	5_prime_UTR5KB_upstreamexonnc_transcriptcoding_sequence	2	0
5	5_prime_UTR5KB_upstreamexonnc_transcriptsplice_donor/acceptor	7	0
5	5_prime_UTR5KB_upstreamexonnc_transcriptNMD_transcript	152	0
5	5_prime_UTR5KB_upstreamexonintronNMD_transcript	3	0
5	5_prime_UTR5KB_upstreamexonintronnc_transcript	65	0
5	5_prime_UTR5KB_downstreamexonnc_transcriptNMD_transcript	1	0
5	5_prime_UTR5KB_downstreamexonintronnc_transcript	14	0
5	5_prime_UTR5KB_downstream5KB_upstreamnc_transcriptcoding_sequence	1	0
5	5_prime_UTR5KB_downstream5KB_upstreamnc_transcriptsplice_donor/acceptor	1	0
5	5_prime_UTR5KB_downstream5KB_upstreamnc_transcriptNMD_transcript	7	0
5	5_prime_UTR5KB_downstream5KB_upstreamintronsplice_donor/acceptor	2	0
5	5_prime_UTR5KB_downstream5KB_upstreamintronNMD_transcript	31	0
5	5_prime_UTR5KB_downstream5KB_upstreamintronnc_transcript	114	0
5	5_prime_UTR5KB_downstream5KB_upstreamexonNMD_transcript	32	0
5	5_prime_UTR5KB_downstream5KB_upstreamexonnc_transcript	109	0
5	5_prime_UTR5KB_downstream5KB_upstreamexonintron	15	0
5	3_prime_UTRexonnc_transcriptNMD_transcriptcoding_sequence	1	0
5	3_prime_UTR5KB_upstreamintronnc_transcriptNMD_transcript	2	0
5	3_prime_UTR5KB_upstreamexonNMD_transcriptsplice_donor/acceptor	1	0
5	3_prime_UTR5KB_upstreamexonnc_transcriptNMD_transcript	3	0
5	3_prime_UTR5KB_upstreamexonintronnc_transcript	148	0
5	3_prime_UTR5KB_downstreamintronnc_transcriptNMD_transcript	157	0
5	3_prime_UTR5KB_downstreamexonnc_transcriptsplice_donor/acceptor	1	0
5	3_prime_UTR5KB_downstreamexonnc_transcriptNMD_transcript	22	0
5	3_prime_UTR5KB_downstreamexonintronNMD_transcript	2	0
5	3_prime_UTR5KB_downstreamexonintronnc_transcript	46	0
5	3_prime_UTR5KB_downstream5KB_upstreamnc_transcriptNMD_transcript	10	0

Table 7. The Genic Profiles of Associated and Non-Associated SNPs contd.

No of Profile Components	Genic Profiles	SNP Counts Non_Assoc	Assoc
5	3_prime_UTR5KB_downstream5KB_upstreamintronNMD_transcript	14	0
5	3_prime_UTR5KB_downstream5KB_upstreamintronnc_transcript	166	0
5	3_prime_UTR5KB_downstream5KB_upstreamexonNMD_transcript	11	0
5	3_prime_UTR5KB_downstream5KB_upstreamexonnc_transcript	94	0
5	3_prime_UTR5KB_downstream5KB_upstreamexonintron	16	0
5	3_prime_UTR5_prime_UTR5KB_upstreamintronnc_transcript	4	0
5	3_prime_UTR5_prime_UTR5KB_upstreamexonnc_transcript	61	0
5	3_prime_UTR5_prime_UTR5KB_upstreamexonintron	1	0
5	3_prime_UTR5_prime_UTR5KB_downstreamintronnc_transcript	1	0
5	3_prime_UTR5_prime_UTR5KB_downstreamexonnc_transcript	3	0
5	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamintron	1	0
5	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexon	3	0
6	5KB_downstreamexonintronnc_transcriptNMD_transcriptsplice_donor/acceptor	3	0
6	5KB_downstream5KB_upstreamintronnc_transcriptNMD_transcriptcoding_sequence	2	0
6	5KB_downstream5KB_upstreamintronnc_transcriptNMD_transcriptsplice_donor/acceptor	16	0
6	5KB_downstream5KB_upstreamexonnc_transcriptNMD_transcriptcoding_sequence	1	0
6	5KB_downstream5KB_upstreamexonnc_transcriptNMD_transcriptsplice_donor/acceptor	4	0
6	5KB_downstream5KB_upstreamexonintronNMD_transcriptcoding_sequence	1	0
6	5KB_downstream5KB_upstreamexonintronnc_transcriptsplice_donor/acceptor	1	0
6	5KB_downstream5KB_upstreamexonintronnc_transcriptNMD_transcript	253	0
6	5_prime_UTR5KB_upstreamexonnc_transcriptNMD_transcriptsplice_donor/acceptor	2	0
6	5_prime_UTR5KB_upstreamexonintronnc_transcriptNMD_transcript	59	0
6	5_prime_UTR5KB_downstream5KB_upstreamnc_transcriptNMD_transcriptsplice_donor/acceptor	1	0
6	5_prime_UTR5KB_downstream5KB_upstreamintronnc_transcriptsplice_donor/acceptor	1	0
6	5_prime_UTR5KB_downstream5KB_upstreamintronnc_transcriptNMD_transcript	64	0
6	5_prime_UTR5KB_downstream5KB_upstreamexonnc_transcriptNMD_transcript	28	0
6	5_prime_UTR5KB_downstream5KB_upstreamexonintronNMD_transcript	7	0
6	5_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcript	43	0
6	3_prime_UTR5KB_upstreamexonintronnc_transcriptNMD_transcript	2	0
6	3_prime_UTR5KB_downstreamintronnc_transcriptNMD_transcriptcoding_sequence	1	0
6	3_prime_UTR5KB_downstreamintronnc_transcriptNMD_transcriptsplice_donor/acceptor	3	0
6	3_prime_UTR5KB_downstreamexonintronnc_transcriptNMD_transcript	34	0
6	3_prime_UTR5KB_downstream5KB_upstreamintronnc_transcriptsplice_donor/acceptor	2	0
6	3_prime_UTR5KB_downstream5KB_upstreamintronnc_transcriptNMD_transcript	23	0
6	3_prime_UTR5KB_downstream5KB_upstreamexonnc_transcriptNMD_transcript	25	0
6	3_prime_UTR5KB_downstream5KB_upstreamexonintronNMD_transcript	4	0
6	3_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcript	82	0
6	3_prime_UTR5_prime_UTRexonintronnc_transcriptNMD_transcript	2	0
6	3_prime_UTR5_prime_UTR5KB_upstreamexonintronnc_transcript	7	0
6	3_prime_UTR5_prime_UTR5KB_downstreamexonnc_transcriptNMD_transcript	1	0
6	3_prime_UTR5_prime_UTR5KB_downstreamexonintronnc_transcript	2	0
6	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamnc_transcriptNMD_transcript	7	0
6	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamintronnc_transcript	6	0
6	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonNMD_transcript	1	0
6	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonnc_transcript	5	0
6	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonintron	1	0
7	5KB_downstream5KB_upstreamexonintronnc_transcriptNMD_transcriptsplice_donor/acceptor	2	0
7	5_prime_UTR5KB_downstream5KB_upstreamintronnc_transcriptNMD_transcriptsplice_donor/acceptor	1	0
7	5_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcriptsplice_donor/acceptor	1	0
7	5_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcriptNMD_transcript	22	0
7	3_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcriptNMD_transcript	19	0
7	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonnc_transcriptsplice_donor/acceptor	1	0
7	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonnc_transcriptNMD_transcript	1	0
7	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonintronNMD_transcript	2	0
7	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcript	11	0
8	3_prime_UTR5_prime_UTR5KB_downstream5KB_upstreamexonintronnc_transcriptNMD_transcript	3	0
6	5_prime_UTR5KB_downstream5KB_upstreamintronNMD_transcriptsplice_donor/acceptor	0	1

Appendix B: Data for nucleotide counts and normalised values of structural and functional features in T1D susceptibility regions

Table 8. Nucleotide counts of structural features in T1D susceptibility regions

Susceptibility Region		Nucleotide counts					
Name	Size	Intronic	Intergenic	Exonic	5' UTR	3' UTR	Non-coding Transcripts
1p13.2	840000	515286	176648	22416	4148	23819	525827
1q31.2	88456	2904	29654	740	65	1211	53882
1q32.1	247391	107236	121874	4650	2442	4469	148222
2p23.3	801217	378861	199529	15560	4796	18937	514513
2q11.2	532426	325595	249428	4800	3260	12385	665263
2q24.2	431888	318094	46951	9959	2004	4519	664417
2q32.3	142329	116140	0	2304	744	591	130070
2q33.2	147249	26720	111543	1263	365	1184	6174
3p21.31	673459	96341	324341	17453	8910	16704	370737
4p15.2	107968	0	105994	0	0	0	1974
4q27	704867	237092	361040	19508	1217	1143	70956
MHC	3808585	1015895	2500273	244932	79194	144016	1910125
6q15	239462	0	40170	0	709	0	199883
6q22.32	981805	944323	0	309	99	405	28677
6q23.3	464566	75481	365861	3575	496	1992	70458
6q25.3	209177	66777	103176	4485	1518	2937	15402
6q27	167120	0	167120	0	0	0	0
7p15.2	547320	237878	280325	9777	3760	12609	239954
7p12.2	328680	220153	71578	7139	2549	13547	274022
7p12.1	773338	295066	299652	4453	282	4587	18675
9p24.2	93008	57488	0	387	704	0	77496
10p15.1	167810	75123	56500	2293	1327	4082	117627
10p15.1	109730	68516	33731	1835	0	1063	0
10q22.3	79492	57529	15287	2906	0	3769	15287
10q23.31	270333	242698	25574	571	0	1490	0
11p15.5	239880	19579	196005	4189	3812	9130	31587
12p13.31	354623	59317	103998	3196	768	6850	186601
12q13.2	453963	191492	111833	40039	23391	45501	341317
12q24.12	1951002	1144967	465226	46296	22020	44819	1673516
13q22.2	301602	242305	13114	5692	1817	3825	233348
13q32.3	191921	70617	68388	1726	82	1620	114170
14q24.1	150324	5051	111955	1860	2179	4288	29556
14q32.2	173692	8580	121178	485	0	5430	52514
14q32.2	40708	0	1371	0	0	0	39337
15q14	182000	47116	120148	1086	934	1492	58412
15q25.1	276556	124070	34937	5761	2590	6140	220912
16p13.13	321999	221530	33052	6446	985	907	323315
16p13.13	41669	31000	0	420	623	108	198
16p11.2	753180	258797	380955	39398	20092	18913	391128
16q23.1	325781	199731	78851	8542	5309	5596	239277

Table 8. Nucleotide counts of structural features in T1D susceptibility regions contd.

		Nucleotide counts					
Susceptibility Region		Intronic	Intergenic	Exonic	5' UTR	3' UTR	Non-coding
Name	Size						Transcripts
17q12	873850	529318	315685	38866	14926	40646	578238
17q21.2	141100	27344	87578	3960	1281	10880	65207
18p11.21	189722	96989	37617	1584	816	3176	164148
18q22.2	92095	80445	0	768	1571	9311	64651
19p13.2	233021	130888	54890	22708	6610	14787	135934
19q13.32	172006	84482	33319	10086	3192	6653	19806
19q13.33	185652	73608	33478	16986	6093	9280	50640
20p13	263117	81330	170156	4468	1609	4148	39582
21q22.3	79544	46882	27615	2555	1475	1017	52025
22q12.2	862028	459735	184993	19170	5093	29034	684547
22q12.3	98116	21988	43937	2946	1421	5899	75577
Xq28	618122	350331	118944	17921	3528	20134	541509

Table 9. Counts of functional features in T1D susceptibility regions

Susceptibility Region		Non-coding	Protein	SNP	Regulatory
Name	Size	Genes	Coding Genes	Counts	Nucleotides
1p13.2	840000	6	10	5744	91129
1q31.2	88456	2	1	772	17249
1q32.1	247391	4	5	1851	60963
2p23.3	801217	10	8	4356	94307
2q11.2	532426	3	4	3251	49587
2q24.2	431888	2	5	1102	36980
2q32.3	142329	2	1	1096	15249
2q33.2	147249	0	2	2162	11817
3p21.31	673459	6	13	5385	108354
4p15.2	107968	1	0	854	21794
4q27	704867	3	4	5599	45444
MHC	3808585	150	157	115045	277448
6q15	239462	2	1	5334	47811
6q22.32	981805	7	2	4025	110101
6q23.3	464566	10	1	1892	26908
6q25.3	209177	3	3	1453	38471
6q27	167120	0	0	2805	11198
7p15.2	547320	8	11	6382	92907
7p12.2	328680	2	4	2747	37270
7p12.1	773338	10	1	4108	72127
9p24.2	93008	1	19	1186	15995
10p15.1	167810	7	3	2026	29658
10p15.1	109730	0	1	1207	11743
10q22.3	79492	0	1	703	20399
10q23.31	270333	0	1	1864	28638
11p15.5	239880	4	5	2561	47247
12p13.31	354623	11	4	2993	40850
12q13.2	453963	16	27	3122	89949
12q24.12	1951002	29	19	12835	195768
13q22.2	301602	5	2	1626	41715
13q32.3	191921	5	2	5526	43122
14q24.1	150324	2	2	1375	46600
14q32.2	173692	0	1	327	3708
14q32.2	40708	4	1	1775	7237
15q14	182000	3	2	1344	22029
15q25.1	276556	5	5	2813	48980
16p13.13	321999	6	3	3331	86921
16p13.13	41669	3	2	3107	13345
16p11.2	753180	20	24	446	116550
16q23.1	325781	8	8	3371	64832

Table 9. Counts of functional features in T1D susceptibility regions contd.

Susceptibility Region		Non-coding	Protein	SNP	Regulatory
Name	Size	Genes	Coding Genes	Counts	Nucleotides
17q12	873850	14	24	6169	167311
17q21.2	141100	3	4	977	15748
18p11.21	189722	6	1	1829	28313
18q22.2	92095	0	2	899	7558
19p13.2	233021	4	14	1923	58846
19q13.32	172006	6	5	1403	45786
19q13.33	185652	3	14	1630	59181
20p13	263117	8	4	2544	22634
21q22.3	79544	0	4	1028	20732
22q12.2	862028	15	15	6778	136086
22q12.3	98116	2	3	1186	35188
Xq28	618122	14	14	1660	62980

Table 10. Standard residuals of structural features in T1D susceptibility regions

Susceptibility Region		Standard Residuals				
ID	Name	3' UTR	5' UTR	Exonic	Intergenic	Intronic
1.1	1p13.2	-0.0719	-0.1769	-0.1353	-0.1595	0.2873
1.2	1q31.2	-0.0478	-0.0358	-0.0333	-0.0447	0.0181
1.3	1q32.1	-0.0618	-0.0465	-0.0576	-0.0340	0.0715
2.1	2p23.3	-0.0925	-0.1588	-0.1535	-0.1403	0.1773
2.2	2q11.2	-0.0748	-0.1092	-0.1293	-0.0654	0.1671
2.3	2q24.2	-0.1001	-0.0993	-0.0827	-0.1112	0.2119
2.4	2q32.3	-0.0627	-0.0410	-0.0406	-0.0582	0.1043
2.5	2q33.2	-0.0602	-0.0470	-0.0461	-0.0329	0.0250
3.1	3p21.31	-0.0789	-0.0741	-0.1134	-0.0566	-0.0391
4.1	4p15.2	-0.0590	-0.0416	-0.0413	-0.0057	0.0110
4.2	4q27	-0.1769	-0.1793	-0.1130	-0.0492	0.0764
6.1	MHC	0.0143	0.0096	0.0205	0.0248	-0.0568
6.2	6q15	-0.0864	-0.0663	-0.0746	-0.0658	-0.0202
6.3	6q22.32	-0.2391	-0.2644	-0.2616	-0.2682	-0.1894
6.4	6q23.3	-0.1218	-0.1267	-0.1171	0.0129	-0.0077
6.5	6q25.3	-0.0628	-0.0484	-0.0486	-0.0322	0.0453
6.6	6q27	-0.0713	-0.0567	-0.0563	0.0049	-0.0031
7.1	7p15.2	-0.0766	-0.1068	-0.1127	-0.0433	0.1145
7.2	7p12.2	-0.0255	-0.0660	-0.0681	-0.0751	0.1509
7.3	7p12.1	-0.1710	-0.2086	-0.1918	-0.0918	0.1108
9.1	9p24.2	-0.0558	-0.0289	-0.0359	-0.0459	0.0647
10.1	10p15.1	-0.0475	-0.0402	-0.0471	-0.0412	0.0624
10.2	10p15.1	-0.0531	-0.0420	-0.0342	-0.0361	0.0704
10.3	10q22.3	-0.0309	-0.0343	-0.0222	-0.0361	0.0680
10.4	10q23.31	-0.0841	-0.0832	-0.0801	-0.0796	0.0486
11.1	11p15.5	-0.0329	-0.0273	-0.0576	-0.0013	-0.0032
12.1	12p13.31	-0.0702	-0.0951	-0.0908	-0.0682	0.0042
12.2	12q13.2	0.1362	0.1651	0.0345	-0.0898	0.0961
12.4	12q24.12	-0.1806	-0.2361	-0.3195	-0.3177	0.5736

Table 10. Standard residuals of structural features in T1D susceptibility regions contd.

Susceptibility Region		Standard Residuals				
ID	Name	3' UTR	5' UTR	Exonic	Intergenic	Intronic
13.1	13q22.2	-0.0769	-0.0683	-0.0671	-0.0926	0.1766
13.2	13q32.3	-0.0670	-0.0621	-0.0555	-0.0423	0.0527
14.1	14q24.1	-0.0426	-0.0249	-0.0444	-0.0138	0.0053
14.2	14q32.2	-0.0408	-0.0584	-0.0560	-0.0158	0.0029
14.3	14q32.2	-0.0449	-0.0243	-0.0242	-0.0322	0.0269
15.1	15q14	-0.0657	-0.0488	-0.0556	-0.0183	0.0346
15.2	15q25.1	-0.0581	-0.0521	-0.0605	-0.0773	0.0793
16.1	16p13.13	-0.0983	-0.0840	-0.0692	-0.0894	0.1537
16.2	16p13.13	-0.0445	-0.0167	-0.0234	-0.0330	0.0268
16.3	16p11.2	-0.0826	0.0467	-0.0440	-0.0531	0.0839
16.4	16q23.1	-0.0716	-0.0304	-0.0616	-0.0714	0.1337
17.1	17q12	0.0199	-0.0495	-0.0768	-0.1239	0.2629
17.2	17q21.2	-0.0019	-0.0339	-0.0335	-0.0216	0.0270
18.1	18p11.21	-0.0574	-0.0522	-0.0555	-0.0545	0.0763
18.2	18q22.2	-0.0009	-0.0177	-0.0341	-0.0456	0.0850
19.1	19p13.2	0.0018	0.0098	0.0197	-0.0581	0.0956
19.2	19q13.32	-0.0332	-0.0177	-0.0163	-0.0518	0.0696
19.3	19q13.33	-0.0206	0.0154	0.0084	-0.0552	0.0568
20.1	20p13	0.9086	-0.0290	-0.0740	-0.0547	-0.0219
21.1	21q22.3	-0.0470	-0.0157	-0.0236	-0.0310	0.0587
22.1	22q12.2	-0.0459	-0.1706	-0.1542	-0.1615	0.2335
22.3	22q12.3	-0.0222	-0.0211	-0.0267	-0.0289	0.0325
X.2	Xq28	-0.0472	-0.1278	-0.0974	-0.1279	0.1959

Table 11. Standard residuals of functional features in T1D susceptibility regions

		Standard Residuals				
Susceptibility Region		Regulatory	Non-coding	Protein	No of SNPs	TFBS
ID	Name	Nucleotides	Genes	Genes	SNPS	SNPs
1.1	1p13.2	0.077140168	-0.195933071	-0.1764042	-0.189265	-0.35908
1.2	1q31.2	-0.000350589	-0.033431576	-0.0481706	-0.04505	-0.0586
1.3	1q32.1	0.117281223	-0.060103141	-0.0619344	-0.075309	-0.10749
2.1	2p23.3	0.098337141	-0.159504489	-0.1795674	-0.191657	-0.32404
2.2	2q11.2	0.004676059	-0.138514838	-0.1386798	-0.134227	-0.22507
2.3	2q24.2	-0.015507202	-0.119875483	-0.1074873	-0.127832	-0.20503
2.4	2q32.3	-0.02109244	-0.046991735	-0.061472	-0.05567	-0.09511
2.5	2q33.2	-0.034698266	-0.061563462	-0.0563173	-0.047631	-0.09714
3.1	3p21.31	0.181060197	-0.154013696	-0.1161764	-0.15085	-0.25229
4.1	4p15.2	-0.138816114	-0.195252612	-0.2067337	-0.198068	-0.31525
4.2	4q27	0.0963705	-0.031676197	-0.0338799	-0.007959	-0.04998
6.1	MHC	0.0026	0.0165	0.0266	0.0201	-0.608
6.2	6q15	-0.114159019	-0.257838583	-0.2682954	-0.227744	-0.4335
6.3	6q22.32	0.240979342	-0.09361811	-0.1335365	-0.109436	-0.219
6.4	6q23.3	0.003930396	-0.010691601	-0.0781802	-0.065627	-0.04921
6.5	6q25.3	0.038070454	-0.064909449	-0.0728484	-0.076925	-0.12111
6.6	6q27	-0.042644533	-0.067290026	-0.0746735	-0.047716	-0.09462
7.1	7p15.2	0.157071943	-0.108930415	-0.0977712	-0.110726	-0.25026
7.2	7p12.2	0.011464647	-0.093897407	-0.0883743	-0.087794	-0.14341
7.3	7p12.1	0.025397643	-0.152487176	-0.21727	-0.18686	-0.32683
9.1	9p24.2	-0.006013851	-0.041244008	0.06535516	-0.042587	-0.07
10.1	10p15.1	0.024440601	-0.020072123	-0.0550245	-0.053941	-0.09133
10.2	10p15.1	-0.025539934	-0.052119703	-0.0534232	-0.046575	-0.07452
10.3	10q22.3	0.013254713	-0.044508616	-0.0459574	-0.043414	-0.06443
10.4	10q23.31	-0.004990294	-0.092544448	-0.0930766	-0.080918	-0.13839
11.1	11p15.5	0.069731751	-0.058212577	-0.0600799	-0.067264	-0.10678
12.1	12p13.31	0.017850893	-0.040427417	-0.0947797	-0.092126	-0.18268
12.2	12q13.2	0.16986245	-0.032098561	0.02718975	-0.115779	-0.16651
12.4	12q24.12	0.175196451	-0.322245648	-0.3933889	-0.404708	-0.74118

Table 11. Standard residuals of functional features in T1D susceptibility regions contd.

Susceptibility Region		Regulatory	Non-coding	Protein	No of SNPs	TFBS
ID	Name	Nucleotides	Genes	Genes	SNPS	SNPs
13.1	13q22.2	0.03428786	-0.067081711	-0.0944276	-0.090785	-0.15367
13.2	13q32.3	0.066911753	-0.039474341	-0.067347	-0.029531	-0.0989
14.1	14q24.1	0.089896908	-0.049004125	-0.0570766	-0.055238	-0.0865
14.2	14q32.2	-0.07056804	-0.068219325	-0.0692156	-0.070176	-0.10805
14.3	14q32.2	-0.024441982	-0.008079783	-0.0363815	-0.024424	-0.05081
15.1	15q14	-0.006621075	-0.050310499	-0.0648975	-0.063408	-0.10195
15.2	15q25.1	0.066764686	-0.060777481	-0.0691353	-0.074221	-0.10048
16.1	16p13.13	0.19209902	-0.065549092	-0.0930942	-0.081051	-0.12875
16.2	16p13.13	-0.002668456	-0.014988339	-0.0302493	-0.013085	-0.03216
16.3	16p11.2	0.190574366	-0.080746619	-0.0657961	-0.213664	-0.25661
16.4	16q23.1	0.111534035	-0.053167711	-0.0621809	-0.081647	-0.1184
17.1	17q12	0.343217982	-0.151119987	-0.0955898	-0.194012	-0.32543
17.2	17q21.2	-0.018985171	-0.040015722	-0.0420603	-0.056397	-0.08746
18.1	18p11.21	0.014088391	-0.032254173	-0.0731735	-0.061118	-0.09085
18.2	18q22.2	-0.0361938	-0.047680867	-0.0426997	-0.044854	-0.07439
19.1	19p13.2	0.113260814	-0.056486126	-0.0010616	-0.071099	-0.05871
19.2	19q13.32	0.081516347	-0.027794949	-0.0433216	-0.060402	-0.07402
19.3	19q13.33	0.126367689	-0.051229729	0.01063396	-0.061833	-0.07488
20.1	20p13	-0.024817673	-0.037394804	-0.0721866	-0.073207	-0.13065
21.1	21q22.3	0.014441875	-0.044521705	-0.0268619	-0.040602	-0.04064
22.1	22q12.2	0.233644143	-0.141477651	-0.1499958	-0.185771	-0.35389
22.3	22q12.3	0.061879909	-0.035863056	-0.0378168	-0.043861	0.006461
X.2	Xq28	0.031420702	-0.086751706	-0.0961441	-0.169428	-0.28185

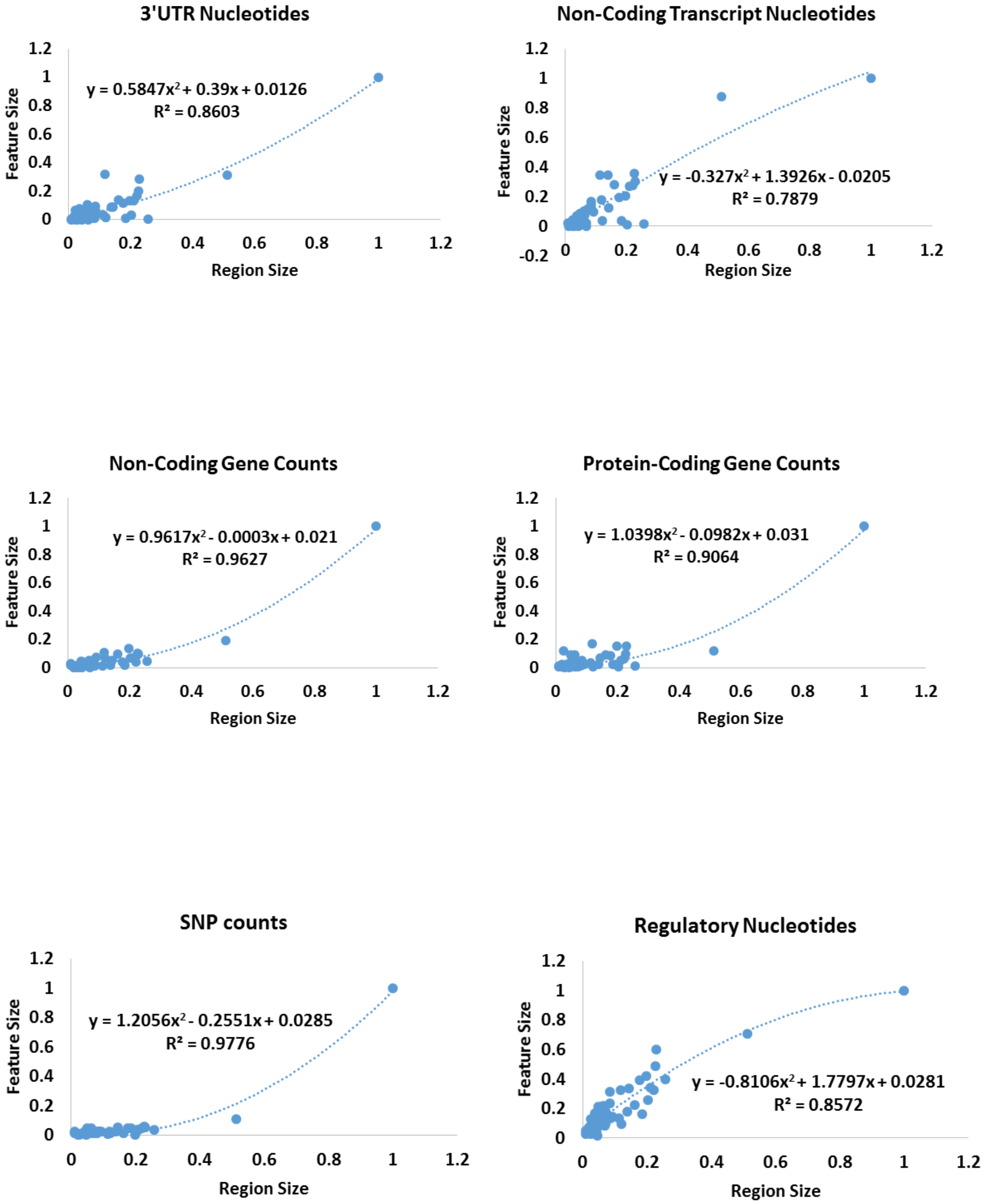


Figure 1. Scatter plots showing data–fitting using a second order polynomial regression for 3' UTR, non-coding transcript and regulatory nucleotides, as well as Non-coding gene, protein coding gene and SNP counts.

Appendix C: Methods for the SNP Sensitivity test

C.1. Establishing the background model of SNP local environment

The simplest model of the DNA sequence assumes that the sequence has been produced by a random process where any of the four nucleotides randomly occurs at each position in the sequence. The probability of choosing any one of the four nucleotides depends on a multinomial sequence model distribution, and the four nucleotides A, C, G, T are chosen with four parameters: p(A), p(C), p(G), and p(T) respectively. However, an assumption of a multinomial model of the DNA sequence is that the probability of choosing a particular nucleotide (for example "A") at a particular position in the sequence only depends on the predetermined frequency of that nucleotide (p(A) here), and does not depend at all on the identity of nucleotides found at adjacent positions in the sequence. This assumption does not hold true for all DNA sequences. In regulatory sequences, which are characterised by recurrent binding motifs, the probability of finding a certain nucleotide at a particular position in the sequence does depend on the identity of the nucleotides are found at adjacent positions in the sequence. In order to characterise this type of DNA sequence a Markov sequence model is used to give a more accurate representation. The Markov model has more parameters than the multinomial model, where the four sets of probabilities p(A), p(C), p(G), and p(T) differ according to whether the previous nucleotide was "A", "C", "G" or "T", and will be later explained

For each regulatory sequence containing a TFBS-SNP therefore, the order of nucleotide dependencies in the sequence is fitted on the basis of the Markov process of order $\boldsymbol{m}$. A model based on Markov orders means that the probability of a nucleotide $\boldsymbol{a_l}$, at position $\boldsymbol{l}$ in the sequence, depends on the $\boldsymbol{m}$ previous nucleotides in the sequence. If $\boldsymbol{m} = 1$, this mean that the identity of a nucleotide depends on the identity of the previous nucleotide. For example, if "A" is the preceding nucleotide, then the probabilities of "A", "C", "G" or "T", are symbolised as p(A|A), p(C|A), p(G|A), or p(T|A) . Similarly, if $\boldsymbol{m} = 2$, this mean that the identity of a nucleotide depends on the identity of the 2 previous nucleotides, if "C" and "A" are the preceding nucleotides, then probabilities of "A", "C", "G" or "T", are symbolised as p (A| CA), p (C| CA), p (G| CA), or p (T| CA). The model is described with a transition matrix.

Given a sequence $\boldsymbol{S_i}$ with nucleotide dependency of order $\boldsymbol{m = 1,}$ the nucleotide dependency is described with a first order transition matrix. For instance the sequence in Figure 2a, is summarized with a first order transition matrix (Figure 2b).

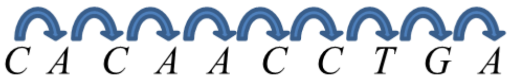


Figure 2a. A short nucleotide sequence assumed to be characterised by first order Markov dependency

next nucleotide, n'

nucleotide, n

	A	C	G	T
A	I	II		
C	II	I		I
G	I			
T			I	

Figure 2b. First order transition matrix summarizing nucleotide dependency of sequence in Figure 1

The rows of the matrix represent the nucleotide found at the previous position in the sequence, while the columns represent the nucleotides that could be found at the current position in the sequence. When $\boldsymbol{m}$ is 1, the transition matrix is computed as the frequencies of di-nucleotide base pairs $\boldsymbol{d_i}$ in $\boldsymbol{S_i}$, when $\boldsymbol{m}$ is 2 the matrix is computed as the frequencies of tri-nucleotide base pairs $\boldsymbol{t_i}$ and so on.

$$\boldsymbol{Transition\,matrix}\ (\boldsymbol{m} = \boldsymbol{1}) = \begin{pmatrix} freq\ (A, A') & freq\ (A, C') & freq\ (A, G') & freq\ (A, T') \\ freq\ (C, A') & freq\ (C, C') & freq\ (C, G') & freq\ (C, T') \\ freq\ (G, A') & freq\ (G, C') & freq\ (G, G') & freq\ (G, T') \\ freq\ (T, A') & freq\ (T, C') & freq\ (T, G') & freq\ (T, T') \end{pmatrix}$$

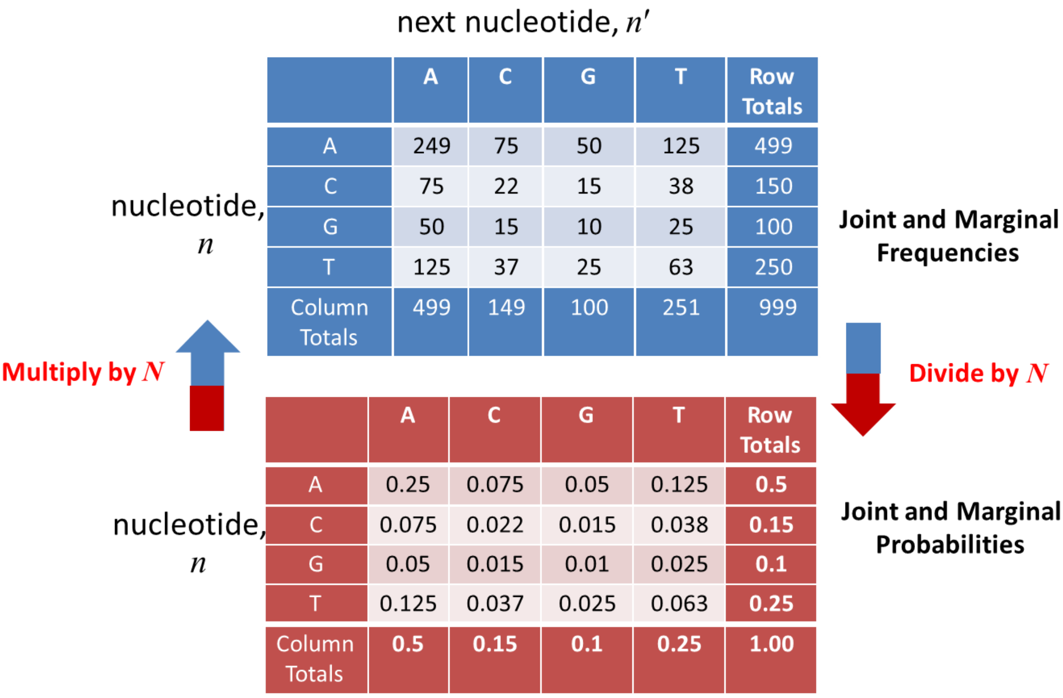
$$\boldsymbol{Transition\,matrix}\ (\boldsymbol{m}=\boldsymbol{2}) = \begin{bmatrix} freq\left(AA,A^{'}\right) & freq\left(AC,A^{'}\right) & freq\left(AG,A^{'}\right) & freq\left(AT,A^{'}\right) \\ freq\left(CA,A^{'}\right) & freq\left(CC,A^{'}\right) & freq\left(CG,A^{'}\right) & freq\left(CT,A^{'}\right) \\ freq\left(GA,A^{'}\right) & freq\left(GC,A^{'}\right) & freq\left(GG,A^{'}\right) & freq\left(GT,A^{'}\right) \\ freq\left(TA,A^{'}\right) & freq\left(TC,A^{'}\right) & freq\left(TG,A^{'}\right) & freq\left(TT,A^{'}\right) \\ freq\left(AA,C^{'}\right) & freq\left(AC,C^{'}\right) & freq\left(AG,C^{'}\right) & freq\left(AT,C^{'}\right) \\ freq\left(CA,C^{'}\right) & freq\left(CC,C^{'}\right) & freq\left(CG,C^{'}\right) & freq\,(CT,C) \\ freq\left(GA,C^{'}\right) & freq\left(GC,C^{'}\right) & freq\,(GG,C) & freq\left(GT,C^{'}\right) \\ freq\left(TA,C^{'}\right) & freq\left(TC,C^{'}\right) & freq\left(TG,C^{'}\right) & freq\,(TT,C) \\ freq\left(AA,G^{'}\right) & freq\left(AC,G^{'}\right) & freq\left(AG,G^{'}\right) & freq\left(AT,G^{'}\right) \\ freq\left(CA,G^{'}\right) & freq\left(CC,G^{'}\right) & freq\left(CG,G^{'}\right) & freq\left(CT,G^{'}\right) \\ freq\left(GA,G^{'}\right) & freq\left(GC,G^{'}\right) & freq\left(GG,G^{'}\right) & freq\left(GT,G^{'}\right) \\ freq\left(TA,G^{'}\right) & freq\left(TC,G^{'}\right) & freq\left(TG,G^{'}\right) & freq\left(TT,G^{'}\right) \\ freq\left(AA,T^{'}\right) & freq\left(AC,T^{'}\right) & freq\left(AG,T^{'}\right) & freq\left(AT,T^{'}\right) \\ freq\left(CA,T^{'}\right) & freq\left(CC,T^{'}\right) & freq\left(CG,T^{'}\right) & freq\left(CT,T^{'}\right) \\ freq\left(GA,T^{'}\right) & freq\left(GC,T^{'}\right) & freq\left(GG,T^{'}\right) & freq\left(GT,T^{'}\right) \\ freq\left(TA,T^{'}\right) & freq\left(TC,T^{'}\right) & freq\left(TG,T^{'}\right) & freq\left(TT,T^{'}\right) \end{bmatrix}$$

The transition frequencies in the transition matrices are converted to transition probabilities by dividing them by the grand total $\boldsymbol{N}$. The row and column totals are also converted to marginal probabilities by dividing by the grand total $\boldsymbol{N}$ (Figure 3).

Transition probabilities: $prob\left(n_i, n_j'\right) = freq\ (n_i, n_j')/N$

Marginal probabilities (Rows): $prob\ (n_i)\ =\ R_i/N\ =\ freq\ (n_i)/N\ =\ \sum_{j=1}^{4} freq\ (n_i, n_j')/$ N

Marginal probabilities (Columns): $prob\left(n'_j\right)\ =\ C_j/N\ =\ freq\ (n'_j)/N\ =$ $\sum_{i=1}^{4} freq\ (n_i, n_j')/N$



next nucleotide, n'

nucleotide, n

	A	C	G	T	Row Totals
A	249	75	50	125	499
C	75	22	15	38	150
G	50	15	10	25	100
T	125	37	25	63	250
Column Totals	499	149	100	251	999

Joint and Marginal Frequencies

Multiply by N

Divide by N

nucleotide, n

	A	C	G	T	Row Totals
A	0.25	0.075	0.05	0.125	**0.5**
C	0.075	0.022	0.015	0.038	**0.15**
G	0.05	0.015	0.01	0.025	**0.1**
T	0.125	0.037	0.025	0.063	**0.25**
Column Totals	**0.5**	**0.15**	**0.1**	**0.25**	**1.00**

Joint and Marginal Probabilities

Figure 3. Data for a fictional sequence of L = 1000 nucleotides, where n_1 = C and n_{1000} = T

The joint probabilities assume independence, that is the occurrence of $\boldsymbol{n'}$ is independent of the occurrence of $\boldsymbol{n}$ before it. However, to measure or establish if there is dependency between nucleotides in the regulatory sequence, a Chi squared test can be done. The Chi squared statistic X^2can be obtained as the sum of residuals using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where:

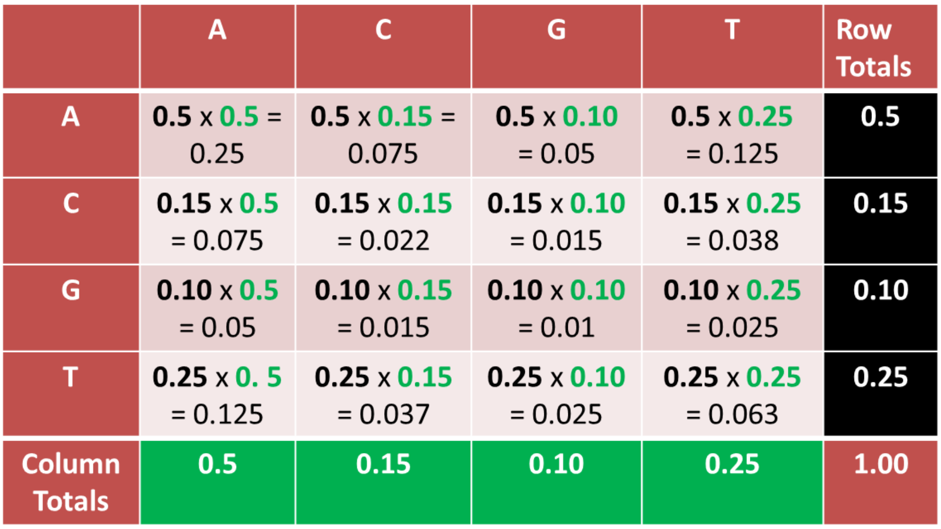
O_i is the observed frequency (number of observations) of the nucleotide pair in each cell, calculated as transition frequencies:

$$O_i = freq\ (n_i, n_j')$$

and:

E_i is the theoretical frequency of each cell given the hypothesis of independence. It is calculated as joint frequencies, which is the product of the corresponding marginal probabilities (Figure 4).

$$E_i = freq\ (n_i, n_j') = freq\ (n_i) \text{ x } freq\ (n'_j)$$

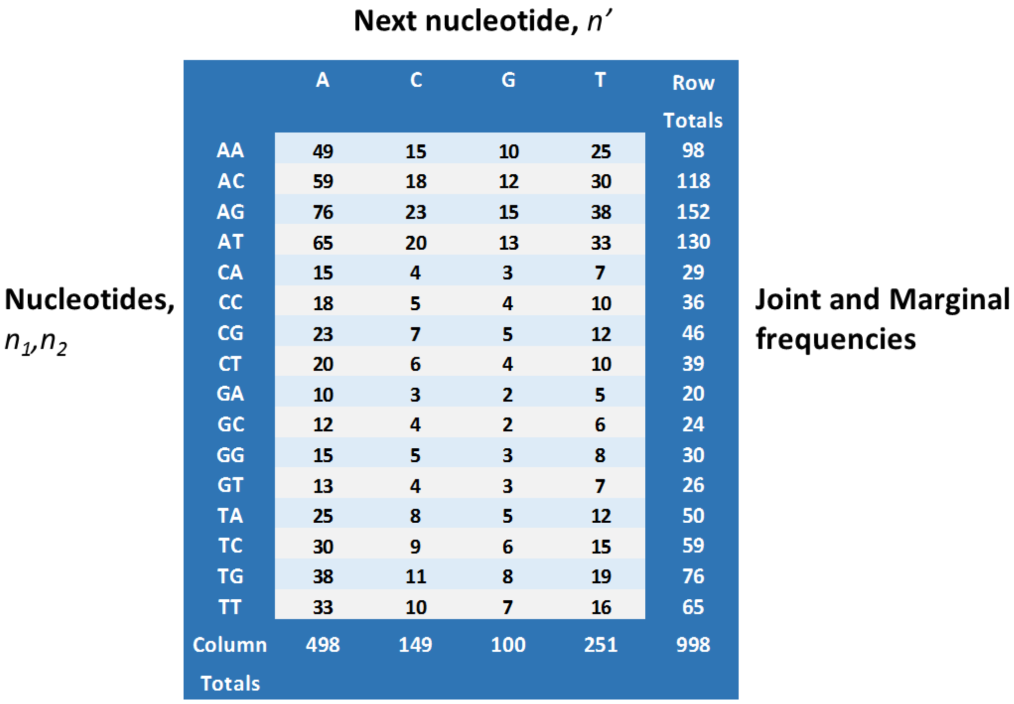


	A	C	G	T	Row Totals
A	0.5 x 0.5 = 0.25	0.5 x 0.15 = 0.075	0.5 x 0.10 = 0.05	0.5 x 0.25 = 0.125	0.5
C	0.15 x 0.5 = 0.075	0.15 x 0.15 = 0.022	0.15 x 0.10 = 0.015	0.15 x 0.25 = 0.038	0.15
G	0.10 x 0.5 = 0.05	0.10 x 0.15 = 0.015	0.10 x 0.10 = 0.01	0.10 x 0.25 = 0.025	0.10
T	0.25 x 0. 5 = 0.125	0.25 x 0.15 = 0.037	0.25 x 0.10 = 0.025	0.25 x 0.25 = 0.063	0.25
Column Totals	0.5	0.15	0.10	0.25	1.00

Figure 35. Joint probabilities = product of the corresponding marginal probabilities

To test for independence, X^2 is used to obtain a $\boldsymbol{p}$-value by comparing the value of X^2 to a Chi-squared distribution (df =[(r-1)(c-1)], $\boldsymbol{\alpha}$ = 0.005). If the $\boldsymbol{p}$-value is higher than $\boldsymbol{\alpha}$ ($\boldsymbol{p} \geq \boldsymbol{\alpha}$), then $\boldsymbol{p}$ is not statistically significant for $\boldsymbol{m} = 1$. This means that the nucleotides in the sequence occur independently of each other and the Markov order of the sequence m is established as zero ($\boldsymbol{m} = 0$). Alternatively, if the $\boldsymbol{p}$-value is less than $\boldsymbol{\alpha}$ ($\boldsymbol{p} \leq \boldsymbol{\alpha}$), then $\boldsymbol{p}$ is statistically significant. This means that there is dependency between the nucleotides in the sequence, and m is not established.

When $\boldsymbol{m}$ is not established, the cylcle is repeated assuming a higher Markov order $\boldsymbol{m} = 2$. An order 2 transition matrix is computed and the O_i and $E_i$values of tri-nucleotide words are calculated (Figure 5). Subsequently, the $\boldsymbol{p}$-value is obtained for X^2 If $\boldsymbol{p} \geq \boldsymbol{\alpha}$, then $\boldsymbol{p}$ is not statistically significant for $\boldsymbol{m} = 2$, and the Markov order of the sequence is established as $\boldsymbol{m} = 1$ (i.e. a first-order Markov model fits the data). This means that the occurrence of nucleotides in the sequence depend on the identity of the preceding nucleotide. Else, if $\boldsymbol{p} \leq \boldsymbol{\alpha}$, then $\boldsymbol{p}$ is significant, and the order of dependency of nucleotides in the sequence has not been established. This cycle is then repeated till the Markov order $\boldsymbol{m}$ is established by a non-significant $\boldsymbol{p}$-value.



Next nucleotide, *n′*

Nucleotides, n_1,n_2

Joint and Marginal frequencies

	A	C	G	T	Row Totals
AA	49	15	10	25	98
AC	59	18	12	30	118
AG	76	23	15	38	152
AT	65	20	13	33	130
CA	15	4	3	7	29
CC	18	5	4	10	36
CG	23	7	5	12	46
CT	20	6	4	10	39
GA	10	3	2	5	20
GC	12	4	2	6	24
GG	15	5	3	8	30
GT	13	4	3	7	26
TA	25	8	5	12	50
TC	30	9	6	15	59
TG	38	11	8	19	76
TT	33	10	7	16	65
Column Totals	498	149	100	251	998

Figure 5. Second order transition matrix summarizing nucleotide dependency for a fictional sequence of L = 1000 nucleotides.

C.2. Algorithm 1 to Fit Markov dependency of nucleotides in SNP-regulatory-sequence

The Markov algorithm is designed to predict the Markov order of dependency m of nucleotides in the local environment of the SNP.

Algorithm 1. Execute Markov dependency ($\boldsymbol{m}$)

Given a sequence $\boldsymbol{S_i}$, that contains a SNP $\boldsymbol{Z_i}$ with reference allele $\boldsymbol{Z_{i,r}}$, and mutant allele/s $\boldsymbol{Z_{i,m}}$;

SNP $\boldsymbol{Z_i}$ is in a TFBS $\boldsymbol{T_i}$, within a regulatory region $\boldsymbol{R_i}$;

```
1: for all $\boldsymbol{S_r}$ (with reference allele $\boldsymbol{Z_{i,r}}$) in the file do steps 2-23
2:        for nucleotide base $n$ in $\boldsymbol{S_r}$ do
3:                compute frequencies "$\boldsymbol{A}$", "$\boldsymbol{C}$", "$\boldsymbol{G}$", and "$\boldsymbol{T}$"
4:                compute transition matrix for all possible duplets $\boldsymbol{d_i}$ in $\boldsymbol{S_i}$
5:                observed frequency ($\boldsymbol{O_i}$) of $\boldsymbol{d_i}$== transition frequency
6:                compute expected frequency ($\boldsymbol{E_i}$) of $\boldsymbol{d_i}$
7:                compute chi-square statistic ( $\boldsymbol{\chi^2}_{observed}$)
8:                compute $\boldsymbol{p}$ value, $\boldsymbol{\alpha} = 0.05$
9:        if $\boldsymbol{p} \geq \alpha$ then
10:               Report $\boldsymbol{\chi^2}_{observed}$ is not significant
11:       end if
12:       Print $\boldsymbol{p}$-value, "Not Significant", "Markov order is 0"
13:       else
14:       if $\boldsymbol{p} \leq \alpha$ , $\boldsymbol{p}$ is significant
15:               compute transition matrix for all possible triplets $\boldsymbol{t_i}$ in $\boldsymbol{S_i}$
16                repeat steps 5-8
17        if $\boldsymbol{p} \geq \alpha$ then
18:               Report $\boldsymbol{\chi^2}_{observed}$ is not significant,
19:       end if
20:       Print $\boldsymbol{p}$-value, "Not Significant", "Markov order is 1"
21:       else
22:       if $\boldsymbol{p} \leq \alpha$ , $\boldsymbol{p}$ is significant
23:       continue cycle till Markov order is established
24: Repeat process for $\boldsymbol{S_m}$ (with mutant allele $\boldsymbol{Z_{i,m}}$) from steps 2 to 23
```

C.2.1. Implementation of Algorithm 1.

The algorithm was developed and implemented using the Python 2.7.3 programming software. The numpy and scipy modules were also applied for scripting. The SNP-sequences $\boldsymbol{S_i}$ are called in fasta file format. In the first instance, the Markov order of dependency of nucleotides $(\boldsymbol{m})$ in of the sequence $\boldsymbol{S_i}$ is assumed to be 1. This means that the occurrence of each nucleotide base (A, C, G, or T) in $\boldsymbol{S_i}$ depends on the identity of the one before it. However, if $\boldsymbol{p}$ is not significant, then m is 0 and the nucleotides occur independently. Else the $\boldsymbol{m}$ of $\boldsymbol{S_i}$ is assumed to be order 2, the cycle is repeated till Markov order is established. The algorithm also computes the chi-squared statistic X^2 using a numpy function and determines the p-value by comparing the value of the X^2 statistic and the degrees of freedom to a chi-squared distribution.

C.3. Motif Representation (probability) of motifs in local background

The motif representation of each binding sequence is computed as a Standard residual value $\boldsymbol{SR_x}$, the calculation is based on the established $\boldsymbol{m}$ of the SNP-regulatory sequence. Motif representation is the probability that a particular motif will occur and entails computing the observed probabilities $(\boldsymbol{O_i})$ and expectancy values $(\boldsymbol{E_i})$.

$Standard\ Residual\ (\boldsymbol{SR_x}) = Residual_i/Sqrt\ (\boldsymbol{E_i})$, $\boldsymbol{SR_i} > |\mathbf{2}|$ is high

$Residual_x = \boldsymbol{O_i} - \boldsymbol{E_i}$

where for $\boldsymbol{m} = 0$;

$\boldsymbol{O}_i =$ observed probabilities $=, prob\ (n)$

$\boldsymbol{E}_i =$ joint probabilities $= prob\ (n_i) * prob\ (n'_j)$

for $\boldsymbol{m} > 0$;

$\boldsymbol{O}_i =$ transition probabilities $= prob\ (n_i, n_j')$

$\boldsymbol{E}_i =$ conditional probabilities $= prob\ (n_i,) * prob\ (n'_j |\ n_i)$

For the theoretical probability $\boldsymbol{E_i}$, joint probabilities assume independence of nucleotides in the sequence. But if there is dependency within the sequence, then $\boldsymbol{E_i}$, is computed as a conditional probability and depends on the established Markov order of the sequence. Thus, the probability of a sequence in which the identity of the second nucleotide depends on the identity of the first one.is the probability that the sequence starts with nucleotide n_i multiplied by the probability of the transition n_i → n′j. Given a word or sub-motif of length 3 ($Ø_{k=3}$), if $\boldsymbol{m}$ = 1, the conditional probability of "ACG"is $prob\ (A)\ *\ prob$ of the transition $\left(A_1{}^{23}\ \rightarrow\ C_2\ \right)\ *\ prob\ (C_2\ \rightarrow\ G_3)$ denoted as:

$\boldsymbol{prob}$ ("ACG") = $prob\ (\ n_{i=1}\ =\ \text{"A"})*prob\ (\ n_{i=2}\ =\ \text{"C"}|\ n_{i=1}\ =\ \text{"A"})*prob\ (\ n_{i=3}\ =\ \text{"G"}|\ n_{i=2}\ =\ \text{"C"})$

For $\boldsymbol{m}$ = 2,

$\boldsymbol{prob}$ ("ACG") = $prob\ (\ n_{i=2}\ =\ \text{"C"}|\ n_{i=1}\ =\ \text{"A"})\ *\ prob\ (\ n_{i=3}\ =\ \text{"G"}|\ n_{i=1,2}\ =\ \text{"AC"})$

For each SNP environment, the observed and conditional probabilities are determined using the appropriate transition matrices, so as to yield the value of $\boldsymbol{SR_x}$

Algorithm 2. Compute Observed Probabilities $(\boldsymbol{O_i})$ and Expectancies $(\boldsymbol{E_i})$ of tri-mers in sub-sequence

Given a sequence $\boldsymbol{S_i}$, that contains a SNP $\boldsymbol{Z_i}$ with reference allele $\boldsymbol{Z_{i,r}}$, and mutant allele/s $\boldsymbol{Z_{i,m}}$;

Create a sub-sequence $\boldsymbol{T_i}$, that also contains SNP $\boldsymbol{Z_i}$ in a TFBS $\boldsymbol{T_i}$, within a regulatory region $\boldsymbol{R_i}$;

```
1: for all $T_i$ in the file do
2:        for nucleotide base in $T_r$ do
3:                generate k-mers (k=3) in single-step sliding window
4:        for each k-mer do
5:                from $S_i$ compute observed probability ($O_i$)
6:                compute Expectancy ($E_i$) given $m$ of $S_i$
7:         else (if $m$ not known)
8:                compute ($E_i$) assuming $m = 0$
9:                 compute ($E_i$) assuming $m = 1$
                  compute ($E_i$) assuming $m = 2$
end
```

[23] where A1 represents a nucleotide of identity "A" at position 1

Repeat process for $\boldsymbol{T_m}$

C.3. Change in motif representation.

The change in representation of the binding sequence is calculated as the difference between the SR value of the reference allele sequence and the SR value of the mutant allele sequence. Thirteen $\boldsymbol{\Delta SR}$ values will be generated per binding sequence. The maximum score is determined from the set of $\boldsymbol{\Delta SR}$ values and denoted as $\boldsymbol{D_{max}}$.

$$Motif\ change\ score = \boldsymbol{\Delta SR} = \boldsymbol{SR_r} - \boldsymbol{SR_m}$$

$$D_{max} = Max(\boldsymbol{\Delta R}) = Max\ [\,\boldsymbol{\Delta R_1}, \boldsymbol{\Delta R_2}, \boldsymbol{\Delta R_3}, \boldsymbol{\Delta R_4}, \boldsymbol{\Delta R_5}, \boldsymbol{\Delta R_6}, \boldsymbol{\Delta R_7}, \boldsymbol{\Delta R_8}, \boldsymbol{\Delta R_9}, \boldsymbol{\Delta R_{10}}, \ldots\ldots \boldsymbol{\Delta R_{13}}]$$

$\boldsymbol{D_{max}}$ is the point with highest difference in the representation between the reference allele binding sequence and the mutant allele binding sequence of the SNP. This score is obtained for all SNPs in the three categories, TFBS-SNPs, REG-SNP and NON-REG-SNP.

C.4 Python Script

Simple scripts were written in python to compute the Markov order of dependency of nucleotides in regulatory sequences, and to compute Dmax scores. The following is a script applied to compute the Dmax of a zero order sequence.

```
#Import the following python functions:
from numpy import *
from scipy.stats import chi2
from Bio.Seq import Seq
from Bio import SeqIO
from decimal import Decimal
import re, itertools,  numpy as np, scipy as sp, random, math

#Create instances, strings and lists, for appending data and calculation:
duplet_list=[], triplet_list = [],raw_sequence='', , transition_freq = [] . . . . . . . . . etc

#define functions:
# function generating duplets to evaluate markov first order dependency
def dups(lst):
   for i in range(1, len(lst)):
      yield lst[i-1], lst[i]

#function generating triplets to evaluate markov second order dependency
def trips(lst):
   for i in range(1, len(lst)-1):
      yield lst[i-1], lst[i], lst[i+1]

#call regulatory sequences in fasta format from sequence files:
for seq_record in SeqIO.parse("TFBS-SEQ1SEPT2014.fasta", "fasta"):
```

```
sequence = seq_record.seq

sequence_list = list(sequence)

#print sequence_list

print seq_record.id

MAXI_ABSI_DIFF_list.append( seq_record.id) # appends SNP-ID to final result

SIGN_OF_list.append(seq_record.id) # appends SNP-ID to final result


#extract upflank of SNP [0-299 /R/M/ 305-605]

up_flank = sequence_list[0:300]

#print up_flank


#extract downflank of SNP [0-299 /R/M/ 305-605]

down_flank = sequence_list[305:605]

#print len(down_flank)


#Extract SNP Alleles from sequence

SNP_Position_REF = sequence_list[301:302]

#print SNP_Position_REF

SNP_Position_MUT = sequence_list[303:304]

#print SNP_Position_MUT


#join both flanks

join_flanks = up_flank + down_flank

#print join_flanks

#print len (join_flanks)


#Create reference allele sequence

ORI_sequence_REF = join_flanks[0:300] + SNP_Position_REF + join_flanks[300:600]

#print ORI_sequence_REF[298:303]

#print len(ORI_sequence_REF)


#Create mutant allele sequence
```

```
ORI_sequence_MUT = join_flanks[0:300] + SNP_Position_MUT + join_flanks[300:600]

#print ORI_sequence_MUT[298:303]

#print len(ORI_sequence_MUT)

sequence_tot = len(ORI_sequence_REF)


#Compute 0 order Markov scores

nucleotide_freq_ref = [ORI_sequence_REF.count('A'),ORI_sequence_REF.count('C'),

....ORI_sequence_REF.count('G'), ORI_sequence_REF.count('T')]

nucleotide_freq1_ref = [float(i) for i in nucleotide_freq_ref]

nucleotide_prob_ref = [i/float(sequence_tot) for i in nucleotide_freq1_ref]

nucleotide_prob_ref = [round(i,4) for i in nucleotide_prob_ref]


#Create 0 order Markov score dictionary

nucleotides = ['A', 'C', 'G', 'T']

markov0_dict = dict(zip(nucleotides, nucleotide_prob_ref))



#Generate overlapping trimers from regulatory sequence using a sliding window method

# SNP POSITION IS 300

# First triplet; [XXX]XXXX(XSX)XXXXXXX

SNP_SUR_MOTIF1 = ORI_sequence_REF[292:295]

#print SNP_SUR_MOTIF1

SNP_SUR_MOTIF1_obs = "".join(str(x) for x in SNP_SUR_MOTIF1)

#print SNP_SUR_MOTIF1_obs


# Second triplet; X[XXX]XXX(XSX)XXXXXXX

SNP_SUR_MOTIF2 = ORI_sequence_REF[293:296]

#print SNP_SUR_MOTIF2

SNP_SUR_MOTIF2_obs = "".join(str(x) for x in SNP_SUR_MOTIF2)

#print SNP_SUR_MOTIF2_obs

        .

        .
```

```
            .
            .
    # Twelveth triplet; XXXXXXX(XSX)XXX[XXX]X
    SNP_SUR_MOTIF14 = ORI_sequence_REF[305:308]
    #print SNP_SUR_MOTIF14
    SNP_SUR_MOTIF14_obs = "".join(str(x) for x in SNP_SUR_MOTIF14)
    #print SNP_SUR_MOTIF14_obs

    # Thirtheenth triplet; XXXXXXX(XSX)XXXX[XXX]
    SNP_SUR_MOTIF15 = ORI_sequence_REF[306:309]
    #print SNP_SUR_MOTIF15
    SNP_SUR_MOTIF15_obs = "".join(str(x) for x in SNP_SUR_MOTIF15)
    #print SNP_SUR_MOTIF15_obs


# Compute Expected probabilities of Trimers in the SNP environment (in this case, Markov Order = 1)
# First triplet
    prob_motif1 = markov0_dict[SNP_SUR_MOTIF1[0]] *
    markov0_dict[SNP_SUR_MOTIF1[1]] * markov0_dict[SNP_SUR_MOTIF1[2]]
    obs_win_1 = markov2_obs_dict[SNP_SUR_MOTIF1_obs]

   # append computed scores to a list
    expected_probs_ref_coll.append(prob_motif1)
    observed_probs_ref_coll.append(obs_win_1)

# Second triplet
    prob_motif2 = markov0_dict[SNP_SUR_MOTIF2[0]] *
    markov0_dict[SNP_SUR_MOTIF2[1]] * markov0_dict[SNP_SUR_MOTIF2[2]]
    obs_win_2 = markov2_obs_dict[SNP_SUR_MOTIF2_obs]
    expected_probs_ref_coll.append(prob_motif2)
    observed_probs_ref_coll.append(obs_win_2)
            .
```

```
          .

          .

# Twelveth triplet

  prob_motif14 = markov0_dict[SNP_SUR_MOTIF14[0]] *

  markov0_dict[SNP_SUR_MOTIF14[1]] * markov0_dict[SNP_SUR_MOTIF14[2]]

  obs_win_14 = markov2_obs_dict[SNP_SUR_MOTIF14_obs]

  expected_probs_ref_coll.append(prob_motif14)

  observed_probs_ref_coll.append(obs_win_14)


# Thirteenh triplet

  prob_motif15 = markov0_dict[SNP_SUR_MOTIF15[0]] *

  markov0_dict[SNP_SUR_MOTIF15[1]] * markov0_dict[SNP_SUR_MOTIF15[2]]

  obs_win_15 = markov2_obs_dict[SNP_SUR_MOTIF15_obs]

  expected_probs_ref_coll.append(prob_motif15)

  observed_probs_ref_coll.append(obs_win_15)



  ##compute SR = (O - E)/ Sqrt (E) ...convert probs to frequencies *599

  expected_probs_ref_coll_1 = array(expected_probs_ref_coll)


  observed_probs_ref_coll_1 = array(observed_probs_ref_coll)

  O_subtract_E_ref = observed_probs_ref_coll_1 - expected_probs_ref_coll_1

  expected_probs_ref_coll_1_sqrt = np.sqrt(expected_probs_ref_coll_1)

  O_subtract_E_over_sqrtE_ref = O_subtract_E_ref/expected_probs_ref_coll_1_sqrt


#Repeat same as above for Mutant allele Sequence


#Compute Dmax (MAX (ABS (REF-MUT)))

    REF_minus_MUT = O_subtract_E_over_sqrtE_ref - O_subtract_E_over_sqrtE_mut

    Absi_REF_minus_MUT  = abs(REF_minus_MUT)

    MAXI_ABSI_DIFF = max(Absi_REF_minus_MUT) ### value of Dmax

    SIGN_OF = MAXI_DIFF == MAXI_ABSI_DIFF #### to know original sign of Dmax
```

#reshuffle regulatory sequence 5000 times to generate random Dmax values in order to determine the significance of the original Dmax.

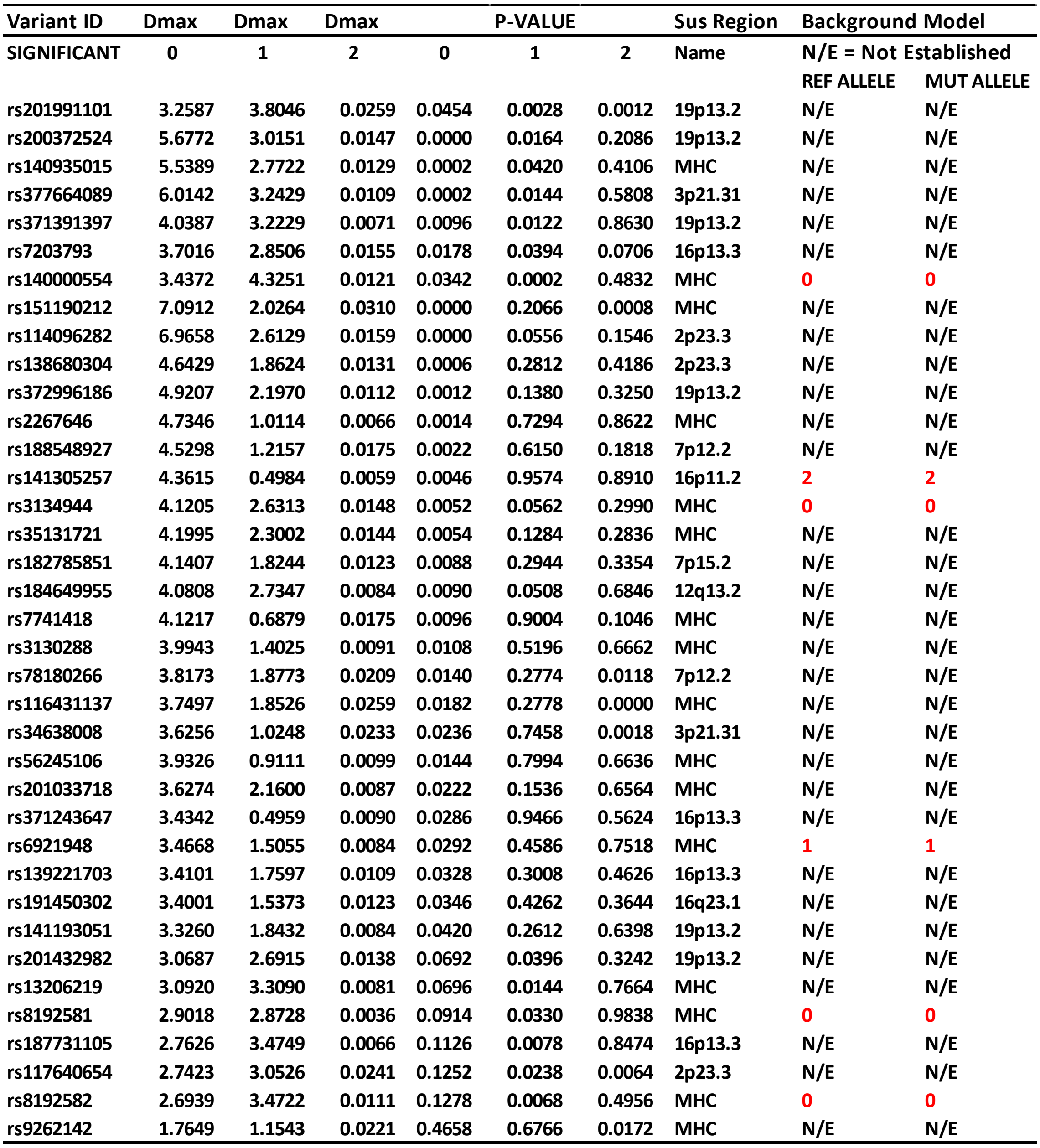
```
 count = 0

 while (count<5000):

 random.shuffle(join_flanks)

#print join_flanks[298:300],join_flanks[300:302]

#mixed_sequence = join_flanks[0:300] + join_flanks[300:600]

#print mixed_sequence

#print len(mixed_sequence)


mixed_sequence_REF = join_flanks[0:300] + SNP_Position_REF + join_flanks[300:600]

#print mixed_sequence_REF[298:303]

#print len(mixed_sequence_REF)


mixed_sequence_MUT = join_flanks[0:300] + SNP_Position_MUT +

join_flanks[300:600]

#print mixed_sequence_MUT[298:303]

#print len(mixed_sequence_MUT)


sequence_tot = len(mixed_sequence_REF)

#print sequence_tot


 count = count + 1
```

#Repeat entire process for new reshuffled sequence, compute O, E, SR and Dmax.

#Refresh lists by deleting data of previous sequence, then call new sequence.

C.5 Statistics for Significant TFBS SNPs

Table 12. D_{max} and *P* values for the TFBS-SNPs that test positive for SNP sensitivity



Variant ID	Dmax	Dmax	Dmax		P-VALUE		Sus Region	Background Model	
SIGNIFICANT	0	1	2	0	1	2	Name	N/E = Not Established	
								REF ALLELE	MUT ALLELE
rs201991101	3.2587	3.8046	0.0259	0.0454	0.0028	0.0012	19p13.2	N/E	N/E
rs200372524	5.6772	3.0151	0.0147	0.0000	0.0164	0.2086	19p13.2	N/E	N/E
rs140935015	5.5389	2.7722	0.0129	0.0002	0.0420	0.4106	MHC	N/E	N/E
rs377664089	6.0142	3.2429	0.0109	0.0002	0.0144	0.5808	3p21.31	N/E	N/E
rs371391397	4.0387	3.2229	0.0071	0.0096	0.0122	0.8630	19p13.2	N/E	N/E
rs7203793	3.7016	2.8506	0.0155	0.0178	0.0394	0.0706	16p13.3	N/E	N/E
rs140000554	3.4372	4.3251	0.0121	0.0342	0.0002	0.4832	MHC	0	0
rs151190212	7.0912	2.0264	0.0310	0.0000	0.2066	0.0008	MHC	N/E	N/E
rs114096282	6.9658	2.6129	0.0159	0.0000	0.0556	0.1546	2p23.3	N/E	N/E
rs138680304	4.6429	1.8624	0.0131	0.0006	0.2812	0.4186	2p23.3	N/E	N/E
rs372996186	4.9207	2.1970	0.0112	0.0012	0.1380	0.3250	19p13.2	N/E	N/E
rs2267646	4.7346	1.0114	0.0066	0.0014	0.7294	0.8622	MHC	N/E	N/E
rs188548927	4.5298	1.2157	0.0175	0.0022	0.6150	0.1818	7p12.2	N/E	N/E
rs141305257	4.3615	0.4984	0.0059	0.0046	0.9574	0.8910	16p11.2	2	2
rs3134944	4.1205	2.6313	0.0148	0.0052	0.0562	0.2990	MHC	0	0
rs35131721	4.1995	2.3002	0.0144	0.0054	0.1284	0.2836	MHC	N/E	N/E
rs182785851	4.1407	1.8244	0.0123	0.0088	0.2944	0.3354	7p15.2	N/E	N/E
rs184649955	4.0808	2.7347	0.0084	0.0090	0.0508	0.6846	12q13.2	N/E	N/E
rs7741418	4.1217	0.6879	0.0175	0.0096	0.9004	0.1046	MHC	N/E	N/E
rs3130288	3.9943	1.4025	0.0091	0.0108	0.5196	0.6662	MHC	N/E	N/E
rs78180266	3.8173	1.8773	0.0209	0.0140	0.2774	0.0118	7p12.2	N/E	N/E
rs116431137	3.7497	1.8526	0.0259	0.0182	0.2778	0.0000	MHC	N/E	N/E
rs34638008	3.6256	1.0248	0.0233	0.0236	0.7458	0.0018	3p21.31	N/E	N/E
rs56245106	3.9326	0.9111	0.0099	0.0144	0.7994	0.6636	MHC	N/E	N/E
rs201033718	3.6274	2.1600	0.0087	0.0222	0.1536	0.6564	MHC	N/E	N/E
rs371243647	3.4342	0.4959	0.0090	0.0286	0.9466	0.5624	16p13.3	N/E	N/E
rs6921948	3.4668	1.5055	0.0084	0.0292	0.4586	0.7518	MHC	1	1
rs139221703	3.4101	1.7597	0.0109	0.0328	0.3008	0.4626	16p13.3	N/E	N/E
rs191450302	3.4001	1.5373	0.0123	0.0346	0.4262	0.3644	16q23.1	N/E	N/E
rs141193051	3.3260	1.8432	0.0084	0.0420	0.2612	0.6398	19p13.2	N/E	N/E
rs201432982	3.0687	2.6915	0.0138	0.0692	0.0396	0.3242	19p13.2	N/E	N/E
rs13206219	3.0920	3.3090	0.0081	0.0696	0.0144	0.7664	MHC	N/E	N/E
rs8192581	2.9018	2.8728	0.0036	0.0914	0.0330	0.9838	MHC	0	0
rs187731105	2.7626	3.4749	0.0066	0.1126	0.0078	0.8474	16p13.3	N/E	N/E
rs117640654	2.7423	3.0526	0.0241	0.1252	0.0238	0.0064	2p23.3	N/E	N/E
rs8192582	2.6939	3.4722	0.0111	0.1278	0.0068	0.4956	MHC	0	0
rs9262142	1.7649	1.1543	0.0221	0.4658	0.6766	0.0172	MHC	N/E	N/E

C.6 Statistics for Significant TFBS SNPs

Table 13. D_{max} and *P* values for the TFBS-SNPs that test negative for SNP sensitivity

Variant ID	Dmax	Dmax	Dmax		P-VALUE		Sus Region	Background Model	
NON-SIGNIFICANT	0	1	2	0	1	2	Name	N/E = Not Established	
								REF ALLELE	MUT ALLELE
rs374210880	3.240573	2.034279	0.010892	0.051	0.2004	0.5398	2p23.3	N/E	N/E
rs78370725	3.160715	1.091354	0.012047	0.0564	0.6782	0.3532	19p13.2	N/E	N/E
rs199581527	3.027160	1.603788	0.014405	0.0776	0.3894	0.2464	12q13.2	N/E	N/E
rs139490960	2.803060	1.273316	0.003930	0.1138	0.5864	0.9802	2p23.3	N/E	N/E
rs114760565	2.676032	0.375281	0.011650	0.126	0.9808	0.514	19p13.2	N/E	N/E
rs181119155	2.356022	2.577291	0.012286	0.2174	0.071	0.4286	16p13.3	N/E	N/E
rs204997	3.176675	0.908030	0.013487	0.0572	0.7946	0.239	MHC	1	1
rs199672847	2.933521	2.353332	0.015793	0.0874	0.1008	0.2212	MHC	0	0
rs113123395	2.901693	1.207738	0.008911	0.0896	0.6278	0.736	MHC	0	0
rs59564381	2.846325	2.362478	0.012286	0.0988	0.114	0.3508	MHC	0	0
rs147592187	2.859507	1.382912	0.011476	0.1	0.5258	0.5146	7p12.2	N/E	N/E
rs148068088	2.824492	2.237404	0.007868	0.1098	0.1376	0.818	MHC	1	1
rs113977555	2.746047	0.702883	0.012286	0.1192	0.8942	0.3112	MHC	N/E	N/E
rs9469383	2.702570	1.527362	0.009876	0.1338	0.445	0.6306	MHC	N/E	N/E
rs73728831	2.543302	1.271656	0.009876	0.1598	0.5924	0.684	MHC	N/E	N/E
rs141920214	2.525023	1.989088	0.007220	0.172	0.228	0.8014	MHC	0	0
rs12194528	2.320206	1.031810	0.012286	0.2354	0.7296	0.3962	MHC	2	2
rs148149314	2.271842	1.912886	0.007121	0.2386	0.2484	0.8572	16p13.3	N/E	N/E
rs183881418	2.297608	0.926283	0.012185	0.245	0.791	0.4006	19p13.2	N/E	N/E
rs373832002	2.260078	0.944472	0.005302	0.2524	0.7518	0.9056	16p11.2	N/E	N/E
rs112027660	2.210722	0.745931	0.010981	0.2638	0.8802	0.5322	3p21.31	N/E	N/E
rs149723334	2.189628	0.723222	0.012291	0.2718	0.8806	0.5062	7p12.2	N/E	N/E
rs794427	2.157357	1.094423	0.012286	0.2904	0.699	0.4722	16p13.3	N/E	N/E
rs202169452	2.114167	1.252146	0.011387	0.2972	0.6124	0.4252	MHC	1	1
rs75810024	2.112495	1.463208	0.012393	0.2976	0.487	0.355	16p13.13	N/E	N/E
rs2735072	2.112382	1.850149	0.008361	0.3046	0.2812	0.7378	MHC	N/E	N/E
rs149780751	2.058868	2.005689	0.012393	0.3246	0.213	0.408	MHC	N/E	N/E
rs375601741	2.024241	1.067739	0.012968	0.3436	0.7152	0.3484	12q13.2	N/E	N/E
rs368672104	1.966531	1.281681	0.015470	0.357	0.5828	0.1888	16q23.1	N/E	N/E
rs111297363	1.968790	0.563793	0.014405	0.3634	0.945	0.1482	MHC	0	0
rs200223154	1.966035	0.319814	0.008552	0.3694	0.9886	0.72	MHC	2	2
rs150428668	1.936227	0.962457	0.012968	0.384	0.781	0.3294	3p21.31	N/E	N/E
rs11833282	1.926165	2.124640	0.007220	0.3858	0.1686	0.8274	12q13.2	N/E	N/E
rs116908088	1.882544	1.475385	0.012968	0.3986	0.4714	0.4712	16q23.1	N/E	N/E
rs77744705	1.833181	1.726470	0.006201	0.4264	0.3444	0.92	16p13.13	N/E	N/E
rs150341510	1.809561	2.100026	0.017093	0.4376	0.173	0.1128	16p13.3	N/E	N/E
rs11575516	1.616444	1.575862	0.015178	0.551	0.4166	0.1428	7p12.2	0	0
rs188878585	1.587201	2.025235	0.007913	0.5522	0.2156	0.7708	MHC	N/E	N/E
rs150127869	1.486201	2.341205	0.006044	0.6064	0.1226	0.9064	2p23.3	N/E	N/E
rs137926274	1.457074	1.097976	0.014405	0.6122	0.7052	0.275	MHC	0	0
rs2735073	1.397497	1.753492	0.009732	0.623	0.311	0.623	MHC	N/E	N/E
rs188429583	1.348052	0.495543	0.004180	0.6552	0.9518	0.9712	7p15.2	N/E	N/E
rs28382772	1.335038	0.545949	0.006201	0.6776	0.951	0.8876	19p13.2	0	0
rs371334332	1.180137	1.432280	0.011558	0.7428	0.4678	0.4256	16p11.2	N/E	N/E
rs117234201	1.170157	0.871440	0.012286	0.7702	0.822	0.4316	16q23.1	N/E	N/E
rs117071466	1.089586	0.921046	0.019169	0.792	0.7792	0.1052	16p11.2	N/E	N/E
rs201503590	1.028949	0.384340	0.008722	0.8146	0.9832	0.7496	MHC	N/E	N/E
rs148180043	0.791948	1.800470	0.006817	0.9186	0.3082	0.8754	12q13.2	N/E	N/E
rs12828657	0.732909	0.922325	0.006004	0.928	0.794	0.9206	12q13.2	N/E	N/E
rs183163194	0.686688	0.453070	0.008402	0.9376	0.9732	0.7074	MHC	N/E	N/E
rs4781062	0.568535	0.552849	0.014405	0.9634	0.9452	0.234	16p13.13	1	1
rs28382773	0.515414	1.725389	0.004488	0.969	0.3278	0.9678	19p13.2	0	0
rs7567804	0.507978	0.549044	0.013589	0.9738	0.9468	0.3142	2q11.2	N/E	N/E
rs185919902	0.420633	0.972615	0.004908	0.9842	0.7514	0.9478	19p13.2	N/E	N/E
rs190388624	0.401348	0.953902	0.009322	0.9874	0.7622	0.6532	16p11.2	N/E	N/E

C.7. Binding motifs in which the significant TFBS-SNPs occur

Table 14. Binding motifs in which the significant TFBS-SNPs occur. The table shows the name of the motif, indicates the SNP position in the motif, and if it is in a high information position. A motif change score obtained from ensembl is also shown.

SNP ID	Motif Family & Motif ID	High information position	SNP position in motif	Motif change score (Ensembl)
rs138680304	USF1:MA0093.2	No	5	Less like consequence
	USF1:MA0281.1	Yes	4	Less like consequence
rs78180266	Max:MA0058.2	Yes	6	Less like consequence
rs141305257	USF1:MA0281.1	Yes	4	Less like consequence
	USF1:MA0093.2	No	5	Less like consequence
rs200372524	USF1:MA0281.1	Yes	7	Less like consequence
rs114096282	Egr1:MA0162.2	Yes	11	Less like consequence
	SP1:MA0079.3	No	11	Less like consequence
rs117640654	USF1:MA0281.1	Yes	7	Less like consequence
rs377664089	Tcf12:MA0521.1	Yes	8	Less like consequence
rs34638008	USF1:MA0281.1	Yes	4	Less like consequence
rs188548927	Egr1:MA0337.1	Yes	2	Less like consequence
rs182785851	Egr1:MA0341.1	Yes	3	Less like consequence
	Egr1:MA0366.1	Yes	3	Less like consequence
rs184649955	Egr1:MA0162.2	No	4	Less like consequence
rs7203793	USF1:MA0093.2	Yes	3	Less like consequence
rs371243647	SP1:MA0079.3	No	2	Less like consequence
rs139221703	Znf263:MA0528.1	No	5	Less like consequence
rs187731105	Znf263:MA0528.1	No	13	Less like consequence
rs191450302	SP1:MA0079.3	Yes	4	Less like consequence
	Egr1:MA0162.2	No	4	Less like consequence
	Egr1:MA0337.1	Yes	4	Less like consequence
rs201991101	USF1:MA0093.2	No	5	Less like consequence
rs371391397	E2F4:MA0470.1	Yes	4	Less like consequence
rs372996186	Egr1:MA0341.1	Yes	3	Less like consequence
	Egr1:MA0366.1	Yes	3	Less like consequence
rs201432982	E2F4:MA0470.1	Yes	4	Less like consequence
rs141193051	E2F4:MA0470.1	No	6	Less like consequence
rs140935015	FOSL2:MA0478.1	Yes	4	Less like consequence
	Jund:MA0491.1	Yes	3	Less like consequence
rs140000554	USF1:MA0093.2	No	2	Less like consequence
	USF1:MA0093.2	No	11	Less like consequence
	USF1:MA0281.1	No	1	No change
rs151190212	EBF1:MA0154.2	Yes	3	Less like consequence
rs2267646	Egr1:MA0366.1	Yes	3	Less like consequence
	Egr1:MA0341.1	Yes	3	Less like consequence
rs3134944	USF1:MA0093.2	No	5	Less like consequence
	USF1:MA0281.1	Yes	4	Less like consequence
rs35131721	EBF1:MA0154.2	Yes	5	Less like consequence
rs7741418	USF1:MA0093.2	No	5	Less like consequence
	USF1:MA0281.1	Yes	4	Less like consequence
rs3130288	SP1:MA0079.3	No	11	Less like consequence
rs116431137	Jund:MA0491.1	Yes	9	Less like consequence
rs56245106	Jund:MA0491.1	Yes	3	Less like consequence
rs201033718	Egr1:MA0341.1	Yes	5	Less like consequence
	Egr1:MA0366.1	No	5	Less like consequence
rs6921948	FOXA1:MA0546.1	Yes	8	Less like consequence
rs9262142	Jund:MA0491.1	Yes	4	Less like consequence
	FOSL1:MA0477.1	Yes	4	Less like consequence
rs8192582	Egr1:MA0162.2	Yes	11	Less like consequence
	SP1:MA0079.3	No	11	Less like consequence
rs8192581	Egr1:MA0162.2	No	9	Less like consequence
	SP1:MA0079.3	No	9	Less like consequence
rs13206219	Jund:MA0491.1	Yes	4	Less like consequence

C.8. Binding motifs in which the non-significant TFBS-SNPs occur

Table 15. Binding motifs in which the non-significant TFBS-SNPs occur. The table shows the name of the motif, indicates the SNP position in the motif, and if it is in a high information position.

SNP ID	Motif Family & Motif ID	High information position	SNP position in motif	Motif change score (Ensembl)
rs112027660	Tcf12:MA0521.1	Yes	4	Less like consequence
rs114760565	ELF1:MA0473.1	Yes	11	Less like consequence
rs11575516	USF1:MA0281.1	Yes	5	Less like consequence
rs116908088	Nrf1:MA0506.1	Yes	3	Less like consequence
rs117071466	CTCFL:MA0531.1	No	4	More like consequence
rs117234201	Nrf1:MA0506.1	No	6	Less like consequence
rs11833282	Egr1:MA0162.2	No	9	Less like consequence
rs12828657	ZBTB33:MA0527.1	No	15	Less like consequence
rs137926274	Yy1:MA0095.2	Yes	8	Less like consequence
rs139490960	USF1:MA0281.1	Yes	4	Less like consequence
rs147592187	Jund:MA0491.1	Yes	3	Less like consequence
rs148068088	Egr1:MA0341.1	No	1	Less like consequence
rs148068088	Egr1:MA0366.1	Yes	1	Less like consequence
rs148149314	Egr1:MA0341.1	Yes	4	Less like consequence
rs148149314	Egr1:MA0366.1	Yes	4	Less like consequence
rs148180043	Yy1:MA0095.2	No	10	Less like consequence
rs149723334	Egr1:MA0423.1	Yes	2	Less like consequence
rs150127869	FOSL1:MA0477.1	No	11	Less like consequence
rs150127869	Jund:MA0491.1	No	11	More like consequence
rs150341510	FOSL1:MA0477.1	Yes	7	Less like consequence
rs150341510	Jund:MA0491.1	Yes	7	Less like consequence
rs150428668	MEF2A:MA0585.1	Yes	6	Less like consequence
rs181119155	MEF2A:MA0052.2	No	12	Less like consequence
rs183163194	SP1:MA0079.3	No	3	Less like consequence
rs183881418	Egr1:MA0341.1	Yes	2	Less like consequence
rs183881418	Egr1:MA0366.1	Yes	2	Less like consequence
rs185919902	Tcf12:MA0521.1	Yes	8	More like consequence
rs188429583	USF1:MA0093.2	No	1	More like consequence
rs188878585	Egr1:MA0341.1	Yes	5	Less like consequence
rs188878585	Egr1:MA0366.1	No	5	Less like consequence
rs190388624	Egr1:MA0162.2	No	2	Less like consequence
rs199581527	E2F4:MA0541.1	No	3	More like consequence
rs201503590	Egr1:MA0162.2	No	4	Less like consequence
rs2735073	Egr1:MA0162.2	No	10	Less like consequence
rs28382772	E2F4:MA0470.1	No	10	More like consequence
rs28382772	E2F4:MA0541.1	No	13	Less like consequence
rs28382773	SP1:MA0079.3	No	6	Less like consequence
rs368672104	Egr1:MA0162.2	No	12	Less like consequence
rs371334332	CTCFL:MA0531.1	Yes	7	Less like consequence
rs373832002	ZEB1:MA0103.2	No	2	Less like consequence
rs374210880	Pax5:MA0014.2	No	8	Less like consequence
rs375601741	ZBTB33:MA0527.1	Yes	5	Less like consequence
rs4781062	HNF4A:MA0114.2	No	12	Less like consequence
rs7567804	USF1:MA0281.1	Yes	3	Less like consequence
rs75810024	USF1:MA0093.2	No	6	Less like consequence
rs75810024	USF1:MA0281.1	Yes	5	Less like consequence
rs77744705	Jund:MA0491.1	No	6	More like consequence
rs78370725	Egr1:MA0366.1	No	5	Less like consequence
rs78370725	Egr1:MA0341.1	Yes	5	Less like consequence
rs794427	Jund:MA0491.1	No	11	More like consequence
rs112027660	Tcf12:MA0521.1	Yes	4	Less like consequence

C.8. Binding motifs in which the non-significant TFBS-SNPs occur contd

Table 15. Binding motifs in which the non-significant TFBS-SNPs occur. The table shows the name of the motif, indicates the SNP position in the motif, and if it is in a high information position.

rs204997	SP2:MA0516.1	Yes	7	Less like consequence
rs204997	SP1:MA0079.3	Yes	7	Less like consequence
rs199672847	Egr1:MA0162.2	No	13	Less like consequence
rs113123395	USF1:MA0281.1	Yes	7	Less like consequence
rs113123395	USF1:MA0093.2	Yes	8	Less like consequence
rs59564381	ELF1:MA0473.1	No	2	Less like consequence
rs113977555	Srf:MA0083.2	No	12	Less like consequence
rs9469383	CTCFL:MA0531.1	No	12	Less like consequence
rs73728831	Srf:MA0083.2	Yes	7	Less like consequence
rs141920214	ELF1:MA0473.1	No	1	More like consequence
rs12194528	SP1:MA0079.3	Yes	7	Less like consequence
rs12194528	Egr1:MA0162.2	No	13	Less like consequence
rs202169452	SP1:MA0079.3	No	11	Less like consequence
rs2735072	Egr1:MA0162.2	No	12	Less like consequence
rs149780751	CTCFL:MA0531.1	Yes	13	Less like consequence
rs111297363	USF1:MA0281.1	Yes	6	Less like consequence
rs111297363	USF1:MA0093.2	Yes	7	Less like consequence
rs200223154	SP1:MA0079.3	No	11	Less like consequence
rs200223154	Egr1:MA0337.1	Yes	3	Less like consequence

C.9. Facts about genes that are in proximity with some of the Significant TFBS-SNPs.

SNP that occurs in an experimentally detected binding site, and is closely linked with a disease associated SNP, is more likely to play a biological role in the genome than other SNPs that occur in parts for which there is no particular known function (Schuab et al., 2012). Through this work, it is found that though the associated T1D-SNPs are not regulatory SNPs that may influence transcription factor binding, there are other nearby non-associated SNPs that can influence this process. Thirty-seven of these rare regulatory TFBS-SNPs have been identified by their testing positive for SNP sensitivity. In addition to significantly changing the representation of their local environment, they are outstandingly closer in proximity to the disease-associated SNPs than the other TFBS-SNPs.

The significant TFBS-SNPs are mostly characterised by C-T transitions, which have previously been shown to cause weaker affinity for transcription factor (TF) binding. Also, they influence 31 different binding sites for 18 transcription factor families. The binding sites for the USF family of transcription factors are the most affected. These proteins, USF1 and USF2, have been linked to genetic problems involving insulin genes and regulation of glucose. These problems are similar to the features that characterise T1D, where insulin is primary auto-antigen[24]. Despite these important findings, further testing in a biological system is necessary to determine whether these SNPs do affect function in vivo. Experimentation can reveal if the recognition and binding of TFs to the affected sites is altered, and how this in turn disturbs the transcription of target genes.

A non-coding SNP that is located at a transcription factor-binding site (TFBS) of a gene could affect the level or timing of gene expression (Xu and Taylor, 2009).

An Ensembl genome browser search reveals that the 37 significant TFBS-SNPs are in the vicinity of about 60 protein-coding and non-coding genes. They occur in flaking regions as well as within gene sequences. A simple identification of the functional annotation of these genes was retrieved from the DAVID (Database for Annotation, Visualization and Integrated Discovery) bioinformatics resource. The result indicates that many of these genes are involved in positive and negative regulation of various biological processes. A good example relates to 4 SNPS, which are adjacent to genes that are involved in positive regulation of T-cell differentiation and activation. A T-cell is a type of lymphocyte[25] that plays a role in cell mediated immunity. This type of immunity involves the activation of cells, called phagocytes, which protect the body by

[24] an antigen that despite being a normal tissue constituent of the body is the target of a humoral or cell-mediated immune response, it stimulates the production of autoantibodies and an autoimmune attack as in autoimmune diseases

[25] A lymphocyte is a sub-type of a white blood cell which are cells of the immune system that are involved in protecting the body against both foreign particles and infectious disease

ingesting harmful foreign particles like bacteria. Auto-reactive T-cells (T-cells produced by an organism and acting against its own cells or tissues) play a major role in the pathogenesis of type 1 diabetes mellitus (Monti et al., 2009). T-cell response against the important insulin-producing beta cells is a main characteristic of T1D, but this process is not yet fully understood. T-cells are a major target of immunomodulatory[26] approaches that are aimed at delaying or even preventing the disease onset (Bluestone and Buor-Jordan, 2012; Monti et al., 2009).

The significant TFBS-SNP rs2267646, in the HLA region, is upstream of the HLA-DMA gene. It is a heterodimeric molecule important for normal antigen[27] presentation (Sanderson et al., 1994). Cells involved in antigen presentation, take up and process antigens into such a form that when displayed at the cell surface is recognized by T cells, and activates an immune response. Rs56245106 and rs13206219 also in the HLA (haplotype: HSCHR6_MHC_COX) are overlapped by intronic parts of transcripts of the HLA-DRB1 gene. This protein is also involved in antigen presentation for recognition by the CD4 T-cells also called mature T-cells (Ayyoub et al., 2004; Janeway et al., 2001). CD4 T-cells are important in orchestrating overall immune responses, they play an important role in modulating immune responses to pathogens (an infectious agent or anything that can produce disease), as well as tumour cells (Macleod et al., 2010). Rs188548927 in region 7p12.2 occurs in intronic regions of 12 coding transcripts of the IKZF1 gene. The protein is a transcriptional regulator of hematopoietic cell differentiation. Hematopoietic stem cells are the blood cells that give rise to all the other blood cells. These include lymphocytes, a type of which is the T-cell. This SNP occurs about 10,000 bps away from an associated T1D-SNP (rs10272724) which is in the 3' UTR of the same gene. This disease-associated SNP is linked with susceptibility to childhood acute lymphoblastic leukaemia, a rare cancer of T-cells (Swafford et al., 2011; Papaemmanuil et al., 2009), in addition to T1D.

Problems in the genes highlighted in the foregoing examples could likely contribute to the aberrant autoimmune reaction that occurs in T1D. In fact, they have already been suggested to be candidate causal/susceptibility genes in the aetiology of T1D (Gillespie, 2014; Noble and Erlich, 2012; Todd, 2011). However, the significant TFBS-SNPs are also adjacent to many other genes that are linked to other conditions and that could possibly be associated with T1D. For instance, rs140000554 is adjacent of the RAGE/AGER gene, it occurs downstream of 14 alternative transcripts of this gene which encodes a pattern recognition receptor. The molecule, is a member of the immunoglobulin superfamily, and is expressed on the surface of different cell types including lymphocytes (Mahajan et al., 2013). It has five binding domains that detect and bind glycoprotein ligands[28] (Neeper et al., 1992), which mediate interaction between white blood

[26] modification of the immune response or the functioning of the immune system, for example, by the inhibition of white blood cell activity or the stimulation of antibody formation.

[27] Simply put, an antigen is any substance that causes one's immune system to produce antibodies against it. An antigen could be a foreign substance from the environment, like, bacteria, viruses, chemicals or pollen. An antigen could also be formed inside the body, as with bacterial toxins or tissue cells

[28] Glycoproteins are proteins that contain oligosaccharide chains (glycans/sugar) covalently attached to the polypeptide side-chains. They are important for white blood cell recognition

cells and inflamed endothelial cells that line the interior surface of blood and lymphatic vessels (Borges, et al., 1997). This interaction leads to a movement of leukocytes towards the site of infection or tissue damage commonly referred to as leukocyte extravasation (Vestweber, 1997). This innate immune response[29] is believed to result in activation of pro-inflammatory genes (Bierhaus et al., 2001). However, this process has been linked to certain chronic inflammatory disorders (Mahajan et al., 2013). The RAGE gene is suspected to have an effect in inflammatory diseases including diabetic complications (e.g. diabetic nephropathy or retinopathy) (Singh et al, 2014; Bierhaus et al., 2001; Hudson et al., 2001), because there is an enhanced level of RAGE ligands in diabetes (Hudson, 2002).

A final example is related to the aforementioned rs2267646. It is also intronic of a second gene, BRD2, which is a transcription factor protein involved in the regulation of another gene, CCND1. The CCND1 protein belongs to the Cyclin family of proteins that are involved in cell cycle progression. Overexpression of this gene can alter cell cycle progression in such a way that contributes to tumorigenesis (the formation of a tumour). CCND1 is frequently observed in a variety of tumours, for instance, in ovarian cancer (Zhang et al., 2014), in the brain (Qin et al., 2014) and in gastric cancer (Kuo et al., 2014).

These random examples highlight the significance of the non-associated regulatory SNPs identified in this work. They may have an influence on the regulation of any of these important nearby genes, as well as other genes that are farther away through gene regulatory networks. These instances also add to the finding that a single SNP can indeed affect more than one process in the genome. By occurring in a binding site and as well as within multiple gene transcripts, the effects of some of these SNPs may not be limited to just one process.

[29] Innate immune systems provide immediate defence against infection. In vertebrates some of the major functions of innate immunity include the identification and removal of foreign substances present in organs, tissues, the blood and lymph, by specialised white blood cells, and activation of the adaptive immune system through antigen presentation.

Appendix D

Ravendbase

Ravendbase was created in my first year of research to bring together data collected from three online genomic databases. The first, T1Dbase, is a T1D dedicated resource containing information about T1D associated regions in the human and mouse genome. The second, Ensembl, stores general genomic data for several species; and the third, DbSNP, stores SNP information made available from biological research and GWAS. Ravendbase was originally created using Microsoft's Access, and contains information collected for T1D susceptibility regions. Nowadays, it is good practice to store research data in a readily available format for view and use. Large amounts of biological data are continuously being generated by advanced biological techniques, even as old information is also being updated. The amount of data available for biological research is now quite large and can become overwhelming and confusing for a researcher if not properly stored in a format that will allow for easy data access, retrieval as well as data querying. Collecting data for my study required having to traverse between the above mentioned resources for information needed. To circumvent having to do this continuously throughout my research, any data retrieved from the larger online databases was formatted and stored in a desired set-up as 'Ravendbase'.

Having this database helped to have relevant data quickly accessible for studies, and made retrieving T1D region and T1D-SNP information easy and straightforward. It also allowed speedy data download for statistical analysis. Ravendbase is now being enhanced for sharing with other researchers or professionals who might be interested in information about T1D-SNPS.

D.1. Reason for database enhancement

The reason for further development of Ravendbase is to make available a simple user friendly resource that gives general information about: (1) T1D susceptibility regions, (2) SNPS in T1D susceptibility regions, (3) Genes and Transcripts in T1D susceptibility regions and (4) Genic positions of SNPS in transcripts. The database is simple, sort of like a reference for T1D SNPs, eliminating the initial confusion first time researchers get when accessing- and having to navigate around the -much larger and highly developed biological databases. Users can find basic information needed and then carry out a more focused search in larger advanced biological resources. This work was done as a BSc (final year project) in computer science at the University of Hertfordshire, which I co-supervised, by Nathan Beka.

D.2. Conversion from MS Access to MySQL

Ravendbase was initially created with Microsoft's Access. This is a popular data management application that allows one to store information in tables in a database format, which it manages directly from the local disk of the computer. Access also has a front end user interface to information in the database. However, this application is constrained in that it can only manage a limited amount of data and is generally used as a personal or single user application. MySQL is an open source relational database management system (RDBMS). It runs as a server providing multi-user access to a database, thereby opening up more possibilities for a database. It is able to manage hundreds of megabytes of data which can become a problem for Access, and is also able to handle many simultaneous users. A MySQL version of the database was therefore created, converting it from a single user system to a multiuser storage management system allowing for further improvement. The database structure is illustrated in the following section showing a data flow chart and Entity Relationship model.

D.3. Database Structure

The data flow chart for Ravendbase is shown in Figure 6. Information in the database is stored in a structure based on ten linked tables as presented in the entity relationship model in Figure 7.

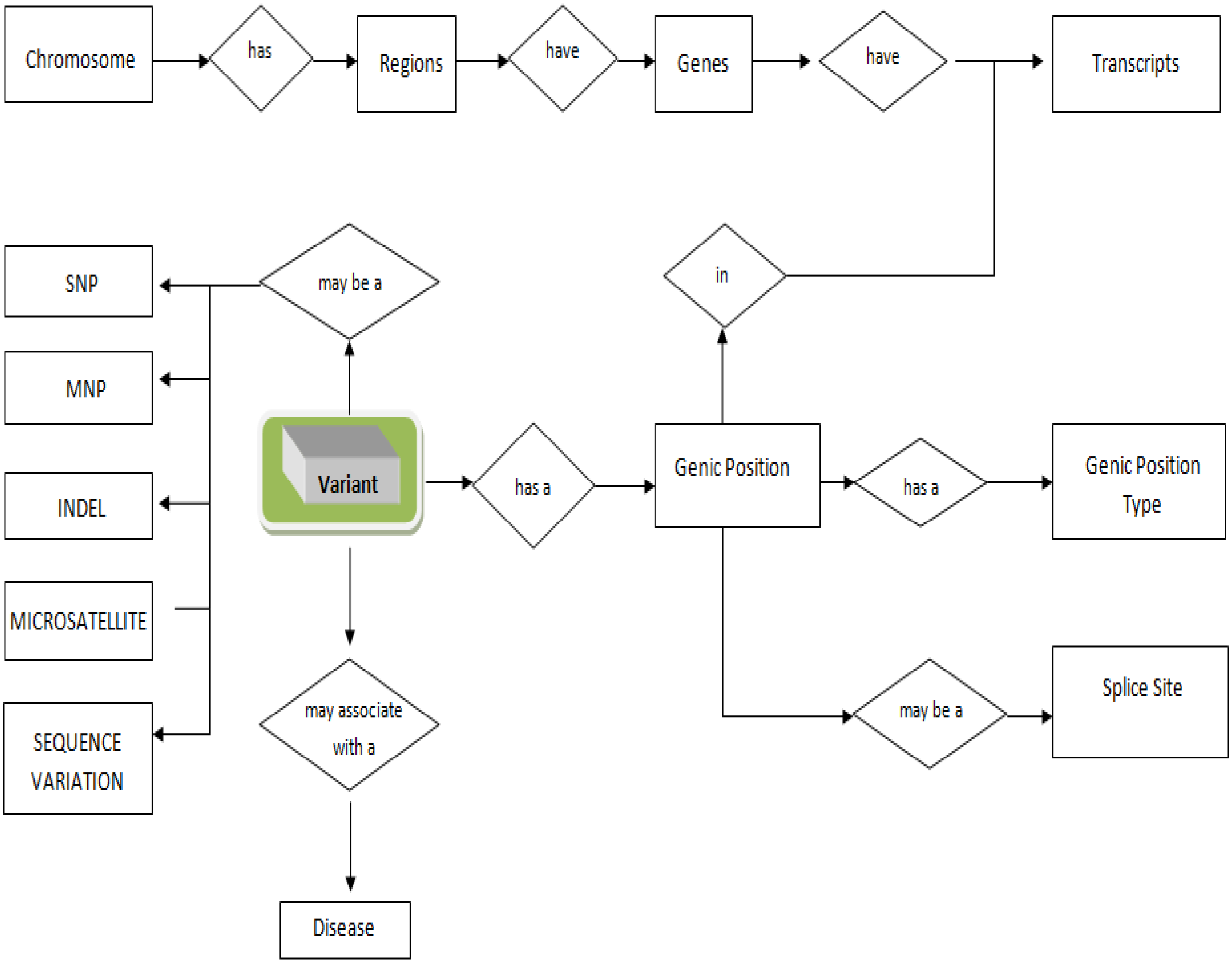


Figure 6. Ravendbase data flow chart

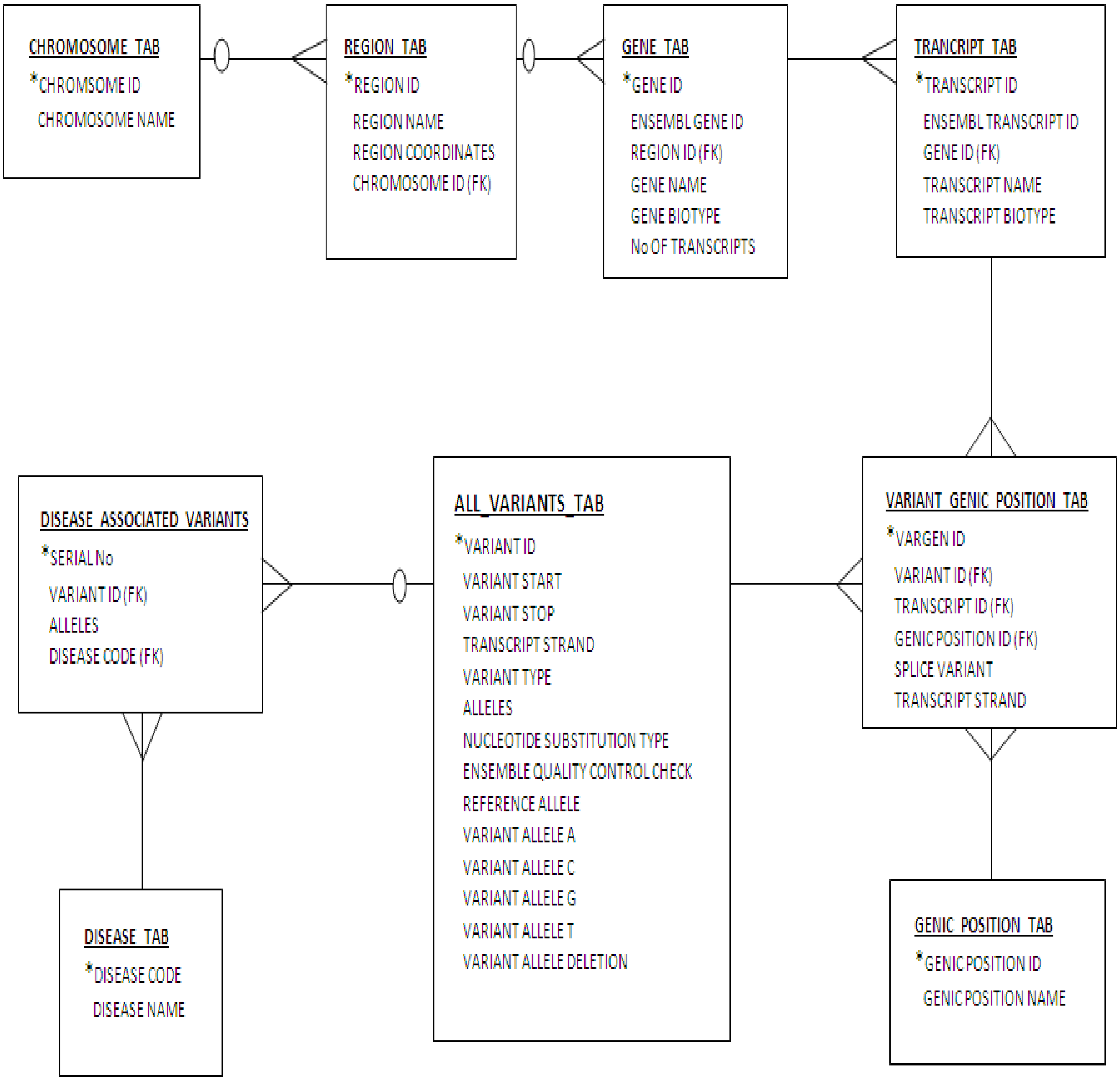


Figure 7. Entity Relationship model for Ravendbase

D.4. Ravendbase Graphical User Interface

In addition to platform conversion, a graphical user interface (GUI) has been designed for Ravendbase. The user friendly GUI was created for easy access to information stored in the database. The database front page is shown in Figure 8. It is kept simple with a short introduction to the database, links to other important web pages within the resource, and an image of the International Diabetes federation logo. More details about the features of the GUI as well as some of the ways by which data can be accessed via the GUI can be seen in Figures 9 to 18. Although the database is still being developed, it is now available online for user access at **http://ravendbase.com/v1/**. Ravendbase is updated regularly as new T1D-SNP information becomes available.

D.5. Future plans for Ravendbase

The database is still undergoing further development. The underlying queries for data retrieval from the database will be fine-tuned to improve retrieval time. Also, all the new data created in this research will be added to the database. These will include the genic profiles and regulatory characteristics of T1D-SNPs.

D.6. Ravendbase Table descriptions

The descriptions of tables that make up the database are shown in Tables 16 to 22

Table 16. REGION_TAB

Column	Type	Default value	Description	Index
REGION_ID	Number		Primary key, internal identifier.	primary key
REGION_NAME	Text		Text containing the names of T1D susceptibility regions	
REGION_START	Number		Gives the chromosomal coordinate of the start position of the susceptibility region	
REGION_START	Number		Gives the chromosomal coordinate of the end position of the susceptibility region	
CHROMOSOMAL_ID	Number		Links region entity to the Chromosome entity	foreign key

Table 17. TRANSCRIPT_TAB

Column	Type	Default value	Description	Index
TRANSCRIPT_ID	Text		Primary key	primary key
ENSEMBL_TRANSCRIPT_ID	Text		Text containing transcript Ensembl ID	

TRANSCRIPT_NAME	Text		Text containing given names	
TRANSCRIPT_BIOTYPE	Text		Text describing transcript biotypes	
GENE_ID	Text		Links transcript entity to the gene entity	foreign key

Table 18. ALL_VARIANTS_TAB

Column	**Type**	**Default value**	**Description**	**Index**
VARIANT_ID	Text		Primary key	primary key
VARIANT_START	Number		Text containing gene Ensembl ID	
VARIANT_STOP	Number		Text containing given gene names	
VARIANT_TYPE	Text		Text describing gene biotypes	
ALLELES	Number		Text showing the alleles of variants	
NUCLEOTIDE_SUBSTITUTION	Text		Text showing if the mutation is a “transition” or a “transversion”	
ENSEMBL_QUALITY_CONTROL CHECK	Text		Text showing if a SNP has “passed” or “failed” the Ensembl quality control check	
REFERENCE_ALLELE	Text		Text showing the reference allele of a SNP	
VARIANT ALLELE_A	Number		Binary numbers indicating if the mutant allele is an ”A”, a“1” indicates positive and a “0” indicates negative	

Table 19. VARIANT_GENIC_POSITION_TAB

Column	**Type**	**Default value**	**Description**	**Index**
VARGEN_ID	Number		Primary key	primary key
VARIANT_ID	Text		Links transcript and genic position entity to the variant entity	foreign key
ENSEMBL_TRANSCRIPT_ID	Text		Links genic position and variant entity to the transcript entity	foreign key
GENIC_POSITION_ID	Text		Links variant and transcript entity to the genic position entity	foreign key
SPLICE_VARIANT	Text		Text showing if any other variant type is also a splice variant	

Table 20. GENIC_POSITION_TAB

Column	Type	Default value	Description	**Index**
GENIC_POSITION_ID	Number		Primary key	primary key
GENIC POSITION_NAME	Text		Text containing names of genic position s	

Table 21. DISEASE_TAB

Column	Type	Default value	Description	Index
DISEASE_CODE	Number		Primary key	primary key
DISEASE_NAME	Text		Text containing names of autoimmune diseases that share susceptibility regions with T1D	

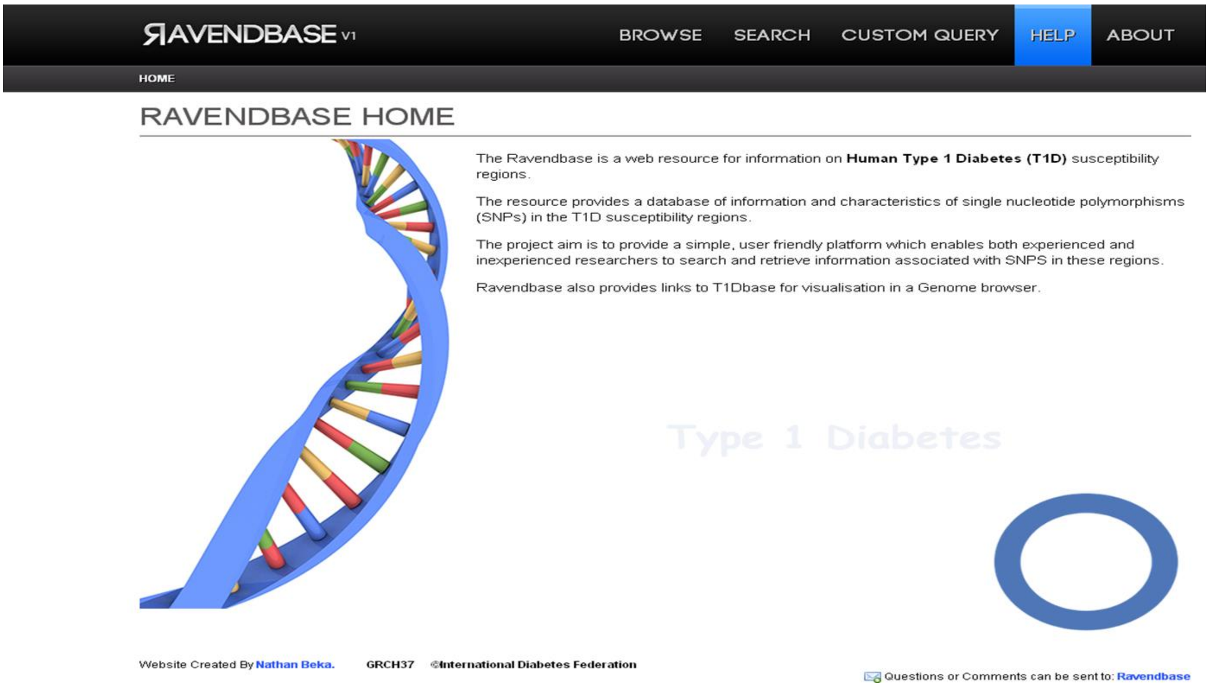
Table 22. DISEASE_ASSOCIATED_VARIANTS_TAB

Column	Type	Default value	Description	Index
SERIAL_No	Number		Primary key	primary key
VARIANT_ID	Text		Links disease marker entity to the variant entity	foreign key
ALLELES	Text		Text showing the alleles of variants (markers)	
DISEASE_CODE	Number		Links disease marker entity to the disease entity	foreign key

D.7. Ravendbase GUI

D.7.1 Browse Page Buttons

The homepage of Ravendbase is shown in Figure 8. This page contains links to other parts of the database. The first link is to the browse page, a user friendly interface for users (scientists and researchers) seeking general information about T1D susceptibility regions. Information sought would include locus coordinates, region size, and number of variants in the region. It also includes named SNPS, genes, and transcripts in the regions. An image of the browse page is shown in Figure 9. There are five link buttons on this web page (T1D Regions, SNPs inT1D Regions, Genes inT1D Regions, Disease Associated SNPs and SNP genic positions) providing region and SNP information. For example, the first button, "T1D regions", displays information about the characteristics of each of the 56 T1D susceptibility regions. The result obtained from clicking the button is shown in Figure 10. A user has the option to download retrieved results from the database as an MS Excel file.



RAVENDBASE HOME

The Ravendbase is a web resource for information on **Human Type 1 Diabetes (T1D)** susceptibility regions.

The resource provides a database of information and characteristics of single nucleotide polymorphisms (SNPs) in the T1D susceptibility regions.

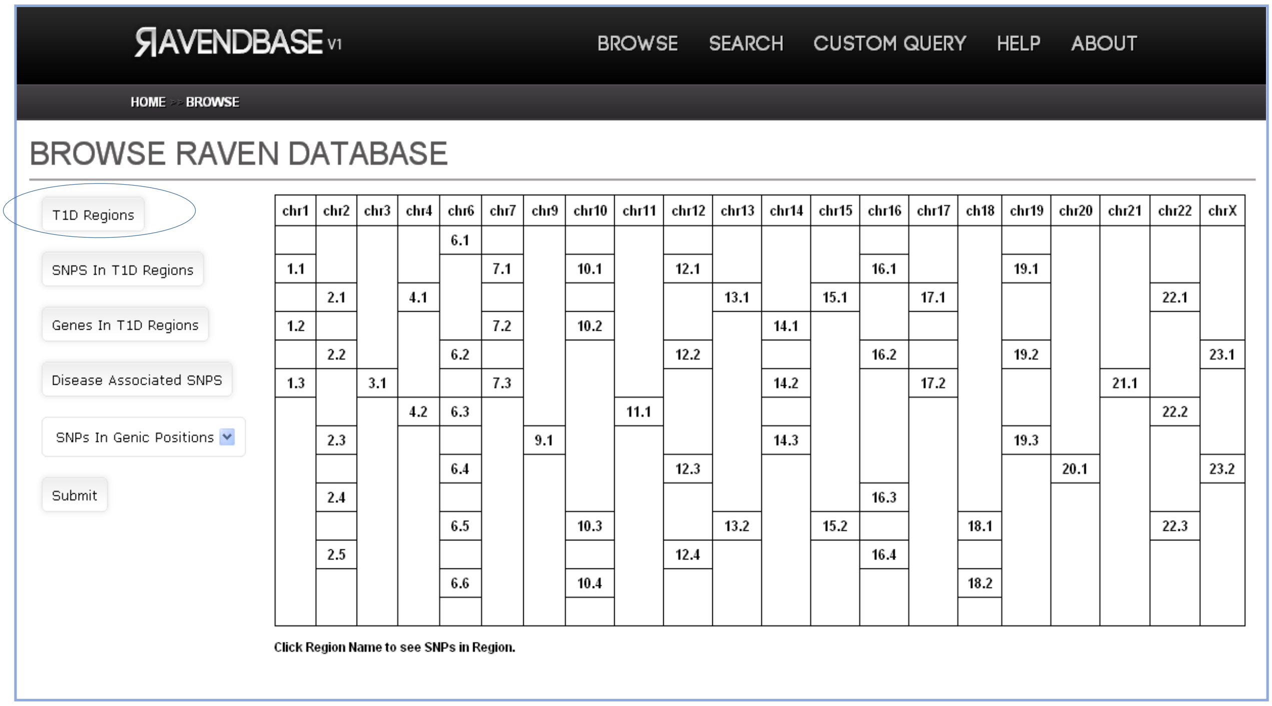
The project aim is to provide a simple, user friendly platform which enables both experienced and inexperienced researchers to search and retrieve information associated with SNPS in these regions.

Ravendbase also provides links to T1Dbase for visualisation in a Genome browser.

Website Created By Nathan Beka.   GRCH37   ©International Diabetes Federation

Questions or Comments can be sent to: Ravendbase

Figure 8. Ravendbase front page

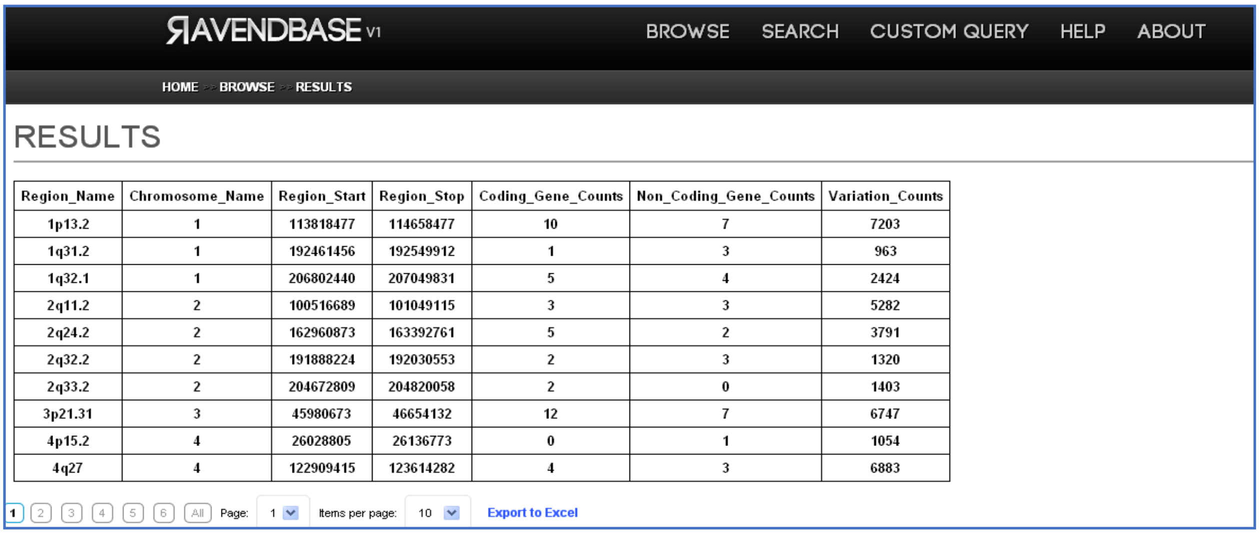


BROWSE RAVEN DATABASE

chr1	chr2	chr3	chr4	chr6	chr7	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrX
				6.1																
1.1					7.1		10.1		12.1				16.1			19.1				
	2.1		4.1							13.1		15.1		17.1					22.1	
1.2					7.2		10.2				14.1									
	2.2			6.2					12.2				16.2			19.2				23.1
1.3		3.1			7.3						14.2			17.2				21.1		
			4.2	6.3				11.1											22.2	
	2.3					9.1					14.3					19.3				
				6.4					12.3								20.1			23.2
	2.4												16.3							
				6.5			10.3			13.2		15.2			18.1				22.3	
	2.5								12.4				16.4							
				6.6			10.4								18.2					

Click Region Name to see SNPs in Region.

Figure 9. Ravendbase browse page



RESULTS

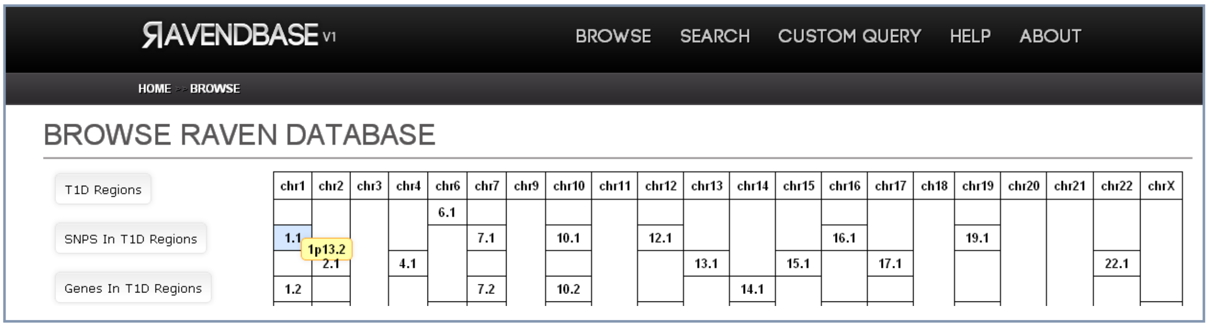
Region_Name	Chromosome_Name	Region_Start	Region_Stop	Coding_Gene_Counts	Non_Coding_Gene_Counts	Variation_Counts
1p13.2	1	113818477	114658477	10	7	7203
1q31.2	1	192461456	192549912	1	3	963
1q32.1	1	206802440	207049831	5	4	2424
2q11.2	2	100516689	101049115	3	3	5282
2q24.2	2	162960873	163392761	5	2	3791
2q32.2	2	191888224	192030553	2	3	1320
2q33.2	2	204672809	204820058	2	0	1403
3p21.31	3	45980673	46654132	12	7	6747
4p15.2	4	26028805	26136773	0	1	1054
4q27	4	122909415	123614282	4	3	6883

Figure 10. Image of result page from Ravendbase's T1D Regions link

D.7.2 Browse page chromosome maze

The chromosome maze is an interesting feature of the database. It is a simple map with T1D region links interspersed in blocks representing chromosomes (Figure 11). Each region has a floating yellow link button which when clicked downloads a result page with general characteristics for all SNPs in the selected region. A small snapshot of region download for

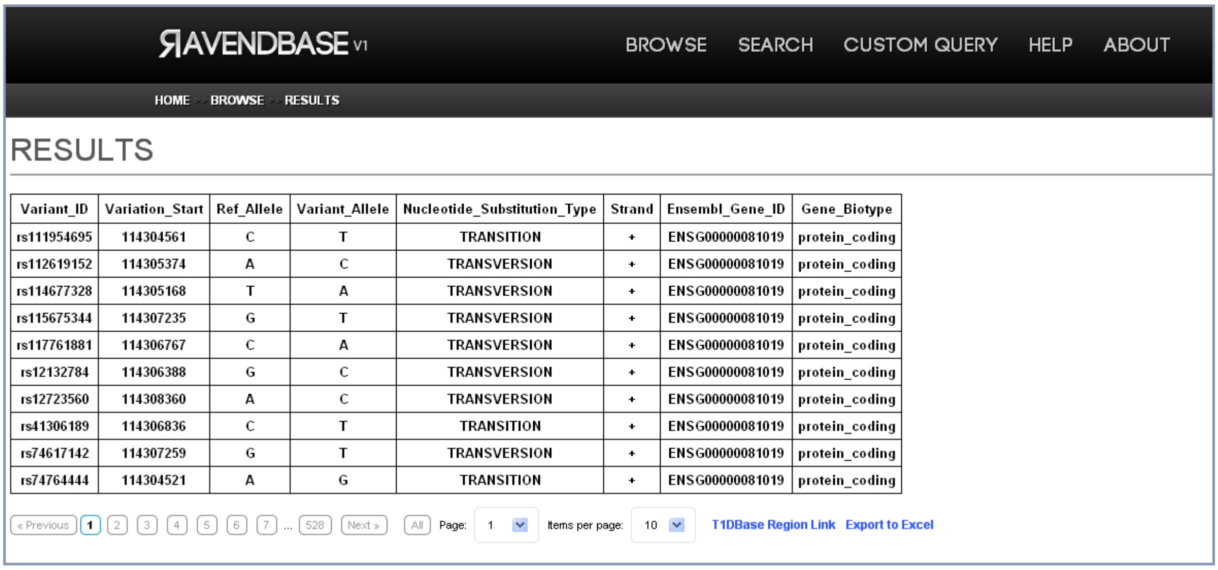
1p13.2 on chromosome 1 is shown in Figure 12. This result page has a link to the same region's page T1Dbase (Figure 13). The reason for this linkage is to allow for viewing the locus in the genome browser provided by T1Dbase (Figure 14).



BROWSE RAVEN DATABASE

chr1	chr2	chr3	chr4	chr6	chr7	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	ch18	chr19	chr20	chr21	chr22	chrX
				6.1																
1.1					7.1		10.1		12.1				16.1			19.1				
	2.1		4.1							13.1		15.1		17.1					22.1	
1.2					7.2		10.2				14.1									

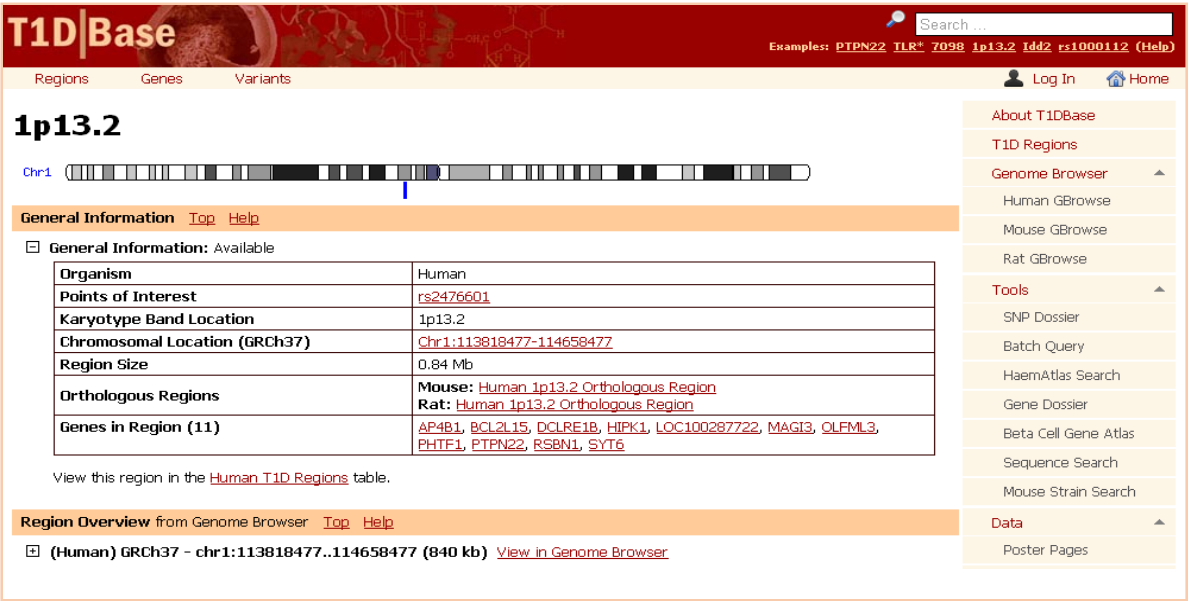
Figure 11. Ravendbase Chromosome maze



RESULTS

Variant_ID	Variation_Start	Ref_Allele	Variant_Allele	Nucleotide_Substitution_Type	Strand	Ensembl_Gene_ID	Gene_Biotype
rs111954695	114304561	C	T	TRANSITION	+	ENSG00000081019	protein_coding
rs112619152	114305374	A	C	TRANSVERSION	+	ENSG00000081019	protein_coding
rs114677328	114305168	T	A	TRANSVERSION	+	ENSG00000081019	protein_coding
rs115675344	114307235	G	T	TRANSVERSION	+	ENSG00000081019	protein_coding
rs117761881	114306767	C	A	TRANSVERSION	+	ENSG00000081019	protein_coding
rs12132784	114306388	G	C	TRANSVERSION	+	ENSG00000081019	protein_coding
rs12723560	114308360	A	C	TRANSVERSION	+	ENSG00000081019	protein_coding
rs41306189	114306836	C	T	TRANSITION	+	ENSG00000081019	protein_coding
rs74617142	114307259	G	T	TRANSVERSION	+	ENSG00000081019	protein_coding
rs74764444	114304521	A	G	TRANSITION	+	ENSG00000081019	protein_coding

Figure 12. Image showing downloaded SNPs in region 1p13.2 using the Chromosome maze



1p13.2

General Information Top Help

⊟ General Information: Available

Organism	Human
Points of Interest	rs2476601
Karyotype Band Location	1p13.2
Chromosomal Location (GRCh37)	Chr1:113818477-114658477
Region Size	0.84 Mb
Orthologous Regions	Mouse: Human 1p13.2 Orthologous Region Rat: Human 1p13.2 Orthologous Region
Genes in Region (11)	AP4B1, BCL2L15, DCLRE1B, HIPK1, LOC100287722, MAGI3, OLFML3, PHTF1, PTPN22, RSBN1, SYT6

View this region in the Human T1D Regions table.

Region Overview from Genome Browser Top Help

⊞ (Human) GRCh37 - chr1:113818477..114658477 (840 kb) View in Genome Browser

Figure 13. Image of T1Dbase page linked to region 1p13.2 in Ravendbase Chromosome maze

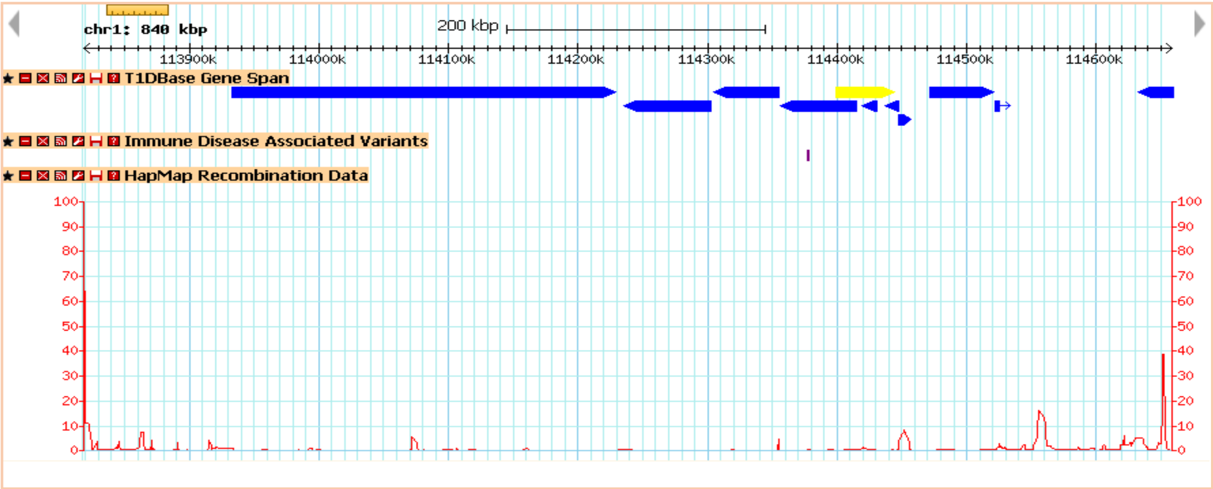
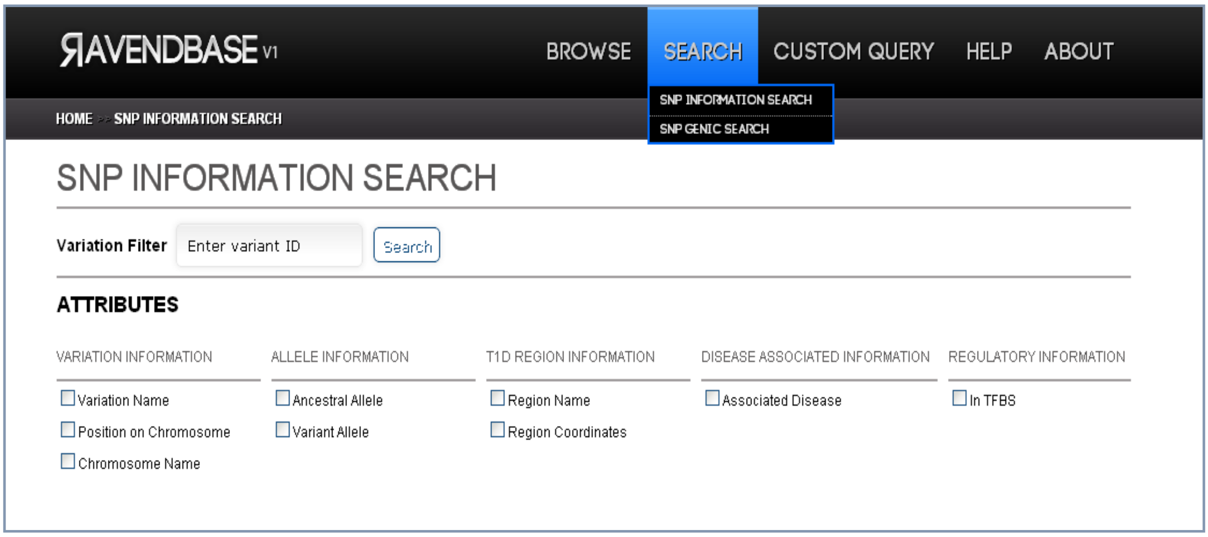


Figure 14. Example of T1D region in T1Dbase genome browser

D.7.3 Search Page: SNP Information Search and SNP Genic Search

The search page was created for users wanting to carry out a more precise searches for a specified SNP. The page has options to select for specific attributes of the SNP of interest which is in a T1D region. The '**SNP Information Search' page** (Figure 15) has a search box were the 'variant ID' of is entered, and nine tick boxes for variant attributes that can be selected for viewing in a result page.



SNP INFORMATION SEARCH

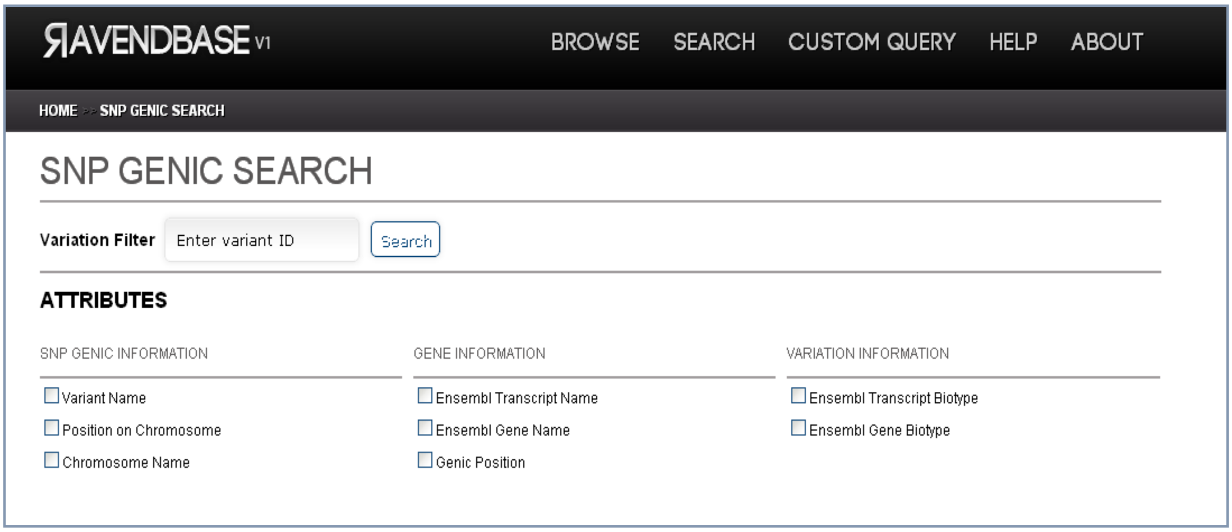
Variation Filter Enter variant ID

ATTRIBUTES

VARIATION INFORMATION	ALLELE INFORMATION	T1D REGION INFORMATION	DISEASE ASSOCIATED INFORMATION	REGULATORY INFORMATION
□ Variation Name	□ Ancestral Allele	□ Region Name	□ Associated Disease	□ In TFBS
□ Position on Chromosome	□ Variant Allele	□ Region Coordinates		
□ Chromosome Name				

Figure 15. SNP Information Search page

The 'SNP Genic Search' page (Figure 16) also has a search box for 'variant ID' input. The page is quite important because it characterises how SNPs sit in the transcripts of genes within its vicinity. Here, there are seven attributes that can be selected for viewing in a result page. Retrieved results from any search such as that shown in Figure 17 also have the option to download as an excel file.



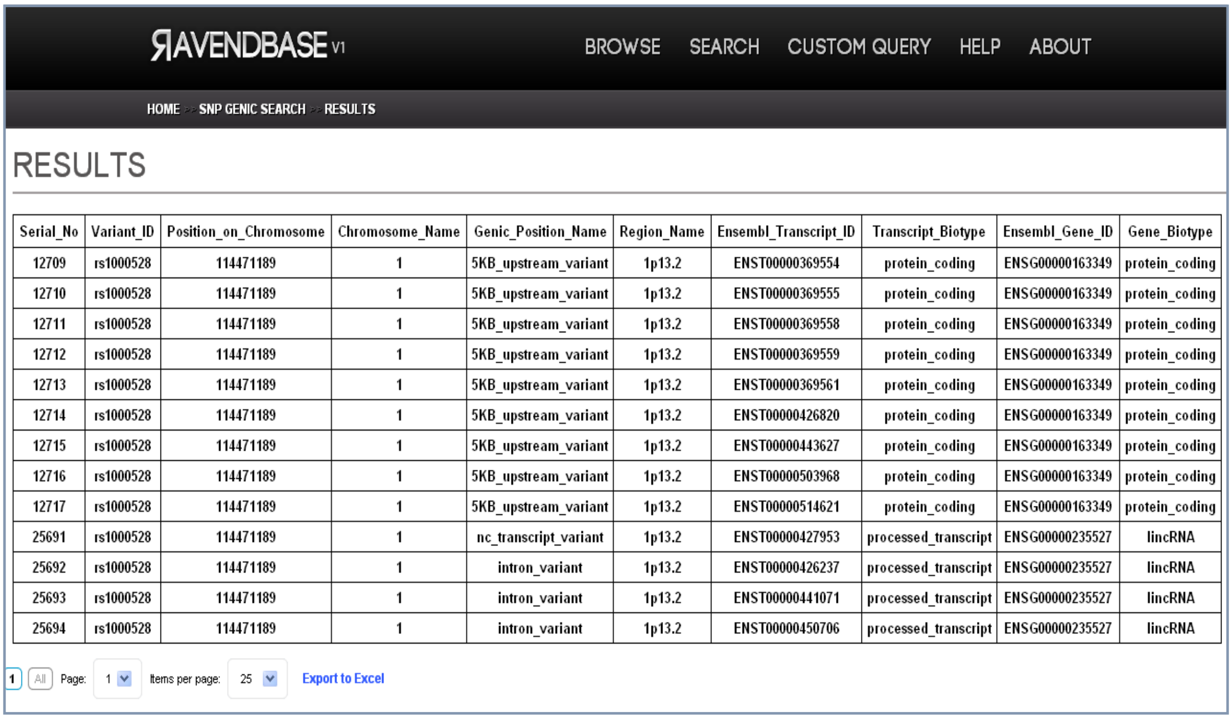
SNP GENIC SEARCH

Variation Filter Enter variant ID

ATTRIBUTES

SNP GENIC INFORMATION	GENE INFORMATION	VARIATION INFORMATION
□ Variant Name	□ Ensembl Transcript Name	□ Ensembl Transcript Biotype
□ Position on Chromosome	□ Ensembl Gene Name	□ Ensembl Gene Biotype
□ Chromosome Name	□ Genic Position	

Figure 16. SNP Genic Search page



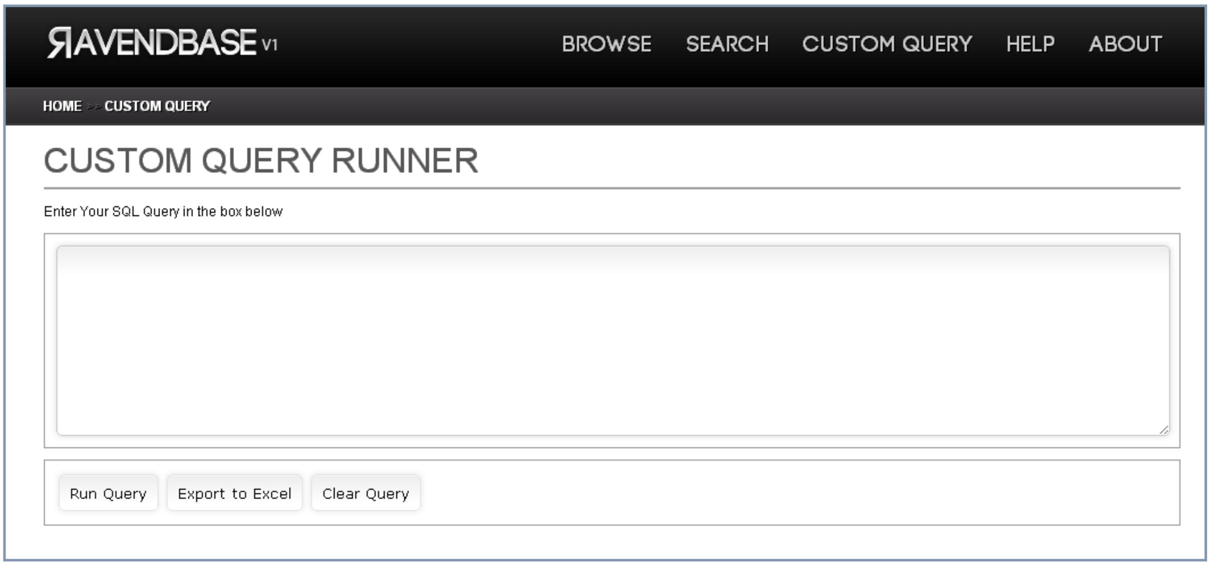
RESULTS

Serial_No	Variant_ID	Position_on_Chromosome	Chromosome_Name	Genic_Position_Name	Region_Name	Ensembl_Transcript_ID	Transcript_Biotype	Ensembl_Gene_ID	Gene_Biotype
12709	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000369554	protein_coding	ENSG00000163349	protein_coding
12710	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000369555	protein_coding	ENSG00000163349	protein_coding
12711	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000369558	protein_coding	ENSG00000163349	protein_coding
12712	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000369559	protein_coding	ENSG00000163349	protein_coding
12713	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000369561	protein_coding	ENSG00000163349	protein_coding
12714	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000426820	protein_coding	ENSG00000163349	protein_coding
12715	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000443627	protein_coding	ENSG00000163349	protein_coding
12716	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000503968	protein_coding	ENSG00000163349	protein_coding
12717	rs1000528	114471189	1	5KB_upstream_variant	1p13.2	ENST00000514621	protein_coding	ENSG00000163349	protein_coding
25691	rs1000528	114471189	1	nc_transcript_variant	1p13.2	ENST00000427953	processed_transcript	ENSG00000235527	lincRNA
25692	rs1000528	114471189	1	intron_variant	1p13.2	ENST00000426237	processed_transcript	ENSG00000235527	lincRNA
25693	rs1000528	114471189	1	intron_variant	1p13.2	ENST00000441071	processed_transcript	ENSG00000235527	lincRNA
25694	rs1000528	114471189	1	intron_variant	1p13.2	ENST00000450706	processed_transcript	ENSG00000235527	lincRNA

Figure 17. Result page showing genic positions of SNP 'rs1000528' in transcripts of two genes

D.7.4 Custom Query Page

The custom query page was designed for the researcher with some programming experience. The user is able to input SQL queries and retrieve specified information from the database. An explanation of the database structure will be made available via the help pages to help users with appropriate query design. The query page is shown below in Figure 18.



CUSTOM QUERY RUNNER

Enter Your SQL Query in the box below

Figure 18. Custom query page

APPENDIX E.

UNPUBLISHED WORK BY ABNIZOVA ET AL.

Computational finding of functional regulatory SNPs

Irina ABNIZOVA
Wellcome Trust Sanger Institute, Cambridge, UK, irina.abnizova@mrc-bsu.cam.ac.uk

Luisa FOCO
University of Pavia, Italy, luisa.foco@unipv.it

Rene te BOEKHORST
University of Hertfordshire, College Lane, Hatfield, UK, r.teboekhorst@herts.ac.uk

Luisa BERNARDINELLI
MRC-BSU, Robinson Way, Cambridge, UK, luisa.bernardinelli@mrc-bsu.cam.ac.uk

This work is devoted to the analysis of human variations in complex human diseases. We present here an in-silico bioinformatics method for inferring possible function of regulatory single nucleotide polymorphisms, SNPs, in human disease development. The research presented here combines the strengths of both genetics and genomics by investigating genetic variants, Single Nucleotide Polymorphisms in regulatory regions instead of genes. By bringing together the computational search and characterisation of regions in DNA that regulate gene expression on the one hand and information about individual variation in the structure of human DNA on the other hand, it aims to identify likely regulatory regions, the individual variation in their molecular make up and the effect this may have in the phenotypic expression of genes.

There is strong recent interest in regulatory SNPs [1-8]. There have been also demonstrated by combining experimental evidence and computation that the promoter regions of human genes provide a rich source of functional single nucleotide polymorphisms [4-8]. As many as 35% of promoter SNPs may be of functional significance [4]. There are, however, currently no computational tools, except of [8] for promoters, which can be used to assess directly from regulatory DNA sequence whether or not a given variant is likely to alter gene expression and hence be of functional significance.

Here, we present the approach that can allow in-silico estimation of the likely functional consequences of single nucleotide changes in putative regulatory DNA. This approach is based on the integration of at least 16 sources of supervised sequence information about a given DNA stretch, with unsupervised methods [9, 10]. We have also incorporated the novel method, which analyse a SNP functionality due to sensitivity of a mathematical model with respect to the SNP variant.

Essentially, the method consists of identifying regions in the human genome that are likely important in the regulation of gene expression and contain motifs that identity them as TFBSs. We then establish whether the motifs contain SNPs and if so, in how far these mutations destroy the signal by which regulatory proteins recognize the motifs as binding sites. Especially these SNPs could be strong candidates for further experimental verification to establish their possible role in the genesis of and susceptibility for particular diseases.

Results. To test the method, we collected several known from literature disease-associated regulatory SNPs [1-3]. We checked if the disease-associated regulatory SNP is within one of the feature-predictions, and thus has a high score. We found that the scores of the disease-associated regulatory SNPs were among the highest scores for all SNPs for all our training sets. Furthermore, these SNPs appeared to be variant sensitive, namely some particular SNP variant changed the results of motif predictions. Interestingly, we found out that known disease-causal SNP variants formed significantly underrepresented motifs within local context.

References

1. Monsuur AJ, de Bakker PI, Alizadeh BZ, Zhernakova A, Bevova MR, Strengman E, Franke L, van't Slot R, van Belzen MJ, Lavrijsen IC, et al. (2005) Nat Genet. 37:1341-4.

2. Ueda H, Howson JM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, Hunter KM, Smith AN, Di Genova G, et al. (2003) Nature 423:506-11.

3. Morahan G, Huang D, Ymer SI, Cancilla MR, Stephen K, Dabadghao P, Werther G, Tait BD, Harrison LC, Colman PG (2001) Nat Genet. 27:218-21.

4. Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, S. K., O'Donovan, M. C. and Buckland, P. R. (2004). Functional analysis of polymorphisms in the promoter regions of genes on 22q11. Hum. Mutat. 24, 35-42.

5. Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief. Bioinform. 6, 44-56

6. Pastinen, T. and Hudson, T. J. (2004). Cis-acting regulatory variation in the human genome. Science 306, 647-650

7. Hudson, T. J. (2003). Wanted: regulatory SNPs. Nat. Genet. 33, 439-440

8.Paul R. Buckland , Bastiaan Hoogendoorn, Sharon L. Coleman, Carol A. Guy, S. Kaye Smith, Michael C. O'Donovan (2005) Strong bias in the location of functional promoter polymorphisms,

9. Khan I, et al. and Chuzhanova N. (2006) In silico discrimination of single nucleotide polymorphisms and pathological mutations in human gene promoter regions by means of local DNA sequence context and regularity, In Silico Biology 6, 0003

10. Irina Abnizova, Alistair G. Rust, Mark Robinson, Rene te Boekhorst and Walter R. Gilks. (2006) Prediction of TFBS using Markov models, J. of Bioinformatics and Comp. Biology, v4, n2, pp 425-441

11. Irina Abnizova, Rene te Boekhorst, Klaudia Walter and Walter R. Gilks. (2005), Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in eukaryotic genomes: the fluffy-tail test. BMC Bioinformatics, 6:109