The Role of 3D Human Genome Architecture in Mutability – From Predicting Penetrance/Gene Fusions to Discovering Novel Schizophrenia-Associated Variants

A thesis by Daniel Buxton submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy.

Department of Physics & Mathematics

Nottingham Trent University

December 2017.

Declaration

I hereby declare that the dissertation submitted for the degree of Doctor of Philosophy in Bioinformatics on “the role of 3D human genome architecture in mutability – from predicting penetrance/gene fusions to discovering novel schizophrenia-associated variants” at Nottingham Trent University, is my own original work and has not previously been submitted to any institution or university, or quoted as indicated and acknowledged by means of a comprehensive list of references.

Copyright Statement

Acknowledgements

I would like to express gratitude to my supervisor, Professor Nadia Chuzhanova, for her invaluable guidance and support during my PhD studies. My journey has been full of new experiences and challenges, and I am grateful to have had a supervisor that was more than willing to help me through each step. I also wish to thank my second supervisor, Dr Jonathan Crofts, for his many important contributions throughout the last four years. I am also indebted to the Mathematics department of Nottingham Trent University for their encouragement and discussions – particularly those who I have shared numerous offices with during my time here! I would also like to thank the Vice Chancellor's bursary for providing me with the funding to undertake this postgraduate study at NTU.

I would like to express my final thanks to my friends and family, for always being there and for occasionally offering a unique perspective on my work. I wish to extend a special gratitude to my parents, and to my partner, Harriet. Without their unwavering support and reassurance, I am sure this process would have been infinitely more difficult. Thank you.

Publications and Presentations

Data described in chapter 2 and methods implemented in chapters 4, 5 and 6 contributed to the writing of an article (see citation below).

> Mayes, M.B., Morgan, T., Winston, J., Buxton, D.S., Kamat, M.A., Smith, D., Williams, M., Martin, R.L., Kleinjan, D.A., Cooper, D.N., Upadhyaya, M. and Chuzhanova, N., 2015. Remotely acting SMCHD1 gene regulatory elements: in silico prediction and identification of potential regulatory variants in patients with FSHD. *Human Genomics,* 9 (1), 25.

The material used in this article was the investigation of the distribution of interaction frequencies in dilution Hi-C data, particularly interactions with the fragment that contained gene *SMCHD1*.

In addition, the analysis carried out in chapter 5 was given as a presentation entitled "Mutability in 3D: exploring the role of spatial organisation of the human genome" at the 9th Science and Technology Annual Research (STAR) Conference, Nottingham Trent University, May 2015.

The material discussed in chapter 6 was given as a poster presentation entitled "Predicting genomic regions linked to schizophrenia using the 3D architecture of the human genome" at ECCB (European Conference on Computational Biology), The Hague, The Netherlands, September 2016.

Abstract

We have become very familiar with the genome being represented as a one-dimensional sequence of the four nucleobases – cytosine, guanine, adenine and thymine. However, in reality this chain folds and is densely packed into the nucleus of eukaryotic cells in a three-dimensional (3D) setting, meaning that pairs of otherwise remote areas of the genome can come into close proximity in 3D space. It is thought that the expression of target genes is influenced by remotely acting regulatory elements, such as enhancers, which are often located several kilobases away from the genes they target.

In our studies we hypothesised that communication between widely spaced genomic elements is facilitated by the spatial organisation of chromosomes that bring genes and their regulatory elements in close spatial proximity. We explored this hypothesis in three distinct contexts: (1) reduced/incomplete penetrance, where disease genotypes do not always induce the expected phenotype; (2) gene fusion events, known to be frequent in cancer; (3) schizophrenia, a complex brain disorder. Whilst previous studies acknowledged the role of polygenic activity in these genetic diseases and phenomena, they did not integrate this idea into existing detection/prediction techniques. Our analysis addressed this oversight by transforming traditionally one-dimensional studies into a contextually relevant, 3D setting.

We utilised data describing the 3D structure of the human genome, alongside prior knowledge of various diseases and genetic phenomena, to predict novel genomic regions of association. Our approaches incorporated network, statistical and computational methods to identify where these regions of interest lie. Identified regions were investigated further to ascertain biological properties, such as an enriched presence of mutations, functionally relevant genes, regulatory elements, or all of the above. Whilst existing approaches tend to fixate on only these static properties, our studies also focused on the communication of otherwise remote regions by creating 3D interaction networks that describe the spatial proximities of genomic fragments. The most important units of such networks were identified via centrality measures and statistical testing, followed by subsequent biological interrogation of so-called candidate regions. This method ultimately confirmed whether regions were genuinely disease-associated via polygenic activity, or not.

A total of 35 novel schizophrenia candidate regions were identified using our approach, 22 of which contained polymorphisms with prior schizophrenia association; most variants found were shown to influence gene expression specifically in brain tissues. We were also successful in showing that cancer-causing gene fusion events are catalysed by paired fusion gene-containing fragments (of lengths 1 megabase and 100 kilobases) sharing small 3D neighbourhoods, particularly for genes residing on different chromosomes. Our transformation of existing approaches into 3D studies has therefore elucidated features and properties of genetic disease and cancer that were otherwise unknown or overlooked.

List of Figures

List of Tables

List of Abbreviations

[1D]	One-dimensional
[3C]	Chromosome conformation capture
[3D]	Three-dimensional
[4C]	Chromosome conformation capture on ChIP
[5C]	Chromosome conformation capture carbon copy
[AML]	Acute myeloid leukaemia
[ANOVA]	Analysis of variance
[aveALL]	Average across all brain regions
[BCT]	Brain connectivity toolbox
[bp]	Base pair(s)
[Braineac]	Brain eQTL almanac
[CAGE]	Cap analysis of gene expression
[CD/CV]	Common disease/common variant
[CEU]	Utah residents from north and west Europe
[CHi-C]	Capture Hi-C
[ChIP]	Chromatin immunoprecipitation
[ChiTaRS]	Chimeric transcripts and RNA-seq
[CML]	Chronic myeloid leukaemia
[COSMIC]	Catalogue of somatic mutations in cancer
[CRBL]	Cerebellar cortex
[DAVID]	Database for annotation, visualisation and integrated discovery
[DGN]	Disease gene network
[DNA]	Deoxyribonucleic acid
[EASE]	Expression analysis systematic explorer
[EGR]	Extended gene region
[ENCODE]	Encyclopedia of DNA elements
[eQTL]	Expression quantitative trait loci
[ExAC]	Exome aggregation consortium
[FANTOM]	Functional annotation of the mammalian genome
[FCTX]	Frontal cortex

[FUSIM]	Fusion simulator
[GE]	Global efficiency
[GWAS]	Genome-wide association study
[GWKR]	Genome-wide Knight & Ruiz
[GWVC]	Genome-wide vanilla coverage
[HDN]	Human disease network
[HGMD]	Human gene mutation database
[HIPP]	Hippocampus
[HK]	Harel-Koren
[IF]	Interaction frequency
[INTERKR]	Inter-chromosomal Knight & Ruiz
[INTERVC]	Inter-chromosomal vanilla coverage
[kb]	Kilobase
[KR]	Knight & Ruiz
[LD]	Linkage disequilibrium
[LE]	Linkage equilibrium
[MAGMA]	Multi-marker analysis of genomic annotation
[Mb]	Megabase
[MEDU]	Inferior olivary nucleus, subdissected from the medulla
[MRI]	Magnetic resonance imaging
[OCTX]	Occipital cortex
[OMIM]	Online Mendelian inheritance in man
[PCR]	Polymerase chain reaction
[PPI]	Protein-protein interaction
[PRS]	Polygenic risk score
[PUTM]	Putamen
[RAW]	Raw interaction count
[RNA]	Ribonucleic acid
[RNA-seq]	RNA sequencing
[rsID]	Reference SNP cluster identification
[SNIG]	Substantia nigra
[SNP]	Single nucleotide polymorphism

[SQRTVC]	Square-rooted vanilla coverage
[SSTAR]	Semantic catalogue of samples, transcription initiators and regulators
[TCGA]	The cancer genome atlas
[TCTX]	Temporal cortex
[TF]	Transcription factor
[THAL]	Thalamus
[TICdb]	Database of translocations in human cancer
[TPM]	Tags per million
[TSS]	Transcription start site
[VC]	Vanilla coverage
[WGS]	Whole-genome sequencing
[WHMT]	Intralobular white matter

Contents

Chapter I

Introduction

Our current understanding of genetics originates from as early as the mid-nineteenth century, with Gregor Mendel's experiments on smooth/wrinkled pea plants (Mendel, 1865). His work was the first to describe inheritance as a culmination of discrete hereditary units (now known as genes) rather than a blend of parental traits. At the time, Mendel focused on the biological output, or observed traits (phenotype), in order to make implications about the nature of the genetic input (genotype).

Nowadays, thanks to the various stages of DNA discovery (reviewed in Dahm, 2008; Watson & Crick, 1953), we know the physical structure of our genome. We are also beginning to understand that changes in the genome, called polymorphisms (that occur in >1% of the population) or mutations (that occur in <1% of the population), may or may not alter our phenotype. Quite often, a single mutation does not change any physical trait, although a specific collection of mutations could work together to induce phenotypic change. There are many well-understood cases of single gene mutations causing disease (monogenic disorders), such as sickle cell disease (Ashley-Koch et al., 2000), cystic fibrosis (Ratjen & Döring, 2003), polycystic kidney disease (Adeva et al., 2006) and Tay-Sachs disease (Neudorfer et al., 2005). However, most diseases are complex, involving a number of different mutations at multiple genes. It is the interactions and relationships between genes in this context that is understood to a lesser extent – particularly when a group of related genes are at seemingly remote distances from one another. This has been previously reported as the *omnigenic model* (Boyle et al., 2017).

Single nucleotide polymorphisms (SNPs) are individual DNA base pair changes common in the human genome. Most SNPs occurring in humans do not induce any negative effects, but it is important to catalogue those that do, whether they are a major participant in a monogenic disorder or a minor participant in a polygenic disorder. The cataloguing procedure is achieved via genome-wide association studies (GWAS): SNPs are measured and analysed from across the genomes of patients and healthy individuals in order to reveal genetic risk factors for disease (Bush & Moore, 2012). For common, monogenic disorders, this process is straightforward (given a large enough sample size). However, for rarer or more complex diseases involving multiple SNPs at multiple genes, the process is less clear, since a number of SNPs could have a large combined effect despite having individually small effect sizes (this idea is explored further in section 6.1.2).

This uncertainty motivated advances in genome sequencing (reviewed in Goodwin et al., 2016) and the interrogation of 3D genomic structure by measuring proximities of DNA fragment pairs (described at length in chapter 2). The ENCODE (Encyclopaedia of DNA Elements) Project has also been responsible for cataloguing various regulatory DNA elements which regulate gene expression; these elements lie in regions that were previously thought to have no functional relevance (so-called *junk DNA*; The ENCODE Project Consortium, 2012).

Improving the accuracy, cost and cell line availability for these new laboratory techniques is now the main focus, especially if we wish to use resulting data to investigate cell-specific diseases. With increasingly comprehensive laboratory data available, we are better able to identify novel long-range relationships between genomic regions by employing various network techniques (described in chapter 3) and other bioinformatics tools. Furthermore, the amalgamation of cutting edge methods for quantifying 3D genomic distances alongside the techniques of network science promises to provide a gateway into identifying previously unknown regulators of genetic disease.

In particular, we focus on the effects of remotely acting SNPs in three areas: (1) schizophrenia, a complex disorder of the brain; (2) reduced penetrance, a phenomenon where one's genetic code does not always predict the expected trait; (3) cancer-associated gene fusion events, where two previously separate (and often remote) genes merge as a result of the rearrangement of large chromosomal segments.

The proceeding sections outline our formulated hypotheses, which is followed by a historical review of existing approaches that partially address the problems we discuss. We then go on to review previous biological applications of network analysis, with the view to incorporate similar techniques in our work.

1.1. Thesis Aims and Objectives

The aim of this study is to combine cutting edge data quantifying the 3D structure of the human genome and network science approaches to identify and investigate previously unknown remotely acting regulatory regions underlying genetic disease and cancer. The following text outlines the hypotheses for each of our studies, which are described in more intricate detail in later chapters. All computational analysis in this thesis was implemented in version R2016a of MATLAB and/or version 3.6 of Python.

For the reduced penetrance phenomenon, we hypothesised that the folding principles of the human genome are at least partially responsible for the level of penetrance exhibited in a population. We aimed to test this hypothesis by showing that (1) remote pairs of genes known to participate in the incomplete penetrance of a disease are within shared 3D neighbourhoods; (2) all or most known genes share communications with common genomic fragments; (3) each unique gene pair communicates with their own distinct genomic fragments.

For cancer-causing gene fusion events, we hypothesised that the 3D structure of the human genome contributes to the choice of fusion partners. This was investigated using two main approaches: (1) we proposed that any region harbouring a fusion gene would show a closeness to other fragments within the cell nucleus, and (2) we hypothesised that one half of a known fusion gene pair would consistently favour a particular gene as a fusion candidate.

Finally, for our analysis of schizophrenia, we hypothesised that both common and individually rare variants exert their influence on gene expression, either (1) directly via long-range looping interactions between fragments that harbour regulatory elements and target genes, or (2) are propagated via the network of communications governed by the 3D architecture of the human genome. We also further investigated any interesting regions found to determine various factors such as the ontology of contained genes, the relative expression of such genes in brain tissues and the presence/absence of mutations. This was in order to build evidence for regions having a genuine and previously unfound association with schizophrenia.

1.2. Penetrance

Penetrance is a genetic phenomenon whereby an individual can harbour a particular disease mutation without actually exhibiting any of the disease traits. The terms *penetrance* and *expressivity* are distinct, but are often mistaken as being identical. In actual fact, penetrance for an individual patient is the binary identifier of phenotype presence, given a particular genotype, whereas expressivity describes the (non-binary) extent to which an existing phenotype is expressed. Hence, despite the distinctions, there is overlap in the two phenomena; it is possible for one to view penetrance as a pole on the expressivity scale (Zlotogora, 2003). It is also important to note that penetrance can be measured on a scale, although this scale would be the ratio of patients from a population that exhibit the binary true phenotype to those that do not.

1.2.1. Approaches that Predict Penetrance

Since penetrance is typically a lifetime measure (i.e. the detection of phenotype presence at any stage during one's lifetime), the ideal way to predict penetrance would be to obtain sequenced genomes from birth and track each patient's disease status until death. This would allow us to group patients with a similar genotype together and identify a subset that developed the disease in question at some point before death, thus giving us the true penetrance. Unfortunately, this method has some logistical problems, such as: (1) it would take a whole lifetime to generate meaningful results, and (2) despite recent improvements, the cost of sequencing a single human genome is still too large to be performed on a mass scale (initially costing approximately $100 million per genome in 2001, to several thousands of dollars today; Wetterstrand, 2017).

Therefore, researchers have devised more practical methods to estimate the penetrance of genetic diseases. To circumnavigate time and cost, it has become common practice to seek out individuals with a disease phenotype and work backwards to ascertain their genotype. Typically, immediate relatives would also be investigated and a level of penetrance would be predicted from a collection of resulting family samples, in what is commonly known as the *family-based* method

(Ford et al., 1998; Wang et al., 2006; Gong et al., 2010). The family-based method, however, is prone to overestimating penetrance based on the biased way in which data is collected (Minikel et al., 2014). The sampling is not representative of the population, since (1) subjects were preferentially chosen based on exhibiting the phenotype of a rare disease, (2) families were preferentially chosen if more than one relative exhibited the phenotype and (3) relatives not displaying the phenotype may not have been genotyped.

The criticisms described above are difficult to avoid, since it is not common practice to test seemingly healthy individuals for rare disease; resources will usually only be used for those that show relevant symptoms. Despite this mentality, attempts have been made to eliminate the preferential sampling bias, such as the kin-cohort method (Wacholder et al., 1998), which obtained the 70-year penetrance of mutations in the *BRCA1* and *BRCA2* genes responsible for breast/ovarian cancer (reported as 63%), by genotyping randomly selected individuals and first-degree relatives. This, however, is only feasible for common mutations, since it is possible that random sampling will not pick up any individuals affected by a rare variant, even for large sample sizes.

Due to the various limitations of family-based techniques, recent research has shifted to *population-based* methods, which is principled on the assumption that a genetic variant with perfect penetrance is no more present in the population than its corresponding phenotype. However, population-based methods for complex diseases rely on the availability of large databases that capture genetic variation. The Exome Aggregation Consortium browser (ExAC) partially solves this problem; it has summary data of 60,706 unrelated sequences obtained from an array of published disease studies (Lek et al., 2016). Allele frequencies for uncommon disease-causing variants can easily be queried using ExAC, which therefore gives the user a clear idea of the expected phenotype frequency and enables them to compare this to an observed phenotype frequency, thus measuring penetrance.

The key advantage with this method is the lack of sampling bias seen for family-based methods, although one should be wary of how variants in population-based databases have been classified as disease-associated. Databases such as the Human Gene Mutation Database (HGMD) have information for approximately 203,000 unique gene lesions obtained from around 2,600 journals (Stenson et al., 2017), although if each of the reported disease-associated lesions resulted in the expected phenotype, the average individual in ExAC would have a plethora of genetic diseases (Lek et al., 2016). It is therefore difficult to distinguish between reports of genuine disease-causing mutations that have incomplete penetrance to those that are simply falsely interpreted as being associated. We suggest that a possible way to alleviate this problem is to use 3D interaction data to determine whether variants participate in polygenic activity. This would act as a vetting process for individual variants that would either strengthen or weaken the claim that they are disease-associated and have low penetrance. For example, variant S_1 has a reported association with bipolar disorder, but is almost completely non-penetrant according to ExAC data. It is possible that S_1 interacts with a remote variant, S_2, which in turn causes an increased penetrance of bipolar disorder. Ordinarily, S_1 may have been reported as being falsely associated, whereas S_2 would claim full credit for the association with bipolar disorder. In reality, S_1 is not discounted, since it participates in long-range polygenic activity that increases the effect of S_2.

1.3. Gene Fusion Events

A *fusion gene* forms from the merging or two previously separate (and often remote) genes as a result of rearrangements of large chromosomal segments. The first discovery of a gene fusion was initially mistaken in 1960 for a large deletion, resulting in patients with chronic myeloid leukaemia (CML) having an unusually small chromosome (Nowell, 2007). Since this initial observation, it was shown that a translocation had occurred between breakpoints at the *ABL1* gene on chromosome 9 and the *BCR* gene on chromosome 22. The resulting *BCR-ABL1* fusion was the first discovery of its kind; the host of the fusion gene itself would later be termed the *Philadelphia chromosome*, named after the location of its origin (Mitelman et al., 2007).

1.3.1. Databases of Fusion Genes

Subsequent years of research have led to a developed understanding of chromosomal rearrangement. Because of the significant ties fusion genes have with carcinogenesis (Edwards, 2010), numerous attempts have been made to comprehensively list all known and potential cases through varied approaches (Novo et al., 2007; Kim et al., 2007; Kim et al., 2006, 2009; Frenkel-Morgenstern, 2012; Forbes et al., 2014; Wang et al., 2015, Lee et al., 2017). The number of putative gene fusion events has increased dramatically from several hundred (358; Mitelman et al., 2007) to thousands (10,534; Forbes et al., 2014).

Initial efforts focused on the identification of translocation breakpoints as a first step towards finding fusion genes. The database of Translocations in Human Cancer (TICdb; Novo et al., 2007) catalogues the genomic locations of 1,225 breakpoints in tumours; 949 were shown at nucleotide level and the remaining 276 were described at intron/exon resolution. In later work, an increasing number of breakpoints were found (8,943 deletion breakpoints) and their association with epigenetic features and 3D proximity was investigated (Abyzov et al., 2015).

The Catalogue of Somatic Mutations in Cancer (COSMIC; Bamford et al., 2004; Forbes et al., 2016) was also being curated at a similar time to TICdb. Originally, this database identified four cancer genes, but has since utilised 12,542 cancer genomes to identify over 10,000 fusion mutations as well as millions of coding and non-coding mutations (Forbes et al., 2014). Both databases validate their entries by searching existing literature via PubMed and online data portals.

Further research incorporated computational techniques in order to predict fusion regions as well as identify them. Developments of sequencing technologies gave way to detections of more fusions, particularly through transcriptome analysis (Wang et al., 2015). The second version of ChimerDB was created using human transcriptome (RNA-seq) data, with fusion candidates pruned through literature mining (Kim et al., 2009). This consequently acted as a reliable reference dataset for computational tools that sought to predict fusion regions, such as the Fusion Simulator (FUSIM; Bruno et al., 2013). The database of Chimeric Transcripts and RNA-seq data (ChiTaRS;

Frenkel-Morgenstern et al., 2012) is an extension of ChimerDB; it includes approximately 2,000 cancer breakpoints from humans, mice and fruit flies. It also describes expression relative to specific tissues, as well as ranking regions based on the strength of evidence for breakpoint presence among repeated experiments.

Recently, ChimerDB has undertaken a further revision (ChimerDB 3.0; Lee et al., 2017). It is a development of version two that includes further data via a three-pronged attack: firstly through a knowledgebase from public resources and experimental evidence (such as COSMIC, ChimerDB 2.0, TICdb, OMIM and Mitelman's database); secondly via keyword mining of PubMed abstracts; and finally from next-generation sequencing data as well as further data on the molecular characteristics of human cancer, such as The Cancer Genome Atlas (TCGA; The Cancer Genome Atlas Research Network, 2008). ChimerDB 3.0 is therefore perhaps the most complete and reliable database as a result of its incorporation of historical and new identification techniques.

1.3.2. Approaches that Predict Fusion Events

Fusion events are characterised by a large variation in linear genome structure, usually via deletions or insertions of relatively long DNA sequences (Wang et al., 2012). Gene fusions are of particular interest because of their strong link to a number of cancers (Watson et al., 2013; Yoshihara et al., 2015). The size of the deletion/insertion is what makes fusion events distinct from other mutations; computational techniques that aim to identify and predict fusion events are therefore designed to detect this signature by comparing various sample sequences to a reference sequence.

In recent years, the efficiency of various sequencing methods, collectively known as *next generation sequencing* (NGS) technologies, has vastly improved. It is no surprise that the development of NGS has given way to advances in detection and prediction of gene fusion events, since computational tools heavily rely on the availability of sequence data. Almost all detection algorithms utilise data from one of three NGS methods: *whole genome sequencing* (WGS), *whole transcriptome sequencing* (commonly referred to as *RNA-seq*) and *targeted sequencing*.

As its name implies, WGS is a thorough method that sequences the entire collection of approximately 3 billion base pairs of coding and non-coding DNA. Bioinformatics tools that utilise data from WGS have identified an array of novel cancer-causing fusion events, including: various fusions of genes *PML*, *RARA* and *LOXL1* affecting acute myeloid leukaemia (AML; Welch et al., 2011); fusion gene *VTI1A-TCF7L2* causing colorectal cancer (Bass et al., 2011); fusion gene *PVT1-CHD7* causing lung cancer (using lung cell lines; Pleasance et al., 2010). The benefit of using WGS over other NGS methods is its comprehensiveness, although this can come at a huge computational cost. The start-to-end process of detecting and validating a fusion gene can take months (for example, WGS identification of gene fusions causing AML took 7 weeks; Welch et al., 2011). Also, the inclusion of non-coding regions implies that an additional step of observing the expression of potential fusion sites is required before classifying them as genuine fusion events.

Mostly because of the computational cost of WGS methods, RNA-seq tools have emerged as the preferred choice to identify gene fusions. The transcriptome is a collection of protein-coding genes derived from the original recipe of the genome and therefore consists of a small subset of the initial ~3 billion base pairs (approximately 2%; Sboner et al., 2010). This poses a significant speed advantage over WGS at the cost of reduced complexity. Nevertheless, it is evident that RNA-seq is the preferred choice of NGS data, since the majority of recent studies use it for detection and prediction. Examples include: a total of 51 novel fusions found to affect breast cancer (Zhao et al., 2009; Maher et al., 2009a; Ha et al., 2011; Edgren et al., 2011; Robinson et al., 2011); 39 fusion genes affecting prostate cancer (Maher et al., 2009a; Maher et al., 2009b; Palanisamy et al., 2010; Nacu et al., 2011; Pflueger et al., 2011); fusion gene *AGTRAP-BRAF* causing gastric cancer (Palanisamy et al., 2010). Despite the popularity of RNA-seq techniques, one must be wary of the limitations that a reduction of complexity causes, such as its inability to detect fusion sites in non-coding regions. Furthermore, genes have typically variable expression profiles across different tissues, hence making the detection of a genuine fusion transcript slightly more complicated than the comprehensive WGS method.

Targeted sequencing is a lesser-used NGS method that is perhaps seen as the middle-ground between WGS and RNA-seq, in that it maintains the complexity of WGS methods, whilst performing at a speed comparable to RNA-seq. The relative speed of this method comes as a result of only a portion of the genome being sequenced. This has been successfully applied in some cases, such as identifying 11 novel fusion genes causing leukaemia (Grossman et al., 2011) and also detecting the fusion gene *C6orf204-PDGFRB* affecting T lymphoblastic lymphoma (Chmielecki et al., 2010). However, it is difficult to apply targeted sequencing to the general detection and prediction of fusion events, because of its reliance on prior knowledge.

It seems that of the two most commonly used NGS techniques, WGS and RNA-seq, the trade-off when considering which to use tends to be speed versus complexity. Hence, given the continued increase of computing power coupled with the optimisation of bioinformatics techniques, it is perhaps feasible to detect and predict fusion events by finding overlaps between both methods. This has already been successfully attempted with a study validating fusion events responsible for prostate cancer (McPherson et al., 2011), although the study cited a large computational time as a limitation, and one could argue that requiring a common agreement between both techniques could lead to genuine gene fusions being missed.

1.4. Schizophrenia

Schizophrenia is a chronic, severe and disabling brain disorder, with common symptoms including hallucinations, delusions and changes in behaviour. Many brain disorders, including schizophrenia, are typically diagnosed based on the observations of a subject's behaviour and functions rather than on genetic factors (McCarthy et al., 2014). This is especially surprising given that schizophrenia is widely established as a heritable disease (Cardno & Gottesman, 2000; Keller et al., 2012). The historical aim has been to be able to diagnose and treat schizophrenia based on knowledge of genetic markers, rather than rely on observed phenotype. Despite huge advances in sequencing techniques, there is still an absence of individual genes known to have a large influence on schizophrenia (Kumar et al., 2014).

1.4.1. Approaches that Predict Schizophrenia-Associated Variants

Genome-wide association studies (GWAS) have, however, succeeded in identifying a community of common variants that have a small effect on the disorder (Ripke et al., 2011, 2013). These variants are among millions of single nucleotide polymorphisms (SNPs) that are assigned schizophrenia association p-values. Furthermore, common SNPs, such as the ones found by GWAS, are responsible for an estimated 23% of total schizophrenia risk (Lee et al., 2012), with this percentage projected to increase after an increase of sample sizes.

Whilst these studies are effective at introducing small-effect SNPs and implying their combined influence on schizophrenia, the interactions between SNPs and underlying mechanisms were seldom explored. Hence, emphasis seems to have shifted to understanding and identifying polygenic activity in schizophrenia development. A case-control study found thousands of common alleles of small effect size contributing to a considerable schizophrenia risk, as well as other brain-related disorders such as bipolar disorder (Purcell et al., 2009). Results like this have motivated the polygenic risk score (PRS; Euesden et al., 2014), which predicts phenotype likelihood based on an accumulation of SNPs with varying individual effect. PRSs were achieved by performing logistic regression analysis (described in detail in section 3.2.1) using SNPs sampled at various association thresholds, hence casting a wider net by including individually small-effect SNPs. Various studies proposed that SNPs with high PRSs were not randomly distributed among genes, rather they resided in or near genes whose functions were related. This hypothesis has been supported via pathway-based and gene set analyses for neurological disorders such as Parkinson’s disease (Holmans et al., 2012), bipolar disorder (Holmans et al., 2009) and schizophrenia (He et al., 2017). Pathway- and set-based analyses are methods that categorise SNPs and expressed genes into distinct biological processes or functions, therefore creating various groups containing affected genes of perceived similarity. For complex disease, these methods are important, since the emphasis lies on identifying interdependent variants that do not have stand-alone disease associations. The Multi-marker Analysis of GenoMic Annotation (MAGMA) tool (de Leeuw et al., 2015) is perhaps one of the strongest examples of analysing aggregated GWAS data to predict genomic elements associated with disease, and has been found to perform well in identifying additional influencers of Crohn’s disease.

A fundamental flaw that is common with these methods, however, is the initial process of pairing SNPs with the correct target genes. Typically, SNPs were paired with the nearest gene on a one-dimensional (1D) chromosome, despite approximately 86% of SNPs targeting genes outside of their nearest 1D neighbour (Mumbach et al., 2017). This gives way to a high risk of incorrect targets, which could be alleviated by incorporating data that captures 3D proximities of the whole human genome in order to correctly identify target genes.

We are now entering a new era of investigation of complex disorders such as schizophrenia. The recent emergence of putative schizophrenia markers and the knowledge that schizophrenia is a disorder of the developing brain (Hannon et al., 2016) has given way to techniques that quantify genomic fragment proximities in young brain tissues (Won et al., 2016). This allows for small-effect SNPs identified by GWAS and groups with high polygenic risk scores to be investigated in a tissue-specific context and as a community of interactions, rather than individually, since schizophrenia is far from monogenic in nature.

1.5. Biological Networks

We have arrived at a point in research where our knowledge of biological components is good, but our understanding of how they communicate in systems is still lacking. Traditionally, we have interpreted systems in a linear and chronological sense, like a chain reaction or domino effect. Nowadays, we understand that complex biological systems communicate in all directions and often in parallel (Buchanan, 2010). In the following text, we describe prevalent examples of networks in biology that aim to elucidate and quantify these communications within biological processes.

1.5.1. Cellular Networks

Gene regulatory networks are a perfect example of what were once thought to flow in a single direction. Our initial understanding was that genes coded for proteins, and proteins carried out a particular function. We have since discovered that genes are regulated by proteins called *transcription factors* (TFs), which bind to promoters that control the transcription of other genes. Clearly, the information flow described here is from protein to gene, therefore the previously-thought chronological process is shown to also communicate in more than one direction. The biological components in gene regulatory networks are classified as either sequence information or proteins, where sequences can include genes, promoters, enhancers, silencers or RNA – all of which influence gene expression. The connections in these networks are representative of the passing of information, ultimately leading to the expression of a gene. In such networks, one must also account for activity within the cell. That is, cells behave differently depending on their cell cycle stage, therefore these networks are time-dependent; information flows that exist at one moment may not necessarily exist at an earlier/later stage.

Since proteins rarely work in isolation (Robinson et al., 2007), it is important to understand how they can interact to form larger structures and affect the speed or nature of a biological process. One can therefore investigate proteins independently from genes or promoters, in what are called *protein-protein interaction networks* (PPIs; Phizicky & Fields, 1995). The connection of proteins is determined by whether they physically or logically interact. Physical interaction constitutes a spatial neighbourhood between single proteins that may result in their merging into a larger structure, whereas logical interactions account for protein complexes. That is, connections exist between proteins that are a part of larger interacting structures, despite them not being individually bound together (Gavin et al., 2002).

Metabolic networks describe reactions caused by specific proteins called enzymes (Jeong et al., 2000). The components of a metabolic network are chemical compounds present in the cell, and typically one-way connections describe a reaction that leads to a new compound. In many cases, the chemical reactions produce more than one output – connections can be assigned weights corresponding to the physical quantity of each output, thus distinguishing between the efficacies of various chemical reactions.

1.5.2. Disease Networks

The relationship between hazardous genotypes and phenotypes (collectively known as the *diseasome*) is often described using *human disease networks* (HDN; Goh et al., 2007), where members of the diseasome connect if a gene has some functional association with the disease phenotype. Since many disorders are complex in nature, it follows that a single disease can be connected to many genes and conversely, a single gene may play a role in more than one disorder.

To this end, the *disease gene network* (DGN) is an adaptation of the HDN that connects genes that are associated with the same disorder (Goh et al., 2007). The DGN is flexible in the sense that the interpretation of 'disorder' can vary – one could connect genes associated with schizophrenia, for example, but one could also simply connect genes that are associated with any brain disorder. The benefit of these networks is that they can reveal clusters of genes that could consequently be investigated in other settings, such as the aforementioned PPIs, regulatory networks or even 3D interaction networks.

1.5.3. Neuronal Networks

Using *neuronal networks* to map both the functional and structural features of the brain has become more commonplace in recent literature, especially because of advances in technologies such as magnetic resonance imaging (MRI; Bullmore & Sporns, 2009). Ideally, a network would consist of individual neurons being connected by synapses, although this soon renders the network too complex for current technology if considering the whole human brain. Therefore, larger-scale networks partition the brain into regions, and connections form as a result of general synaptic activity, which sacrifices specificity for the benefit of computational efficiency.

The common denominator with all of the existing biological networks described above is the unique ability of networks to be able to represent communications, rather than observing biological units in isolation. This is exactly why we have chosen to utilise networks for our own analysis – we are interested in the long-range communications between sections of the human genome, as opposed to simply observing the features of DNA fragments independently.

1.6. Thesis Overview

This thesis employs mathematical, statistical and computational procedures to identify and catalogue novel genomic regions which have an association with particular diseases and genetic phenomena. Our search criteria for identifying sets of SNPs is primarily focused on the 3D architecture of the human genome; the premise is that one-dimensionally remote regions of the genome can be spatially close in a 3D setting, which can consequently affect mutation behaviour and therefore overall function.

Chapter 2 gives an in-depth description of our most fundamental resource, Hi-C data, and articulates the history of methods that led to this means of quantifying the proximities of all genomic regions. A timeline of these procedures is given; the efficacy of such experiments has a correlation with developments in computing and consequently sequencing power. As such, we find that the pool of up-to-date Hi-C procedures offer the most reliable means of analysing the folding patterns of the human genome. Emphasis is given to *interaction frequencies* and reported *normalisation techniques* implemented to avoid potential bias in any experimental procedures, so that clear and accurate conclusions are made.

In Chapter 3, we introduce the mathematical techniques that are applied to our Hi-C data that aid in the identification of important genomic regions. Firstly, we define what a *network* is and provide a number of examples of the different types of network models that have been used to describe biological data to date. We then go on to describe various *network measures*, focussing on those that are useful for identifying the most important/influential genomic regions. We then consider the implementation of two different network studies: a *top-down* (global) and *bottom-up* (targeted) approach, where regions that are detected by both approaches are considered particularly significant, thus highlighting the power of such a two-pronged attack. The second section of this chapter describes the *statistical* analysis undertaken. *Hypothesis testing* and *predictive modelling* are the two predominant methods that we use – both of which have *p-values* at their core. As well as describing each statistical method, we also emphasise the importance of the p-values found, show how they can be used to interpret results and consider how to overcome multiple testing problems common in this area. Finally, we outline the implementation of these methods with respect to our aims and hypotheses – the results of which are discussed in subsequent chapters.

In Chapter 4 we observe the first disease-causing genetic phenomenon of our investigations – *reduced (or low/incomplete) penetrance*. A detailed definition is given first, followed by a description of how we are able to quantify levels of penetrance within populations. We give an outline of our aims and hypotheses for the proceeding analysis, and continue on to explain our motivations for choosing such hypotheses. A comprehensive database of genes known to influence clinical penetrance is introduced, which forms the basis for targeted analysis. For each hypothesis, we describe our intended approach in an algorithmic manner. We end this chapter by summarising our results and discussing their implications, with the end goal of addressing whether we have improved the knowledge of penetrance by performing our 3D analysis.

The second disease-causing genetic phenomenon that we investigate – *gene fusion events* – is the focus of Chapter 5. We first describe a gene fusion by showing the three possible processes in which one can occur: a *translocation*, *deletion* or *inversion*. Descriptions of hypotheses as well as our motivations are once again outlined, followed by the introduction of a database, *ChimerDB*, which catalogues known fusion events. We then present our methods via tailored algorithms and finish with a summary and discussion of results obtained from the execution of these algorithms. Our discussion includes a critique of our implemented methods as well as a conversation about the implications of such results. Most importantly, we discuss whether our findings show an association between the 3D structure of the human genome and the incidence of gene fusion events.

Chapter 6 reports our investigation of the well-known brain disorder *schizophrenia*. We first give an overview of the disease, including some key statistics regarding incidence and cause. We follow this with an introduction of *single nucleotide polymorphisms (SNPs)* and explain their role and importance in schizophrenia development. Alongside the stating of our hypotheses, we also introduce and describe the various tools which aid us in our global and targeted approaches, such as *DAVID*, *Braineac* and *FANTOM*. Whilst Hi-C data is now available for brain tissues, at the time of study we only had access to Hi-C for blood cell lines. Hence, we justify the use of blood Hi-C in this chapter by showing that there is an association between the data we used and brain-specific data. Various data sources, including schizophrenia-associated genes and SNPs found from *genome-wide association studies (GWAS)*, are described. These datasets form the basis for our analysis, allowing us first to focus only on regions of the genome relevant to the disease in question. We then outline the design of our algorithms and finish by describing our findings. Here we apply network techniques to study Hi-C and schizophrenia data. Beginning with the global approach outlined earlier, we attempt to find novel regions with schizophrenia association. We then use a range of targeted approaches to provide supporting evidence which implicate these regions in schizophrenia development. That is, an accumulation of positive results for a given region indicates strength of association with the disease. We conclude this chapter by highlighting the novel candidate regions most likely to be associated with schizophrenia, and suggest possible amendments to our analysis given the emergence of cell-specific Hi-C data.

Chapter 7 knits all of our analysis in preceding chapters together; our results and the efficacy of our methods is summarised, which leads to a discussion regarding routes for future work.

Chapter II

Hi-C Data Description

This thesis centres on the use of Hi-C data and its application in both network and statistical approaches. In what follows we shall present a chronological history of molecular techniques that serve to quantify the three-dimensional architecture of the human genome. We follow this by reviewing existing types of Hi-C protocol in order to ascertain strengths/weaknesses of each procedure. Finally, we investigate methods that aim to remove pre-existing bias through matrix normalisation techniques.

2.1. A History of Chromosome Conformation Capture Techniques

The molecular techniques in this chapter all probe the spatial organisation of chromatin, which is a complex of DNA, RNA and proteins found in the nucleus of eukaryotic cells. Human DNA of a single cell totals an approximate length of two metres (reviewed by de Wit & de Laat, 2012), and is therefore densely packed into structures called chromatin. We can imagine DNA as a piece of string being wrapped around biological spools, which in this case are the alkaline proteins called histones. Considering the appropriate resolution, this can resemble a “beads on a string” structure. Finally, at a lower resolution and with the addition of so-called scaffold proteins, we see DNA packed into a familiar “two-pronged” shape that we recognise as a chromosome.

Understanding the folding mechanisms of chromatin and hence the overall structure of chromosomes is the main purpose for developing chromosome conformation capture techniques. The resolution of these molecular approaches far outweighs what is possible with microscopic techniques (50-100 nanometres (nm); reviewed by Barutcu et al., 2015), which is why we see a rapid development of protocols described in this section.

2.1.1. Chromosome Conformation Capture (3C)

The chromosome conformation capture (3C) technique is known as a one-to-one method, since the protocol measures the interaction frequency of two specified regions of the genome. The pair of fragments can be either inter- or intra-chromosomal and the interaction frequency between them correlates to their spatial closeness in 3D space, achieved by capturing a population average proximity of the locus pair (Dekker, 2006).

The process for all chromosome conformation capture techniques is initially similar in nature: DNA is crosslinked at the location of interest, the chromatin is then digested with a particular restriction enzyme, ligation then occurs to “tie up” the loose ends in dilute conditions and DNA is then purified. The final stage for 3C is to detect ligated products using polymerase chain reaction (PCR), which is a method that "photocopies" small DNA fragments in order to reveal the interaction frequency of pairs from sequence information.

The choice of restriction enzyme is important when designing a 3C experiment. A large-base cutter will digest a fragment of DNA less frequently, resulting in a lower resolution because of the larger distance between cuts. A small-base cutter gives improved resolution at the expense of increased complexity (figure 2.1). The essence of choosing the appropriate restriction enzyme lies in the hypothesis that one may wish to test. That is, if the interactions are expected to be over several kilobases (kb), a 6-base cutter would suffice (digesting DNA approximately once every 4 kb) whereas if the experimenter was interested in a fine mapping of a particular element over a few hundred base pairs (bp), a 4-base cutter would be more suitable (digestion occurs approximately once every 256 bp; reviewed in Barutcu et al., 2016).

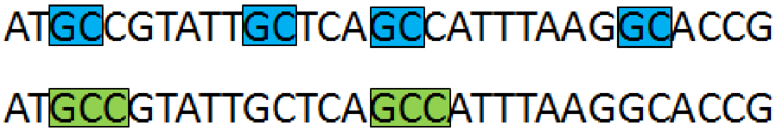

ATGCCGTATTGCTCAGCCATTTAAGGCACCG
ATGCCGTATTGCTCAGCCATTTAAGGCACCG

Figure 2.1 Visual representation of 2- and 3-base cutters. A 2-base cutter (GC, blue) digests DNA more frequently than a 3-base cutter (GCC, green), resulting in smaller DNA fragments between cuts and therefore a higher and more complex resolution. The larger the restriction enzyme, the lower the 3C resolution.

With enough 3C experiments, an all-to-all library of interaction frequencies is possible, but this would prove to be inefficient in this case. For this reason, it is accepted that 3C serves as a technique to use when some prior information is already known about the region/s in question, since the most efficient way to utilise this protocol would be to specify one region (or very few regions) of interest to investigate.

It is also advised to be mindful when considering the limitations of the 3C procedure. For example, the quantified spatial proximity between a given pair of regions from 3C does not specify which type of long-range contact is made. Therefore, the contact of chromosomes could be any or all of the following types: paternal-maternal, paternal-paternal or maternal-maternal. Furthermore, 3C may be able to give an indicator of proximity, but it lacks the ability to infer the functional relevance of such regions. Finally, 3C is reliable with contacts within a range of approximately 1 megabase (Mb), but a larger distance affects the accuracy of detection. One should therefore be cautious before interpreting results from a 3C experiment – the use of additional techniques to validate hypotheses are recommended.

2.1.2. 4C Methods

The key to a successful 3C experiment is having prior knowledge about the regions of interest before commencing the one-to-one method. The next natural step for 3C technologies is therefore to suppose that the interacting regions are not explicitly known. For example, perhaps we have a gene that is known to contribute to a particular disease, but we are aware that the expression of this gene is regulated by a region elsewhere on the genome. The notion of one-to-all approaches is therefore necessary to interrogate the entire genome for interactions with a specified region.

A number of one-to-all approaches were independently developed, all termed as 4C methods: 3C on chromatin immunoprecipitation (ChIP) (Simonis et al., 2006), circular 3C (Zhao et al., 2006), open-ended 3C (Wurtele & Chartrand, 2006) and olfactory receptor 3C (Lomvardas et al., 2006). The initial steps of crosslinking, digestion, ligation and purification are shared, albeit with particular differences between each protocol. For example, the most popular method, 3C on ChIP (which we will now exclusively term as 4C), performs two digestion and ligation events in order to trim initially large molecules into DNA fragments suitable for PCR (van de Werken et al., 2012). The defining difference between 4C and 3C techniques is the application of "bait" to the region of interest, which in turn results in a pool of interacting regions both *in cis* (inter-chromosomal) or *in trans* (intra-chromosomal) (reviewed in Sati & Cavalli et al., 2017).

Whilst 4C is a powerful tool to probe genome-wide interactions for a given region, one should pay close attention to the results of a 4C experiment. One-dimensional neighbours of the baited region should always be represented in 4C libraries with high interaction frequencies relative to typically long-range regions. If this is not the case, the 4C experiment could be the victim of poor crosslinking, which leads to a high false positive rate from frequent random ligation events (reviewed in Barutcu et al., 2015).

2.1.3. Chromosome Conformation Capture Carbon Copy (5C)

Prior knowledge of particular fragments of DNA are required for the one-to-one 3C technique, as is the case for the chromosome conformation capture carbon copy (5C) method (Dostie et al., 2006). The difference, however, is the size of the predefined region/s of interest. The 3C method requires a pair of single fragments to be queried, whereas the 5C method can probe a group of neighbouring regions, thus converting a one-to-one approach into a many-to-many approach.

Initially, the 3C steps of cross-linking, digestion and ligation are performed, followed by the introduction of 5C primers. Hundreds of primers are used as a means of simultaneously recombining DNA back into double-helix form at ligation junctions from the original 3C library. Once ligated, fragments are amplified using PCR, and genomic location is detected with the aid of deep sequencing (Ferraiuolo et al., 2012).

This technique seems favourable in comparison to 3C experiments not just because of the number of fragments being interrogated, but rather the biological implication this has. Transcription often occurs as a result of the interactions of a network of neighbours. The neighbourhood of a gene commonly includes interacting enhancers, silencers or promoters which are likely to be separated by a region of length ranging from a few base pairs up to 1 Mb in distance. Therefore, a many-to-many approach encapsulates all of these important genetic elements.

Due to its similarities to the 3C method, the drawbacks of the 5C design are also shared. In particular, the detection range is limited to approximately 1 Mb, hence, the exposure of remote interactions is restricted. Moreover, the 5C technique is not genome-wide, despite its ability to handle a sample of fragments. This implies that one cannot simply go into a 5C experiment blindfolded – some prior knowledge is required in order to receive effective results.

2.1.4. Hi-C

We finally arrive at Hi-C – a method that comprehensively maps intra- and inter-chromosomal interactions of the entire human genome simultaneously (Lieberman-Aiden et al., 2009; Rao et al., 2014). The general method follows the 3C protocol, except changes are made at the point of ligation. Digested ends are filled with biotin prior to ligation and unbiased interactions are detected post-ligation with the aid of streptavidin – a protein with a high non-covalent affinity to biotin (Green, 1975).

This all-to-all approach allows for a global view of chromosomal folding, which benefits studies that may not have prior knowledge about the region/s of interest for the particular genetic feature or disease in question. Of all 3C-based methods described, Hi-C is a tool which is perhaps the most suited to find novel regions associated with a given disease or feature because of its genome-wide detection window. Furthermore, there exists a range of resolutions available for this type of data, thanks to the use of several restriction enzymes.

Interaction frequencies from Hi-C techniques are obtained from millions of cells, and are therefore resultants of a so-called snapshot or population average of the genome. Thus, the data itself cannot be used to make informed decisions about 3D proximity within specific cells/tissues or even the persistence of certain long-range interactions over a full cell cycle. One must also be aware of the heterogeneous nature of fragment digestion throughout the genome. As discussed earlier, restriction enzymes target a specified sequence; therefore guarantees cannot be made about their whereabouts. The possibility of high resolution fragment digestion at fixed intervals is appealing but not currently viable, and so until this becomes plausible, we must understand the limitations of variable fragment lengths and choose binned region sizes appropriately.

The Hi-C method, amongst all preceding methods, is described graphically in figure 2.2.

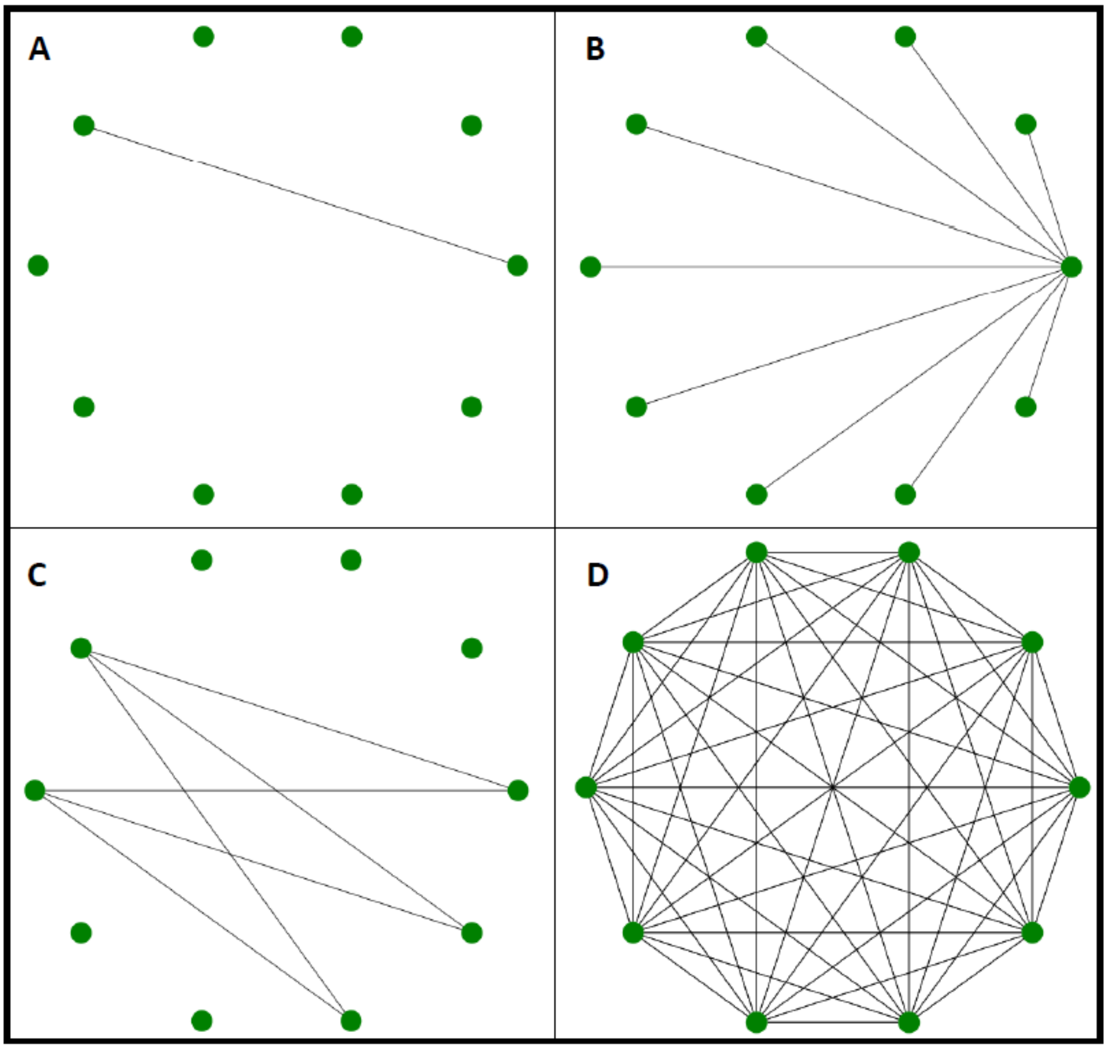


Figure 2.2 Graphical interpretation of chromosome conformation capture methods: (**A**) 3C method (one-to-one; Dekker, 2006), (**B**) 4C method (one-to-all; Simonis et al., 2006), (**C**) 5C method (many-to-many; Dostie et al., 2006), (**D**) Hi-C (all-to-all; Lieberman-Aiden et al., 2009). Green nodes represent regions of the genome, and edges connect them if interaction frequencies are obtained between said regions from the method in question.

2.2. Hi-C Methods

2.2.1. Dilution Hi-C

The first all-to-all 3C-based approach is now commonly referred to as dilution Hi-C (Lieberman-Aiden et al., 2009) due to the ligation step of the method being performed under highly diluted conditions. The rationale for this is to reduce the frequency of illegitimate ligation pairs from molecules that are not initially crosslinked. An overview of the dilution Hi-C experiment design is given below.

After crosslinking of DNA, genomic fragments are digested using 6-cutter restriction enzymes HindIII or NcoI, which cleave DNA at palindromic nucleotide sequences A^AGCT_T and C^CATG_G respectively (read from 5' end to 3' end; symbols "^" and "_" represent exact separation sites). These fragments are ligated in dilute conditions if they were crosslinked (hence spatially close), and their genomic positions are identified through alignment with the reference human genome, revealing the 1 Mb or 100 kb fragments (bins) in which the pair of ligation products lie. This constitutes a single *interaction frequency* between these bins. The process is repeated for all ligated pairs and the numbers of interactions are counted. Each of these binned regions is represented by a pair of co-ordinates (chromosome, bin number), where the bin number corresponds to a genomic range in which a particular fragment lies according to the hg18 reference genome (released in March 2006 and accessed using the Human Genome Browser; Kent et al., 2002). As an example, consider a ligated pair of fragments: fragment F_1 lies on chromosome 7 between positions 9,645,937 and 9,647,029 (inclusive) and fragment F_2 lies on chromosome 10 between positions 3,019,234 and 3,024,938. Assuming that the for the 1 Mb binning process, fragments within positions 1-1,000,000 are labelled bin 1, fragments within positions 1,000,001-2,000,000 are labelled bin 2 (and so on), it follows that F_1 has genomic co-ordinates (7, 10) and F_2 has co-ordinates (10, 4). Similarly, for the 100 kb binning process, F_1 and F_2 have genomic co-ordinates (7, 97) and (10, 31), respectively.

The binning process is executed for DNA regions of size 1 Mb and 100 kb for inter- and intra-chromosomal interactions, and the resulting cumulative counts, or interaction frequencies (IFs), are collected. Each IF is an indicator of the spatial proximity between two given fragments of the genome. That is, a high IF corresponds to a pair of fragments being relatively close together in three-dimensional space, and so the IF can be seen as inversely proportional to distance. The ij^{th} entry of contact matrix $M'(\alpha, \beta)$ gives the IF between the i^{th} and j^{th} binned regions of chromosomes α and β, respectively. Assuming data is available for identical bin sizes of all α and β combinations, we can concatenate each chromosome-chromosome contact matrix into a genome-wide contact matrix, M, describing IFs of any pair of regions:

$$M = \begin{bmatrix} M'(1,1) & \cdots & M'(1,23) \\ \vdots & \ddots & \vdots \\ M'(23,1) & \cdots & M'(23,23) \end{bmatrix}.$$

Entries of M are denoted m_{ij}, with each binned region, regardless of chromosome, now having a unique index. Suppose $B = \{b_1, b_2, \ldots, b_{23}\}$ is the set of values that represents the number of bins in a given chromosome (e.g. $M'(1,1)$ is of size b_1-by-b_1, $M'(1,2)$ is of size b_1-by-b_2, etc.). Indices i and j of M can be easily recalculated from the original $M'(\alpha,\beta)$ as follows:

$$i = \sum_{k=0}^{\alpha-1} b_k + i', \qquad j = \sum_{k=0}^{\beta-1} b_k + j',$$

where i' and j' are the original indices of $M'(\alpha,\beta)$, and $b_0 = 0$.

For example, bins of matrix $M'(1,1)$ would be identical in M, although this changes after chromosome 1: bin 154 of $M'(2,2)$ ordinarily has an index of 154. However, for the concatenated matrix M, its unique index would be 154 greater than the highest index found in $M'(1,1)$. That is, $i = b_1 + 154$. Indices for subsequent chromosomes in matrix M follow the same rule.

Both chromosome and genome-wide matrices (sometimes referred to as libraries) exist for both karyotypically normal (GM06990 – female blood) and aberrant (K562 – female blood with chronic myeloid leukaemia (CML)) cell lines, obtained from a total of approximately 10 million individual cells (available at http://www.ncbi.nlm.nih.gov/geo/, accession number: GSE18199).

2.2.2. *In Situ* Hi-C

The *in situ* Hi-C protocol (Rao et al., 2014) follows a similar methodology to the dilution Hi-C procedure, but with some key advantages. Four-cutter restriction enzymes, MboI and DpnII, are used to digest the genome into small fragments, which consequently allows for a higher resolution of Hi-C library. We have intra-chromosomal bins ranging from 1 kb to 1 Mb for the *in situ* method, which achieves a resolution 100 times better than the smallest dilution Hi-C bin size. There is also inter-chromosomal data available at 100 kb resolution. Also, and perhaps most importantly, this new Hi-C method is not performed in dilute solution (*in vitro*), but rather in a true biological context (*in situ*). This results in a much-reduced false positive rate of ligation products, since the frequency of random ligation *in situ* is far lower than *in vitro* because of chromatin crosslinking being frozen in its natural environment (hence there is less unnatural intrusion in the protocol).

Although the *in situ* Hi-C experiments devised by Rao et al. span 8 unique human cell lines (available at http://www.ncbi.nlm.nih.gov/geo/, accession number: GSE63525), we chose to focus on the most comprehensive of these libraries (GM12878 – female blood cell line; between 2-5 million individual cells used) for three reasons. Firstly, this particular Hi-C library has the highest and most varied resolution of the 8 cell lines (ranging from 1 Mb to 1 kb bins). Secondly, the GM12878 cell line is most similar to the previous GM06990 cell line used in the dilution Hi-C experiments. This enables us to make more confident comparisons between results from both sets of libraries. Finally, the *in situ* Hi-C procedure has an improved signal to noise ratio in comparison to dilution Hi-C because of the relative lack of spurious ligations, as described above (reviewed in Sati & Cavalli, 2017).

It is also worth noting that the *in situ* Hi-C libraries align with the hg19 reference genome (released in February 2009 and accessed using the Human Genome Browser: http://genome.ucsc.edu/cgi-bin/hgGateway). Thus, when comparing results obtained from both dilution and *in situ* Hi-C procedures, we can use the Lift Genome Annotation program available at http://genome.ucsc.edu/cgi-bin/hgLiftOver on gene datasets in order to attain accurate genomic positions between both hg18 and hg19 reference genomes.

2.2.3. Capture Hi-C

The cost of Hi-C experiments is dependent on the number of paired ends analysed. Therefore, the larger the genome, the less likely it is to be able to map genome-wide interactions at a high resolution efficiently. The Capture Hi-C (CHi-C) assay (Mifsud et al., 2015) is designed in order to circumvent this issue. This approach combines the typical Hi-C approach with hybridisation techniques as a means of capturing predefined regions of the genome. One could see this technique as a hybrid of Hi-C and other 3C-based approaches, since this is a hypothesis-driven Hi-C design.

The long-range interactions of approximately 22,000 gene promoters are captured for two cell lines (GM12878 and CD34, both blood cell types; 30 million total cells analysed). For our analysis, we decided to exclusively use reads from the GM12878 cell line (available at http://www.ebi.ac.uk/arrayexpress/experiments/, accession number: E-MTAB-2323), since this sample is common in both CHi-C and *in situ* Hi-C protocols. Baits are used to target promoter regions in the genome and subsequently probe interacting fragments, which typically span a few thousand base pairs. With this in mind, it is clear that CHi-C provides an extremely high resolution of interacting sequences in comparison to the 100 kb and 1 Mb bins seen in dilution Hi-C procedures. However, due to the specific nature of CHi-C, we do not have complete genome-wide libraries at binned intervals, but rather a comprehensive list of bait regions and their interacting partners. Thus, CHi-C is likely to be used as a secondary tool for analysis after initial findings from all-to-all Hi-C investigations. Furthermore, both CHi-C and *in situ* Hi-C align with the hg19 reference genome, which allows for seamless analysis; the need to convert genomic positions is removed.

2.3. Normalisation of Hi-C Data

The Hi-C procedures described above are not experimentally flawless; there are two common biases in particular that exist with most Hi-C experiments. Firstly, it is possible that after crosslinking and digestion, the loose ends of a pair of genomic fragments may not ligate with one another. Ligation may occur with preference to a site which is outside of the intended pair (Yaffe & Tanay, 2011). This results in the presence of false positive reads: a higher than expected IF is observed, leading to the potential identification of regions being in a 3D neighbourhood when in actual fact, they may be remote. Secondly, the pool of digested fragments is not uniform in length, since restriction enzymes target a nucleotide sequence rather than a specific position. Hence, crosslinked pairs can have variable lengths which consequently affects ligation efficacy (Yaffe & Tanay, 2011). For example, the loose end of a long fragment may be too far away to ligate to the intended shorter fragment end, resulting in the absence of an interaction where there should be one. Without repeating the Hi-C experiment with a catalogue of different restriction enzymes, this bias is likely to be predominant in specific genomic regions, since the same nucleotide sequence is being queried for cleavage each time.

There are two ways in which these biases can be treated. Firstly, changes could be made to the experimental procedure, such as the choice of restriction enzymes. Secondly, a dry-lab solution would be to post-process the resulting data via matrix normalisation. Since we are not the creators of Hi-C data and hence cannot change experimental procedure, we introduce published techniques which serve to balance the contact matrices obtained from Hi-C experiments as a means of lessening the effects of experimental bias. These are outlined and described in the proceeding text.

2.3.1. Raw Data

Chromosomal fragments are cut at irregular intervals by restriction enzymes, resulting in the crosslinked fragments having variable lengths. With this method being executed for all regions of the genome, we are left with a pool of ligation pairs of different sizes, which are then sorted into an appropriate bin for our contact matrix, M . For example, a crosslinked pair of genomic fragments – locus i from chromosome 6, positions 2,387,049-2,388,067 and locus j , also from chromosome 6, positions 37,587,185-37,587,980 would have an interaction recorded in 100 kb contact matrix M corresponding to row 24 and column 376, corresponding to genomic regions 2,300,000-2,400,000 and 37,500,000-37,600,000, respectively. All other ligated pairs are indexed using the same procedure until we have a complete raw library of interactions between all loci. In cases where the digested fragment is present in two consecutive bins, the interaction frequency is divided evenly between both intervals. In rare cases where a fragment may span 3 consecutive bins, the central bin will be empty, with the surrounding bins sharing the recorded contact. Duplicate contacts, reads that correspond to unligated fragments, or pairs that do not uniquely align to the genome are excluded from the completed library.

2.3.2. Vanilla Coverage Normalisation

Vanilla coverage (VC) normalisation is used for contact matrices produced by both the dilution Hi-C procedure (Lieberman-Aiden et al., 2009) and the *in situ* Hi-C procedure (Rao et al., 2014); it is considered a single iteration version of the Sinkhorn-Knopp matrix balancing algorithm (Sinkhorn & Knopp, 1967). Each entry of a raw intra-chromosomal contact matrix is given a normalised value corresponding to the sum of the row and column it lies upon. The normalised value corresponding to the i th row of square matrix M is given by

$$r_i = \frac{1}{\sum_{p=1}^{n} m_{i\,p}},$$

where n is the total number of columns (or rows) in M. The denominator in this expression equates to the *degree* of region i, typically denoted k_i. Similarly, we have the normalised value corresponding to the j th column of matrix M given by

$$c_j = \frac{1}{\sum_{p=1}^{n} m_{p\,j}}.$$

Therefore, each entry of the normalised matrix M^* is denoted as

$$m_{i\,j}^{*} = r_i m_{i\,j} c_j = \frac{m_{i\,j}}{k_i k_j}.$$

The advantages of this normalisation technique are its simplicity, since it is computationally inexpensive (even for large $(n \times n)$ matrices), and also its ability to handle sparse data, which can cause problems for more complex normalisation algorithms.

Square Root Vanilla Coverage (SQRTVC)

The VC normalisation technique is prone to an over-adjustment of raw interactions; the scalar multipliers r_i and c_j have a large influence on entries of M^*. In order to reduce the effects of these multipliers, each normalisation vector is square-rooted such that

$$m_{ij}^* = r_i^{0.5} m_{ij} c_j^{0.5} = \frac{m_{ij}}{\sqrt{k_i k_j}}.$$

This variant of the VC procedure is known as the *square root vanilla coverage (SQRTVC)* normalisation – it produces normalised counts that share a similarity to much more complex and expensive balancing algorithms (Rao et al., 2014).

Genome-wide Vanilla Coverage (GWVC)

Ordinarily, the VC procedure performs normalisations per matrix. That is, each raw interaction matrix consists of all interactions between a pair of identical chromosomes. The *genome-wide vanilla coverage (GWVC)* method calculates vectors r and c after concatenating raw matrices of all 529 (23-by-23) chromosome pairs, whether inter- or intra-chromosomal.

This has the advantage of incorporating information for all interactions, so that normalisation can be a truer representation of the global folding behaviour of the genome rather than focusing on local activity.

Inter-chromosomal Vanilla Coverage (INTERVC)

The issue with incorporating both inter- and intra-chromosomal interactions into a matrix balancing procedure is that intra-chromosomal counts tend to dominate, since their raw interaction frequencies are on average much higher than their inter-chromosomal counterparts.

Hence, the *inter-chromosomal vanilla coverage (INTERVC)* normalisation procedure is created by concatenating all chromosome pairs (as seen for GWVC), but then setting all intra-chromosomal interaction frequencies to zero before calculating vectors r and c. Inter-chromosomal interactions are therefore no longer washed out by intra-chromosomal counts – this gives a global balancing procedure without the bias of one-dimensional proximities.

2.3.3. Knight and Ruiz Matrix Balancing

The *Knight & Ruiz (KR)* normalisation procedure (Knight & Ruiz, 2013) is used predominantly with *in situ* Hi-C data (Rao et al., 2014). Provided that a raw matrix is square and has no negative entries, the KR method will normalise such that M^* is doubly-stochastic (all rows/columns have an equal sum).

This was originally achieved by repeating the VC normalisation procedure on a matrix until convergence (Sinkhorn & Knopp, 1967), but has since been adapted by Knight & Ruiz with improved efficiency. Assuming D is the operator that converts a vector into a diagonal matrix and e is a vector of ones, the KR method on a square, symmetric and non-negative Hi-C matrix, $M = (m_{i\,j}) \in Z^{n \times n}$, is designed to take an initial guess for a balancing vector, x_o (usually a vector of ones), and execute iterations (similar to Newton's method) until M is doubly-stochastic:

$$x_{k+1} = D(Mx_k)^{-1}e.$$

After convergence, the final balancing vector, x, is used to create the doubly-stochastic normalised matrix, M^*:

$$M^* = D(x)MD(x).$$

Sparse matrices can be problematic with this method, although this can be circumnavigated by removing rows/columns with the highest sparsity.

Genome-wide Knight and Ruiz Normalisation (GWKR)

This procedure behaves identically to GWVC (seen in section 2.3.2), with the exception that the normalisation technique used is KR rather than VC.

Inter-chromosomal Knight and Ruiz Normalisation (INTERKR)

This procedure behaves identically to INTERVC (seen in section 2.3.2), with the exception that the normalisation technique used is KR rather than VC.

Chapter III

General Methods and Implementation

The analysis carried out in this thesis centres around two main approaches based on network theory and statistical testing. This chapter serves as a guide for understanding these approaches by giving a description of each measure or test, sometimes with example applications. We also briefly discuss the implementation of these approaches for our analysis, with supplementary explanations being given in the methods sections of subsequent chapters.

3.1. Network Theory

3.1.1. Types of Network

A *network* (or *graph*), G, is a pair of disjoint sets (V,E), where V is a non-empty finite set of elements called *nodes* (or vertices) and E is a finite set of distinct pairs of elements (v_i,v_j) of V called *edges* (Wilson, 1970; Newman, 2010). The decision to connect a pair of nodes with an edge depends upon the context of a network. Typically, nodes represent a set of objects (such as people, genes or websites) and edges connect them if there exists some meaningful relationship between a pair of such objects (such as functional similarity for a set of genes).

Often, we can display these as *simple* networks. A simple network has the property of having no *loops* or *multiple edges*. That is, a single node cannot have an edge connecting to itself (a loop) or a pair of nodes cannot share any more than one edge between them (multiple edges). Networks containing multiple edges are called *multigraphs* and those that contain both multiple edges and loops are called *pseudographs* (figure 3.1).

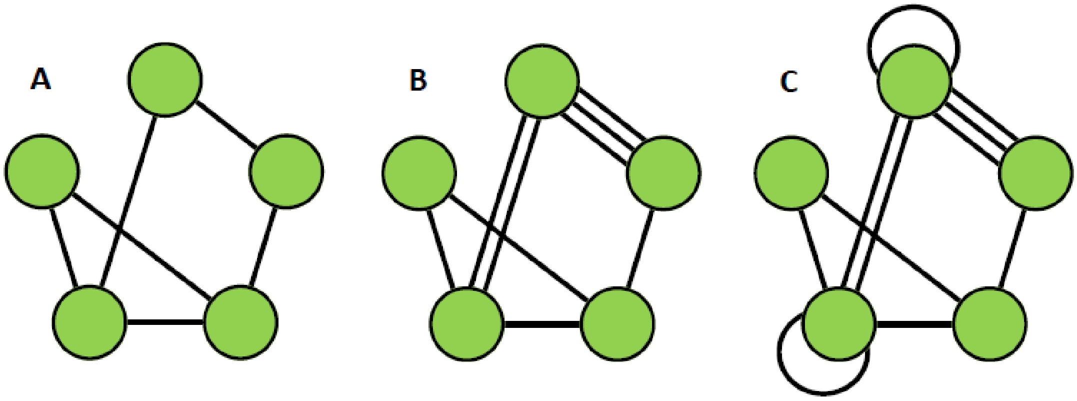


Figure 3.1 Types of network: (A) a simple graph, (B) a multigraph containing multiple edges and (C) a pseudograph containing multiple edges and loops.

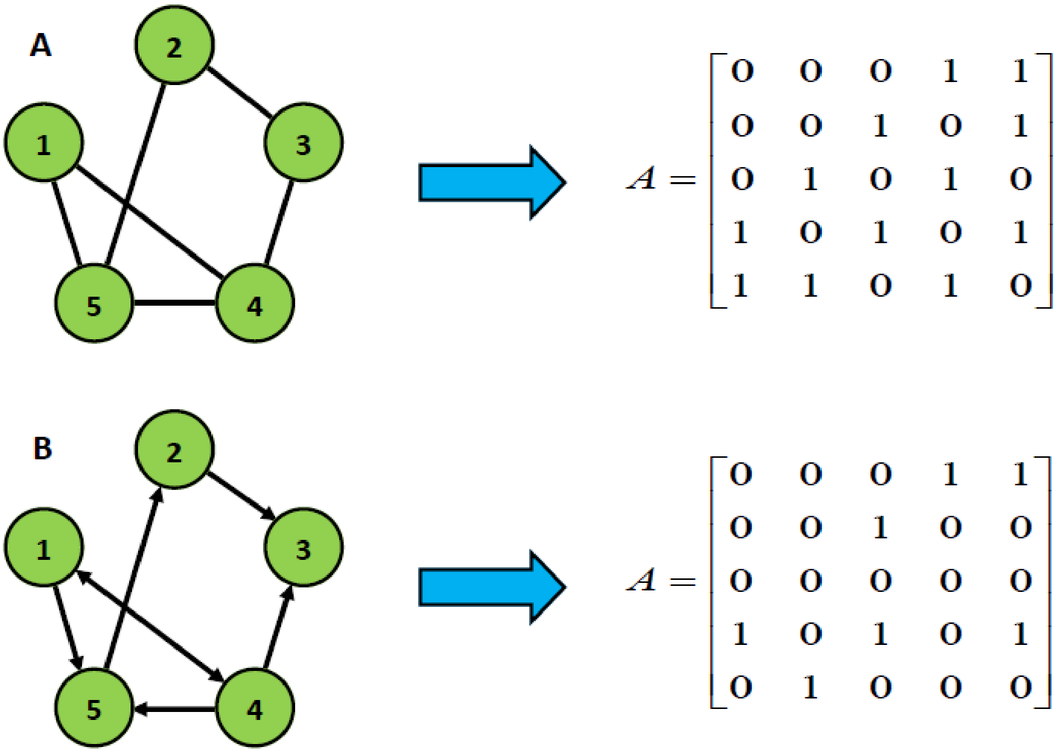
Whilst it is possible to display fairly small networks as a diagram of points joined by lines, as networks become larger or more complex, this means of display becomes unsuitable and does not offer a mathematically viable way of performing analysis. Hence, the *adjacency matrix* representation of a network offers a mathematical alternative by allowing us to employ a wealth of tools and techniques from linear algebra. A simple network G with n nodes $V = \{v_1, v_2, \ldots, v_n\}$ has an adjacency matrix, A, of size $n \times n$ whose ij th entry is either zeroes or ones depending on whether there exists an edge from set $E = \{E_1, E_2, \ldots, E_m\}$ that connects the vertex pair $\{v_i, v_j\}$. Nodes connected by an edge are *adjacent*, whereas edges are *incident* if they share the same node. Entries of A are given by

$$A_{ij} = \begin{cases} 1, & \text{if } v_i \sim v_j, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $v_i \sim v_j$ denotes a connection between two nodes.

Directed/Undirected Networks

A network may also have *undirected* edges. This means that a relationship between a pair of nodes is mutual and has no specified direction (figure 3.2A). For example, vehicle access between a pair of locations via road systems is undirected, although this property could be lost if the connecting roads were part of a one-way traffic system. In this context, our simple network changes to a simple *digraph*, where edges are *directed*. As a result of directed edges, our adjacency matrix, A, loses the property of being symmetric (where $A_{ij} = A_{ji}$), since a connection from node v_i to node v_j is not necessarily reciprocated (from v_j to v_i; figure 3.2B).



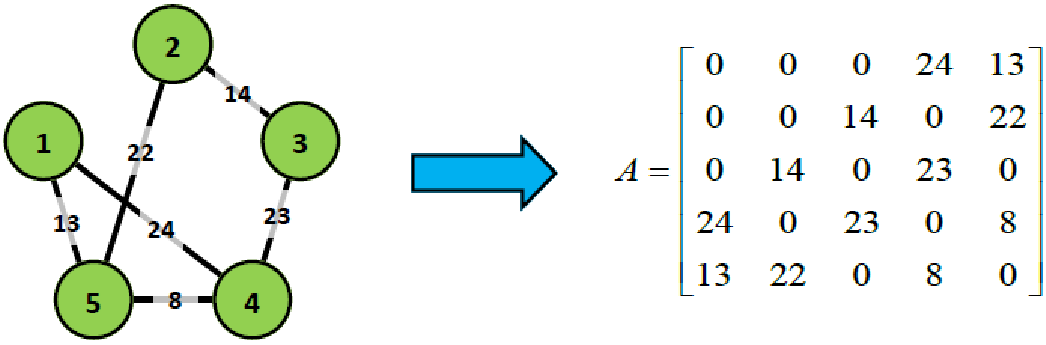
$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Figure 3.2 Visual representations and corresponding adjacency matrices of (A) an undirected network and (B) a directed network.

Binary/Weighted Networks

We initially defined an adjacency matrix as having entries of either zero or one – either an edge exists between a pair of nodes, or it does not. This is the *binary* case. Networks can, however, describe connections between nodes in a non-binary manner. It is possible to attach discrete or continuous values to each edge to give them *weighting*; connections may exist between particular node pairs, but some connections may have more importance or weight than others (figure 3.3).



$$A = \begin{bmatrix} 0 & 0 & 0 & 24 & 13 \\ 0 & 0 & 14 & 0 & 22 \\ 0 & 14 & 0 & 23 & 0 \\ 24 & 0 & 23 & 0 & 8 \\ 13 & 22 & 0 & 8 & 0 \end{bmatrix}$$

Figure 3.3 A weighted network with corresponding adjacency matrix.

A multigraph can be transformed into a discretely weighted, simple network by counting the numbers of shared edges between node pairs and assigning the frequency to a single edge, for example. If we consider the network of locations connected by road systems, a simple network with continuous weights can be created by calculating the physical distances and labelling each edge with this value. As a result, discrete or continuous weighted networks are conducive to attaining more detail by making distinctions between connected node pairs.

K-partite Networks

Nodes in a simple network can be partitioned into smaller groups according to which other nodes they share an edge with. Consider node set V being partitioned into two distinct subsets, $U = \{u_1, \ldots, u_n\}$ and $W = \{w_1, \ldots, w_m\}$, where $U \cap W = \emptyset$. That is, $u_i \sim w_j$ is possible, but $u_i \sim u_j$ and $w_i \sim w_j$ are not. These conditions satisfy what we call the *bipartite* property (figure 3.4A), which is the *k-partite* property at $k = 2$. Note that this rule can be extended for any choice of k corresponding to the number of distinct subsets (a tripartite example is shown in figure 3.4B).

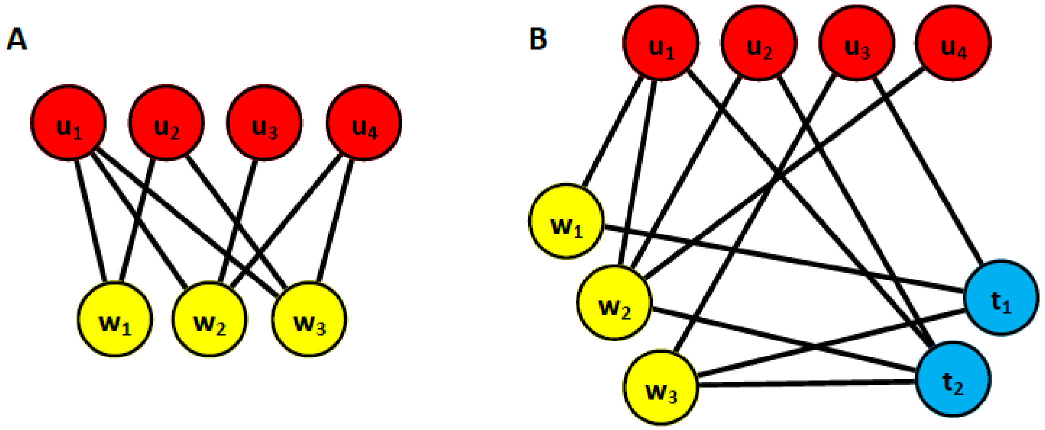


Figure 3.4 Examples of k-partite networks: (A) a bipartite network, showing two distinct node groups (red and yellow) connecting between one another but not within their own groups, and (B) a tripartite network, showing three distinct node groups (red, yellow and blue) with the same connective properties.

3.1.2. Network Measures

Recent literature has shown a sharp increase of the practical use of network theory thanks to its ability to model a diverse range of systems; nodes can be representative of a wide variety of discrete objects and defining the relationship that an edge symbolises is equally as flexible (Havlin et al., 2012). Various network measures act as tools to identify the most interesting and influential nodes in a network (Liu et al., 2016). In the proceeding section, we describe four of the most popular network measures that are used to make these identifications. We also determine why each measure has its own importance and explain how aggregating such measures can help to recognise influential nodes.

Degree Distribution

Any node in an undirected and unweighted network, has a *degree*, which is measured by counting the number of edges incident to v_i. In this case, the degree, k_i, can be found by summing the i th row of the adjacency matrix:

$$k_i = \sum_{j=1}^{n} A_{ij}.$$

We can consider the complete collection of node degrees as a *degree distribution* by plotting the probabilities of choosing a node with degree k:

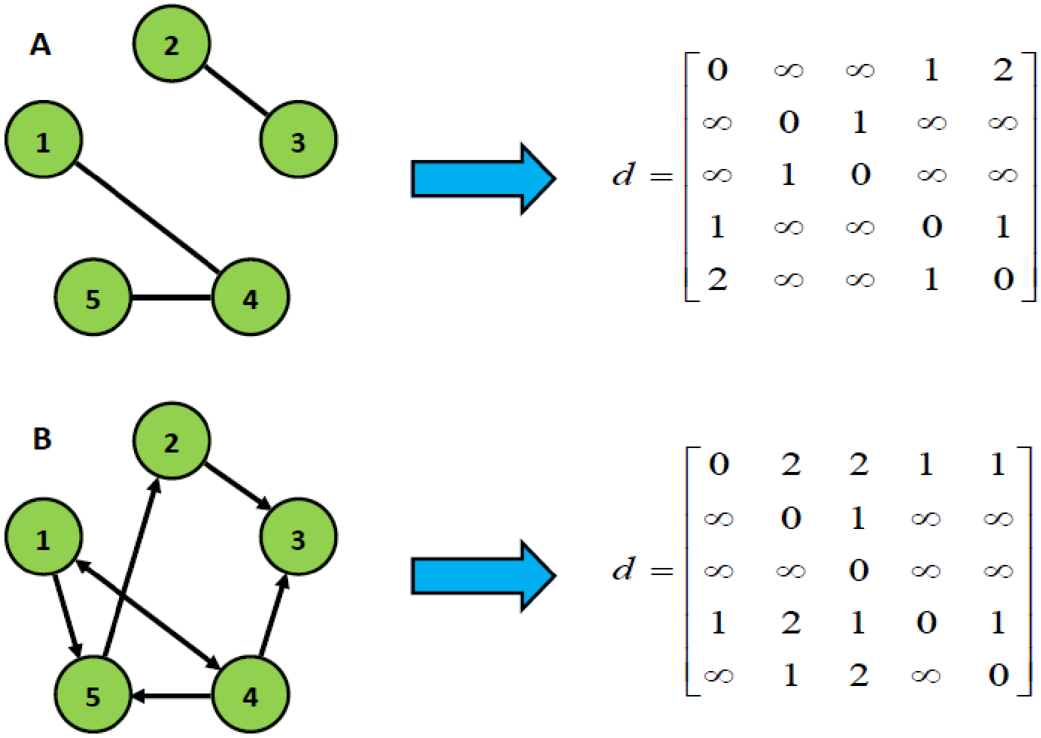
$$P(k) = \frac{n_k}{n}.$$

Note that n is the total number of nodes in the network and n_k is the number of nodes with degree k. Our definition of degree is modified if we consider digraphs; an edge directed from v_i to v_j implies that node v_i has an *out-degree*, and an edge directed from v_j to v_i implies that node v_i has an *in-degree*. Separate in- and out-degree distributions are computed in this case.

Path Length

A finite sequence of edges connecting node v_i to node v_j through a chain of distinct nodes is called a *path*. Although it is possible for many paths to exist that connect v_i to v_j, the path containing the fewest edges is known as the *shortest path* for unweighted networks. This is denoted by a distance measure, d_{ij}, which corresponds to the number of edges in the shortest path. For weighted networks, the calculation of d_{ij} is modified such that the sum of the edge weights is prioritised, rather than simply the number of edges. For example, if edges are weighted according to physical distance, such as in road network described earlier, two roads of length 6 and 7 miles would be shorter than a single 15-mile road. Hence, the shortest path for weighted networks does not necessarily have the fewest edges. It is also important to understand the context of weighted networks before rushing into calculations such as path lengths. A network that is weighted according to the strength of connections between pairs of nodes would behave in the opposite manner to networks where edges described distance. That is, the most important connections would have the highest weights. Therefore, we would calculate path lengths by summing the reciprocals of each weighted edge connecting v_i to v_j, so that the shortest path still represents the journey of minimum cost or maximum efficiency.

It is possible that no paths exist connecting particular node pairs. This can occur as a result of an unconnected network, where nodes are completely isolated from the rest of the network (figure 3.5A). Alternatively, this can be a consequence obtained from digraphs, where v_i connects to v_j in only one direction (figure 3.5B). In both cases, $d_{ij} = \infty$.



$$d = \begin{bmatrix} 0 & \infty & \infty & 1 & 2 \\ \infty & 0 & 1 & \infty & \infty \\ \infty & 1 & 0 & \infty & \infty \\ 1 & \infty & \infty & 0 & 1 \\ 2 & \infty & \infty & 1 & 0 \end{bmatrix}$$

$$d = \begin{bmatrix} 0 & 2 & 2 & 1 & 1 \\ \infty & 0 & 1 & \infty & \infty \\ \infty & \infty & 0 & \infty & \infty \\ 1 & 2 & 1 & 0 & 1 \\ \infty & 1 & 2 & \infty & 0 \end{bmatrix}$$

Figure 3.5 Visual representations and corresponding shortest path matrices for (A) a disconnected network and (B) a directed network.

We can also take a more global approach by considering the *average distance* of a node with respect to the connections made to all other nodes. This is achieved by calculating the mean of all shortest paths corresponding to v_i:

$$\ell = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}.$$

At this stage, this measure is still problematic for isolated nodes (i.e. node pairs can still satisfy $d_{ij} = \infty$). There are a number of ways to circumnavigate this issue: we could (1) omit these shortest paths to avoid divergence; (2) study only the largest component of the network (i.e. a sub-network where no nodes are isolated); (3) remove the directionality from the network. However, all of the above remedies would fundamentally change the properties of the network and/or would leave us with an incomplete representation of our population of nodes. Therefore, a more elegant solution known as the *global efficiency*, *GE*, is used (Latora & Marchiori, 2001):

$$GE = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}}.$$

This is still a measure that correlates with the average distance, but instead scales to boundaries of zero and one, where a higher efficiency (tending to one) corresponds to a shorter average distance and vice versa.

Whilst path length alone is not directly indicative of the general importance of a node, the calculation serves as an intermediary to a larger process that ultimately identifies global properties, such as the aforementioned global efficiency (which gives us an idea of a network's communicative ability) as well as some centrality calculations, described later in this section.

Clustering Coefficient

The *clustering coefficient* is a measure of how connected a node's neighbours are with each other in a network. A *complete* network has the property of all nodes connecting to one another – we are looking for near-completeness within sections of complex networks as a way to identify clusters. Clustering is, of course, relative. Although a large, complete network has the property we are looking for, there is no distinction between any nodes/sub-networks here because the network has no areas of sparseness. Hence, the *local clustering coefficient*, C, was introduced to measure the relative clustering of each node (Watts & Strogatz, 1998). It defines the clustering strength of a node by the ratio of triangles (complete networks with $n = 3$ nodes, C_3) that neighbour v_i to connected triples (incomplete, connected networks with $n = 3$ nodes; commonly called 2-paths, P_2) with centre v_i:

$$C(i) = \frac{C_3(i)}{P_2(i)}.$$

Note that P_2 can be calculated using information of a node's degree, such that

$$P_2(i) = \frac{k_i(k_i - 1)}{2},$$

and knowing that the adjacency matrix raised to the r^{th} power counts the number of routes between two nodes via exactly r edges (known as r*-walks*), one can find the number of triangles, C_3:

$$C_3(i) = \frac{1}{2} A_{ii}^3.$$

The total number of 3-walks from v_i to v_i is halved to account for walks travelling across the same triangle in opposite directions.

Hence, the local clustering coefficient of node v_i can be expressed as

$$C(i) = \frac{A_{ii}^3}{k_i(k_i - 1)}.$$

This satisfies $0 \le C \le 1$, where a high value of C corresponds to a node exhibiting clustering properties, and vice versa.

The *global clustering coefficient*, $\overline{C}$, is a measure attributed to a network as a whole, rather than each node. It acts as a way of giving context to local clustering coefficients by describing how *small-world* a network is, which is a measure of how connected a network is relative to its number of nodes, n. To be considered small-world, the average shortest path of a network typically satisfies (or falls below) $\ell \propto \log(n)$ (Watts & Strogatz, 1998). The global clustering coefficient is computed by taking an average of local coefficients:

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} C(i).$$

This measure is important for comparing local clustering coefficients between different networks. For example, two nodes from separate networks (i.e. $u_i \in G_1$ and $w_i \in G_2$) may have an identical local clustering coefficient, but node u_i carries a higher importance if $\overline{C}(G_1) < \overline{C}(G_2)$. A combination of local and global measures is therefore important in order to attain a reasonable understanding of a node's clustering status.

Connected Components

If any pair of nodes from a sub-network are connected by at least one path and do not connect to any node from elsewhere in the global network, the sub-network is known as a *connected component*. Although not an actual measure, observing the connected components of a large network enables the reduction of analysis complexity by focusing on smaller communities of well-connected nodes. Connected components are often used as a preliminary step before calculating network measures, so that isolated (and therefore trivial) nodes are rightfully omitted when considering which are most important, thus increasing computational efficiency.

Centrality

For complex networks, a common method to identify influential nodes is through *centrality* measurements. Generally speaking, this is a measure of the importance of nodes relative to the wider network (Estrada, 2012); a classic example of high centrality would be a node participating in a large number of shortest paths between node pairs. However, centrality is a diverse term which can be applied in a multitude of ways. We outline a handful of the most popular centrality measures in the proceeding text.

Degree centrality is a simple score, which counts the number of edges incident to a node (in other words its degree, k). Digraphs have both in- and out-degree measurements, as detailed previously. Whilst the degree can be an efficient means of measuring a node's importance, it is apparent that there is a flaw regarding its ignorance to the wider network. That is, degree centrality only takes into account nearest neighbours and ignores broader mechanisms, such as connective traits that extend beyond the first point of contact.

Eigenvector centrality is a measure that scores nodes based on their connections to other highly central nodes. For example, node v_i with degree $k=6$ can have a higher eigenvector centrality than node v_j with degree $k=6$ because v_i connects to other nodes which have a high centrality score, whereas v_j does not. This can be likened to the concept that someone can be considered influential based on another important person perceiving them as such (Kiss & Bichler, 2008). The eigenvector centrality, c_i, is defined as

$$c_i = x_i = \frac{1}{\lambda}\sum_{j=1}^{n} A_{ij} x_j ,$$

where x is the principal eigenvector corresponding to the maximum eigenvalue λ. This method is recursive in nature, as all nodes initially start with a score of one and iterations are performed until centrality scores converge.

PageRank centrality follows a similar process, although the output is a score based on the probability of arriving at a particular node during a random walk and therefore uses a modified Markov matrix (the *Google matrix*, G) instead of the adjacency matrix (Brin & Page, 1998). This special Markov matrix is created via three main steps. The first is to modify the adjacency matrix such that

$$H_{ij} = \begin{cases} {}^{1}\!/_{k_i}, & if\ v_i \sim v_j, \\ 0, & otherwise. \end{cases}$$

Now, the matrix H is almost row-stochastic, containing transition probabilities at each entry. The only place where H is not row-stochastic would be in cases of dead-ends (typically found for nodes with zero out-degree in a directed network; the row sum would be zero). Therefore, the second step replaces entries of the zero-sum rows of H, with a constant, $1/n$, thus creating a new matrix S. Every node now contains a transition probability to at least one other node, thus removing the possibility of encountering dead-ends. This also extends trivially for undirected networks (a dead-end node has zero in- and out-degree and would therefore be remedied in the same way). The third step accounts for a *damping factor*, δ, which is the probability that the walk will teleport to a random node in the network. This is a predefined input which is typically set to $\delta = 0.85$ for analysis of website networks (Page et al., 1999). The damping factor was introduced in order to guarantee the existence of an equilibrium solution for the Markov process; directed networks typically have weak connectivity and there is hence a motivation to avoid the possibility of not reaching certain areas of the network. The Google matrix accounting for the damping factor is therefore given by

$$G = \frac{1-\delta}{n} J_n + \delta S,$$

where J_n is the $n \times n$ matrix of ones and serves as an important term in ensuring all nodes of the network are accessible. The PageRank centrality of node v_i is hence obtained by observing the i^{th} entry of x^{r+1}:

$$x^{r+1} = x^r G,$$

which is the power method for computing the eigenvectors of G. Given that the maximum eigenvalue for a stochastic matrix is $\lambda = 1$, x is the corresponding left eigenvector – the approximation of which improves with increasing r.

So far, each centrality measure has been strongly associated with the degree of a node. *Betweenness centrality*, however, has less emphasis on this property and more emphasis on how well it connects other nodes. Consider a social network, containing work colleagues and sports teammates of subject A. Assuming that these two circles do not communicate, subject A would act as a node with high betweenness centrality – without this person, the groups would be isolated, hence this node carries high influence. Consider also that healthy brain networks, describing both structure and function, rely heavily on this type of interconnectedness. It was found that these networks become less connected in post-stroke patients, thus emphasising the communicative importance of remaining nodes, measured using betweenness centrality (Li et al., 2014).

Mathematically, the betweenness centrality score, b_i, correlates with the number of times it is a participant in the shortest path between two other nodes (Freeman, 1977):

$$b_i = \sum_{h \neq i \neq j} \frac{\sigma_{hj}(i)}{\sigma_{hj}},$$

where σ_{hj} is the number of shortest paths between nodes v_h and v_j, and $\sigma_{hj}(i)$ is the number of those paths which pass through node v_i.

3.1.3. Comparison of Network Measures

Particularly with the choice of centrality measures, we have looked to previous publications (described below) that compare measures in terms of performance and sensitivity to change, to inform us of the best way to identify important nodes.

We settled on four centrality measures: betweenness, degree, eigenvector and PageRank, based on various findings. Comparing three centrality measures in social task-based networks, it was found that betweenness centrality was the most accurate indicator of the perceived leader (Freeman et al., 1979). This result was supported by further research into larger social networks: both betweenness and degree-based centrality measures displayed a high sensitivity to random variation in network structure, hence capturing small changes well (Bolland, 1988). Betweenness centrality, however, seemed to perform less effectively when networks were incomplete or sampled. Eigenvector-based methods handled this problem better, displaying a better correlation between full and sampled networks than other measures (Costenbader & Valente, 2003).

The four selected measures represent a good diversity in terms of their method of identifying important nodes. For example, betweenness correlates the least with other measures because high ranking nodes do not necessarily have high degree. One could argue that the eigenvector and PageRank methods are so well correlated, that only one is required. However, the damping factor of PageRank provides an opportunity for variation between these measures. Whilst the application of the damping factor is typically explained with a web surfer’s chance of resetting to a random webpage, this property is perhaps conceivable for some biological networks and therefore important to include for our analysis.

3.2. Statistical Approaches

The approaches described in this chapter follow a chronology in our analysis. Generally speaking, network measures are employed first and interesting nodes are then interrogated further using statistical tests and methods. This section gives an overview of such methods. Particular emphasis is given to statistical p-values from our results; it is therefore important to understand how they are calculated so that the correct interpretations can be made.

3.2.1. Logistic Regression

Logistic regression is a modification of linear regression. Both types of model share common properties, such as predictors being continuous and/or categorical, but the key difference is that the response for logistic regression is transformed into a binary decision, i.e. success or failure (reviewed in Bush & Moore, 2012), whereas the response variable for linear regression is continuous. Logistic regression therefore enables us to distinctly categorise cases based on a whole range of characteristics, which is especially useful in biology with factors affecting the presence/absence of disease, for example.

The general logistic regression model is given by

$$y_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \underline{x}_i^T \underline{\beta},$$

where the response, known as the *logit* function, describes the log-odds of success (i.e. for random binary variable X, $P(X=1)=\pi$ and $P(X=0)=1-\pi$), $\underline{x}_i^T$ is a vector of measurements corresponding to predictors and $\underline{\beta}$ is a vector of coefficients (Dobson & Barnett, 2008). Note that in this form, logistic regression is versatile to any number of predictors.

Not only does logistic regression make it possible to categorise cases based on the predictive traits, the probability of obtaining success, π, can also be calculated by rearranging the logit function:

$$y_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) \Rightarrow \pi_i = \frac{\exp(y_i)}{1+\exp(y_i)}.$$

This is a powerful tool for diagnosis; given a patient's characteristics, one can predict the likelihood of developing a specified disease. These models are, however, at the mercy of sample size – the predictive power of logistic regression is proportional to the initial number of entries used to form the model.

3.2.2. Tests for Associations

The *chi-square* test investigates the level of dependence/independence between categorical variables. Contingency tables are used to count frequencies of events (observations), which are tested against expected frequencies.

Fisher's exact test is a modification of the chi-square test, which is applicable for 2-by-2 contingency tables and also small observation counts. Consider two categorical variables, A and B. The corresponding contingency table for observations of these variables is shown in table 3.1.

Table 3.1 Contingency table showing success/failure of categorical variables A and B.

	Variable A success	**Variable A failure**	Total
Variable B success	a	b	$a+b$
Variable B failure	c	d	$c+d$
Total	$a+c$	$b+d$	n

The probability of obtaining the observed set of frequencies under the assumption that the proportions between A and B are equal is given by a p-value. This is an indicator for the level of association between A and B. For a one-tailed test, a small p-value corresponds to a high success count for both A and B (frequency a) compared to other observed frequencies – this in turn causes a difference in proportions. The calculation for p is given by

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

If p is less than a level of statistical confidence (typically $p = 0.05$), we reject the null hypothesis in favour of variables A and B having an association.

3.2.3. Tests for Differences in Means

The *two-sample t-test* is a parametric test that investigates the difference of means between two independent, normally distributed variables. Considering two samples, X_1 and X_2 of size n_1 and n_2, respectively with equal variances, we arrive at the test statistic, t, using the following steps:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}},$$

and s^2 is the sample variance. An increase in the test statistic is proportional to a decrease in p-value, since the probability density function for the Student's t-distribution, given by:

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{t^2}{2}\right)^{-\frac{v+1}{2}}, \text{ where } \Gamma(x) = (x-1)!,$$

follows a similar shape to the normal distribution centred at zero (figure 3.6).

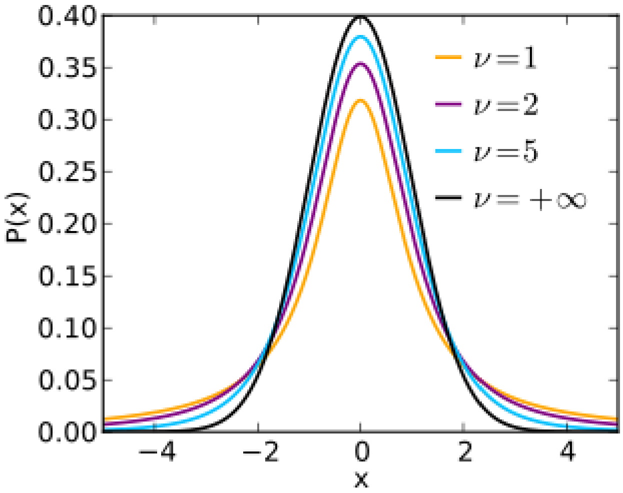


Figure 3.6 Probability density functions of Student's t-distribution at various degrees of freedom, ν.

A small p-value ($p < 0.05$ for 95% confidence) points to a rejection of the null hypothesis in favour of there being a significant difference in means between X_1 and X_2. In cases where multiple comparisons are made, one must make a Bonferroni correction. That is, if there are m comparisons made, the p-value threshold would change to $p = \frac{0.05}{m}$.

3.3. Implementation

The overarching aim of this thesis is to find novel long-range genomic interactions that influence disease or genetic phenomena. The various network measures and statistical approaches described above were our tools to help find such regions. This implementation section therefore serves to explain the preceding measures in the context of 3D genomic interactions and also highlights its relevance to our hypotheses. We also evaluate the computational hurdles that arose from our analysis and explain how we overcame them.

3.3.1. Hi-C Networks

The networks used in this thesis were all constructed from various Hi-C datasets. In all cases, nodes represented the collection of binned genomic regions and edges connected them if there existed at least one interaction between a pair of regions. Henceforth, we refer to such networks either as *Hi-C networks* or *3D interaction* *networks*. Without any further treatment, these networks were weighted. That is, the higher the weighted edge connecting two nodes, the smaller the 3D distance between the corresponding regions. In some cases, we used a threshold to convert a Hi-C network from weighted to unweighted; connections therefore constituted a pair of regions being sufficiently nearby within the nucleus, without distinguishing how close.

3.3.2. Network Features

It is hypothesised that in any of our created networks, well-connected nodes were representative of regions of the genome with most influence over other remote regions. Hence, our network measures were chosen specifically to identify such nodes. We utilised two computational toolboxes to calculate various network measures: the Brain Connectivity Toolbox (BCT) written in MATLAB (Rubinov & Sporns, 2010) and the Python-based NetworKit (Staudt et al., 2016). Both toolboxes contained the means to calculate all measures mentioned above, so we decided to run network analysis using each toolbox where possible and compare results to ensure they consistently agreed.

We further justify the choice of centrality measures by considering how they each align with our underlying hypothesis. The degree centrality, for example, directly measures the connectedness of a node. Hence for an inter-chromosomal Hi-C network, we could identify a node with high degree centrality (i.e. many inter-chromosomal connections) as a region that spends the majority of its time on the periphery of a folded chromosome. We could then use this information to explain how a previously found long-range interaction influences a disease, rather than simply knowing that it does. Furthermore, highly central nodes found using eigenvector/PageRank centrality measures could also indicate that particular loops preferentially locate at the surface of a compacted chromosome. Finally, nodes of high betweenness centrality would correspond to regions that connect otherwise disconnected components of the network. Polygenic diseases may not express the expected phenotype as a result of the absence of a node with high betweenness centrality, for example.

3.3.3. Computational Complexity

The measures we present provide varying degrees of computational complexity, which could be problematic given that we are working with large networks (tens of thousands of nodes and millions of edges). The calculation of degree centrality lies on the faster end of this scale, with computational complexity $\mathrm{O}(m + n\langle k\rangle)$, where m, n and $\langle k\rangle$ are the number of edges, the number of nodes and the average degree of the network, respectively (Liu et al., 2016). The opposite end of this scale is where we find measures such as betweenness, since the nature of the method is not well designed for scalability. For large networks, it is clear that the computational cost of calculating betweenness centrality (given by $\mathrm{O}(n^3)$; Liu et al., 2016) can quickly become problematic.

There are a number of ways that we have approached these problem cases for our analysis. We first suggested that we could pre-process our Hi-C networks by converting the corresponding adjacency matrices from weighted to binary (i.e. zeroes and ones), based on a chosen threshold. This would preserve the size of a network, whilst significantly reducing the file size of the adjacency matrix. It also speeds up calculations such as path lengths, since path length algorithms are much faster on unweighted networks. Our alternative suggestion was to randomly sample path length calculations (for betweenness centrality in particular) in order to obtain a de facto average path length. The former seems much more desirable than the latter, since we were able to investigate using all available data. We have also seen that some centrality measures do not perform accurately under sampling conditions (described in section 3.1.3), which is more reason to choose the thresholding method to improve computational efficiency. We were careful not to completely discount sampling methods, however, since a handful of measures still perform well under these conditions.

Chapter IV

Reduced Penetrance

It is known that an individual can harbour a particular disease mutation without actually exhibiting any of the disease traits. Such an individual may be termed a carrier, although biologically speaking, this is known as reduced (or incomplete) penetrance. To describe this phenomenon in basic quantifiable terms, let G be the frequency of individuals from a population with a hazardous genetic variant (disease genotype) and let P be the frequency of those individuals that express the ill effect of the variant (phenotype) at some point during their lifetime. It follows that reduced penetrance occurs when $G > P$. We measure the penetrance of disease by calculating $\frac{P}{G}$, where values can lie in the range $[0,1]$ (a value of 1 indicates complete penetrance). To summarise, disparity between G and P is an indicator for a level of penetrance, therefore genotype is not directly predictive of phenotype (Cooper et al., 2013).

Reduced penetrance poses somewhat of a conundrum when considering heritable disease; implying that DNA language codes for a subsequent trait is not completely true. A confounding example would be a parent and child carrying the same mutation, but only the child exhibiting the subsequent trait. We are effectively saying that reduced penetrance has synonymous properties with the type 2 error, or false negative. That is, a genetic variant exists that influences disease formation, but we do not witness the manifestation of this disease and we therefore fail to reject our null hypothesis that the variant is not associated with the disease. This interpretation is also distinctly binary; we assume that variants either do or do not influence a disease, without considering the possibility that disease may develop only when a group of variants exist within an individual. Consider a hypothetical disorder that develops as a result of three variants at separate loci. The absence of one of these variants would mean an absence of the disorder. Therefore, does this mean that the other two variants have zero effect? It is likely that variants are interdependent and investigating interactions between variants is extremely important to understand the true mechanisms of penetrance.

Perhaps there may also be concern about identifying whether a genotype has a significantly low penetrance or if it simply has no effect on the suggested disease. This notion is quashed by observing the genotypes and phenotypes of carriers' ascendants/descendants in what is called *cascade genetic screening* (Berge et al., 2008), thus providing hereditary evidence for this phenomenon. The idea of variable expression efficacy of genotype also supports this, whereby studying a collection of variants and their interactions is likely to be a key component in identifying causes of penetrance.

In complex disease, the presence of a single gene variant is not sufficient for a resulting phenotype, despite it coding for a function which clearly affects the disease. Quite often, we find that variants in *cis* or in *trans* are responsible for altering the function of the *primary gene*, whether that is by amplifying, silencing or even completely changing the gene's final function. Such cases are termed *modifier genes*. The mechanism of primary and modifier gene interactions affecting expression is known as epistasis; we argue that this phenomenon is a key component in explaining variable levels of penetrance for disease.

Interdependent genes in complex disease can also manifest themselves slightly differently. In contrast to the domino-effect nature of primary and modifier genes (which implies a chronology of cause and effect), gene pairs exists that influence *digenic* inheritance. This is where variants at two unlinked genes cooperate to consequently produce a phenotype. There is no order of one gene affecting another, and the absence of either variant results in a zero-penetrance of the other, hence an absence of expected phenotype.

In this chapter, we hypothesised that: (1) the level of penetrance can be partially determined by the occurrence of distal variants, such as modifier genes, coming into close contact with known primary disease genes via the folding of the genome in 3D space; (2) genes associated with reduced penetrance share common third-party interacting fragments; (3) primary-modifier and digenic reduced penetrance gene pairs are regulated by such third-party fragments, if not more. The latter in particular alludes to our earlier suggestion of interdependency between variants associated with low penetrance. We utilised Hi-C data describing 3D proximity of genomic fragments (Lieberman-Aiden et al., 2009; Rao et al., 2014) to investigate all three of our hypotheses.

4.1. Background

4.1.1. Factors Affecting Penetrance

Historically, reduced penetrance was thought to be a phenomenon associated only with autosomal dominant disorders, such as glaucoma (Morissette et al., 1998), retinitis pigmentosa (Saini et al., 2012) and long QT syndrome (Mathias et al., 2013). However, developments in molecular techniques have revealed the presence of disease-causing mutations in healthy phenotypes for autosomal recessive disorders, such as hemochromatosis (Beutler, 2003) and deafness (Fairley et al., 2008). This revelation has opened the floodgates for identifying factors affecting penetrance, since a much larger pool of genotype profiles now have reported associations. The various factors affecting penetrance are discussed below.

Mutation Type

Mutation type is one such factor. Whilst there are few studies which investigate specific mutation types and their effect on penetrance, we do have exposure to a multitude of studies which report on mutations affecting particular diseases such as missense mutations, nonsense mutations, insertions and deletions (described in figure 4.1).

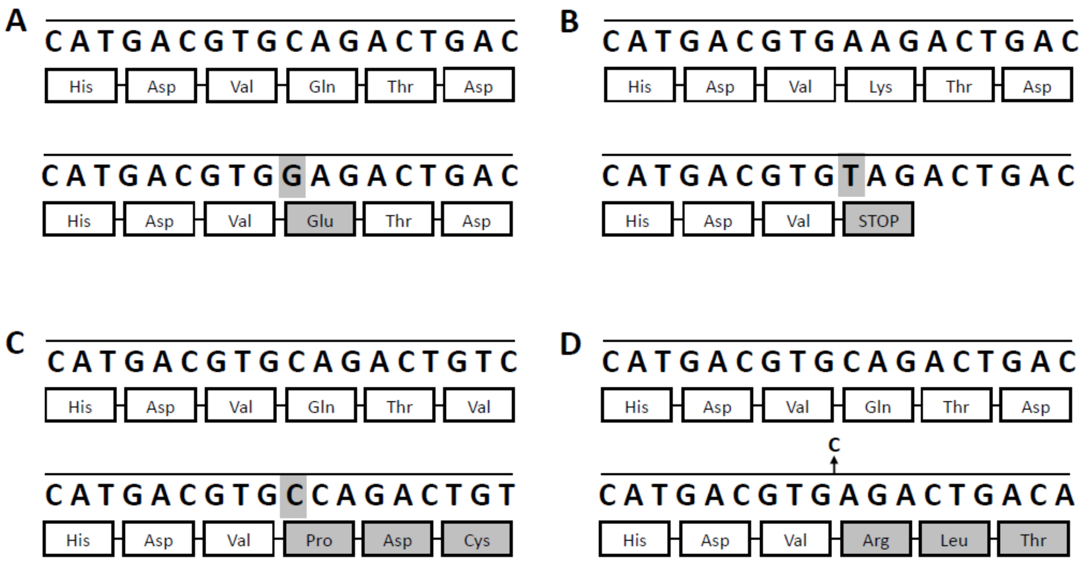


Figure 4.1 Pairs of ordinary and mutated sequences that illustrate types of mutation: (A) a missense mutation, caused by a single base change and resulting in an amino acid change; (B) a nonsense mutation, caused by a single base change and resulting in a premature signal to stop building a protein; (C) an insertion of one (or more) bases, which causes a frame shift, resulting in a change of the subsequent amino acid sequence; (D) a deletion of one (or more) bases, which causes a frame shift, resulting in a change of the subsequent amino acid sequence. Changes in DNA or amino acid sequence are highlighted in grey.

A well-known example of a highly penetrant mutation is the in-frame deletion in the *CFTR* gene (removal of CTT; rs113993960), causing cystic fibrosis (found in approximately 70% of cystic fibrosis patients; Kerem et al., 1989), whereas there exists a missense mutation in the same gene (change from G to A; rs78655421) which at best has a very mild clinical consequence (penetrance in the range .03-.06%; Thauvin-Robinet et al., 2009).

The type of mutation need not be different in order to observe differences in penetrance. Consider missense mutations in the *BRCA1* gene, causing a change of amino acid from arginine to glutamine. The 70-year penetrance of breast/ovarian cancer from this type of mutation is 24% (Spurdle et al., 2012) compared to the average penetrance for breast and ovarian cancer via *BRCA1* mutations being 71.4% and 58.9%, respectively (van der Kolk et al., 2010). This suggests that perhaps the effect of the mutation (resulting amino acid sequence), as opposed to the type of mutation (insertion, deletion, missense, nonsense, etc.) is a driving factor in penetrance.

Digenic Inheritance

The occurrence of digenic inheritance is dependent upon the interaction of two variant genes and a subsequent expression of phenotype (Shawky, 2014). Attempts have been made to catalogue cases of digenic inheritance and reduced penetrance (table A4.2; Cooper et al., 2013), although one must be wary of distinguishing between true digenic inheritance and a coinheritance of two genes that exacerbates a phenotype. Despite this, there are well-studied examples of digenic inheritance causing variable penetrance, such as the interactions of mutations in the *FGFR1* and *NELF* genes causing idiopathic hypogonadotropic hypogonadism (Pitteloud et al., 2007) and even three mutations in any two of four *BBS* genes (*BBS1, BBS2, BBS4* and *BBS6*) causing Bardet-Biedle syndrome (Schäffer, 2013).

Modifier Genes

It is possible for a phenotype to exist given the presence of mutation/s in a single gene. However, a second modifier gene may interact with the original primary gene which consequently affects the severity of the phenotype and ultimately, penetrance (Cooper et al., 2013). This has similarities with digenic inheritance, whereby an interaction of a pair of genes is necessary, although the key difference is that this mechanism has a chronology as well as dependence.

Cystic fibrosis is a typical example – predicting the phenotype using only information from a mutated *CFTR* gene is unreliable. There is reported to be a minimum of seven associated modifier genes, such as *TGFB1* (Drumm et al., 2005), that each adjusts phenotype expression and therefore penetrance (Badano & Katsanis, 2002).

Epigenetic Factors

The penetrance of a phenotype may also be controlled by modifications outside of changes in genetic code. Epigenetic factors, such as histone modifications and DNA methylation are able to activate or repress gene expression by altering the environment in which DNA resides (Dong & Weng, 2013; Kass et al., 1997).

Monozygotic twin studies are able to best explain differences in phenotype despite observed similarities in genotype (Wong et al., 2005). A higher methylation of *BRCA1* was observed in one patient affected by childhood leukaemia (12%) compared to her unaffected sister (3%), for example (Galetzka et al., 2012). Furthermore, methylation of *SLC6A4* correlated with a presence of bipolar disorder in another twin study (Sugawara et al., 2011). This was replicated in a case versus control setting by finding higher methylation in post-mortem brains of patients suffering from bipolar disorder.

Sex

Differences in penetrance for autosomal disorders between male and female patients can be observed in one of two ways. A disorder can be described as either *sex limited* or *sex influenced* (Shawky, 2014). A sex limited disorder can only occur in males or females; one gender has a 0% penetrance and the other displays at least some penetrance. Precocious puberty as a result of *LCGR* mutations is an example of this. Females are unaffected by the hazardous mutation, whereas males exhibit the phenotype (Coleman & Tsongalis, 2010). By contrast, a sex influenced disorder affects both males and females, but to varying degrees. As an example, duplications and deletions on chromosome 16, positions 14,800,000-16,800,000 are associated with a set of brain-related disorders including autism, epilepsy, attention deficit hyperactivity disorder and schizophrenia. However, there is a reported bias of penetrance towards men (Ramalingam et al., 2011; Tropeano et al., 2013), although a proportion females still experience the phenotype.

4.2. Materials

4.2.1. Dataset of Primary and Modifier Genes

In our studies, we utilised a dataset of 37 primary and modifier gene pairs, where variants in the modifier genes have been found to modulate the penetrance of a variety of inherited diseases (table A4.1; Cooper et al., 2013). These mutations have been found to be a cause for a range of diseases, such as breast cancer (Antoniou et al., 2007), cystic fibrosis (Drumm et al., 2005) and Parkinson’s disease (Gan-Or et al., 2011). There are 32 inter- and five intra-chromosomal gene pairs in this dataset.

4.2.2. Dataset of Genes Responsible for Digenic Inheritance

We have a comprehensive dataset of 133 unique gene pairs (split into 121 inter- and 12 intra-chromosomal pairs) that are known to contribute to disease via digenic mutations (table A4.2; Cooper et al., 2013). Gene pairs in table A4.2 are unambiguous cases; the digenic mutations are strictly from genes that have a functional association with the corresponding disease, thus reducing the chance of pairs being found by coincidence. Examples of diseases affected by digenic activity include chronic lung disease (Bullard & Nogee, 2007), colorectal cancer (Uhrhammer & Bignon, 2008; Li-Chang et al., 2013) and polycystic kidney disease (Pei et al., 2001; Dedoussis et al., 2008). There is also some overlap between this dataset and our list of primary/modifier genes with regards to the underlying disease, such as breast cancer (Pern et al., 2012) and Parkinson's disease (Dächsel et al., 2006).

4.3. Methods

4.3.1. Proximity of Gene Pairs

Our first step was to determine whether gene pairs associated with reduced penetrance had high interaction frequencies (IFs) in Hi-C data relative to randomly selected control regions. If this was indeed the case, we could confirm our hypothesis that reduced penetrance gene pairs are chosen non-randomly and the 3D architecture of the human genome is associated with this.

We defined a pair of genomic fragments on chromosome α, bin i and chromosome β, bin j to be relatively close in 3D space if their IF was higher than most interactions between chromosome α, bin i and other fragments. This led us to construct a method to rank the IFs of fragment pairs in descending order (figure 4.2) and compare them to a rank threshold, in order to categorise an interaction as being either high (thus being close in 3D space) or not high (implying remoteness).

R1.	**[Construct counter vector.]** Let f be the index of the genome-wide, $k \times k$ contact matrix, M, that corresponds to chromosome α, bin i. It follows that $M_f = [m_{f1}, m_{f2}, \ldots, m_{fk}]$, where m_{fl} is the interaction frequency between region f and any other region, l. We construct our counter vector, $v = [v_1, v_2, \ldots, v_n]$, where v_1 is a counter for the number of interactions (or entries of M_f) with frequency 1, v_2 with frequency 2, and so on. The last entry, v_n, corresponds to the maximum interaction frequency recorded for fragment f.
R2.	**[Find IF of pair.]** Find m_{fg}, where g is the index of M corresponding to chromosome β, bin j.
R3.	**[Find IF rank.]** If $c = m_{fg}$, then the interaction frequency rank, r, is found by counting the number of non-zero entries of v between v_c and v_n. If $c = 0$, then $r = n + 1$.

Figure 4.2 Algorithm R: pseudocode ranking fragment pairs according to their interaction frequency relative to other interacting fragments. A higher rank implies the tested pair of regions are relative neighbours in 3D space.

We set a number of rank thresholds in order to find an appropriate stringency; a small rank threshold could miss out potential neighbouring pairs, whereas a large threshold would cause a high false positive rate. That is, remote fragment pairs could be recorded as being 3D neighbours. After all interactions corresponding to gene pairs in our datasets were considered, we tabulated the frequencies of pairs that were 3D neighbours and those that were not. We also recorded the same information for controls (figure 4.3). The control algorithm was repeated 1,000 times and frequency entries were averaged, thus giving values on the continuous scale.

C1.	**[Retain primary fragment and secondary fragment chromosome.]** Given a test fragment pair (α, i) and (β, j), the corresponding control pair is given by (α^{*}, i^{*}) and (β^{*}, j^{*}), where $\alpha^{*} = \alpha$, $i^{*} = i$ and $\beta^{*} = \beta$.
C2.	**[Select random secondary fragment bin.]** Randomly generate a bin, j^{*}, within the bin range appropriate for chromosome β.
C3.	**[Is secondary fragment associated with reduced penetrance?]** If the control region (β^{*}, j^{*}) does not exist in the test dataset, algorithm terminates; (β^{*}, j^{*}) is our control region. If the region does exist in the dataset, repeat C2.

Figure 4.3 Algorithm C: pseudocode generating control fragment pairs to test against reduced penetrance gene pairs.

The motivation for retaining the primary fragment in our controls was to ensure we focused on the choice of partner. Randomising the primary fragment in this instance would not have been logical, as we would then not be looking at a choice in partner, rather a completely disassociated pair.

After tabulation of frequencies, we performed chi-square proportion tests to measure the statistical difference in the proportions of neighbouring regions between reduced penetrance pairs and controls. Small p-values ($p < 0.05$) would lead us to reject our null hypothesis, in favour of concluding that the proportion of reduced penetrance partner choices within highly interacting fragments is greater than for controls.

4.3.2. Identification of Penetrance Regulators

We hypothesised that there are common fragments in the human genome that remotely interact with genes associated with reduced penetrance. Our rationale was that there could be distinct genomic regions which inhibit the expression of indiscriminate phenotypes, rather than disease-specific variants affecting penetrance.

For each reduced penetrance gene forming an inter-chromosomal interaction, we found the corresponding top ten interacting fragments throughout genome-wide dilution Hi-C data (using 1 Mb bins). The list of top ten interacting fragments excluded intra-chromosomal interactions in order to avoid the bias introduced by chromosome territories. From the resulting lists, we identified particular interacting fragments which were dominant throughout and compiled a table of these most commonly featuring fragments using prevalence values. These values were measured as a proportion; the number of times a region appeared in the top ten list of interacting fragments was divided by the total number of genes interrogated. It is worth noting that our prevalence calculation would adjust the total number of genes interrogated for cases where a commonly interacting region was on the same chromosome as some genes from the dataset. To clarify, 100% prevalence can be achieved by an interacting fragment on chromosome α if it is found in every list of top ten interactions except for those corresponding to genes which are on the same chromosome. We also compiled a similar table for all bins throughout the genome to act as a control and to see whether our prevalent fragments are in fact over- or under-represented in comparison.

In order to statistically test the prevalence proportions between cases and controls, we also created a contingency table to count the frequencies of regions being either within or outside of a top ten list. We then used Fisher's exact test to conclude whether a difference in proportions existed and ultimately decided whether we had found regions that indiscriminately influenced genes associated with incomplete penetrance.

4.3.3. Identifying Third-Party Regulators of Specific Gene Pairs

Our final motivation was to determine whether both genes within a reduced penetrance pair shared the same interacting fragment/s. We termed such cases as *third-party fragments*. Although this seems similar to our previous analysis, the key difference here was that we were looking for unique regulators between pairs, as opposed to finding single fragments that influenced the population of pairs (figure 4.4). Hence, we formulated the following hypothesis: *reduced penetrance gene pairs are regulated by at least one third-party interacting fragment*.

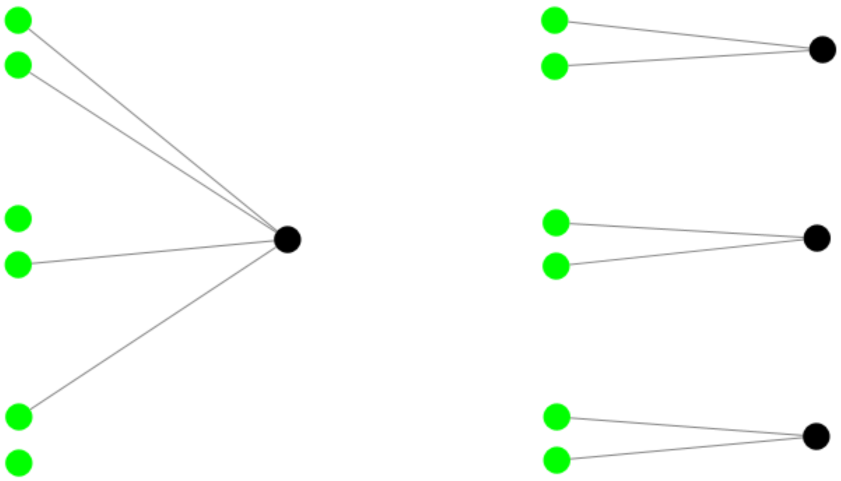


Figure 4.4 Graphical interpretation of two distinct reduced penetrance analyses. Green nodes represent reduced penetrance genes (in pairs) and black nodes represent fragments which have a high interaction frequency with genes in Hi-C data. Edges form a connection if a fragment is found to be a close neighbour of a gene in 3D space. The network on the left depicts the identification of penetrance regulators (section 4.3.2) whereas the network on the right depicts analysis for third-party fragments (section 4.3.3).

Using all reduced penetrance gene pairs that form an inter-chromosomal interaction, we found the top *h* interacting fragments for each gene using *in situ* Hi-C data at 100 kb resolution. We then examined each reduced penetrance pair and noted instances where a particular fragment was found in the top *h* list of both the primary and modifier gene (or gene 1 and gene 2 for the dataset of digenic mutations). We termed these cases intersects; a reduced penetrance pair was then recorded as 'containing at least one intersect' or 'containing no intersects' and frequencies were entered into a contingency table of cases and controls. Controls were generated using algorithm C (figure 4.3), with the random generation being repeated 1,000 times and an average frequency being taken.

Finally, we performed Fisher's exact test to determine whether the ratio of intersects/no intersects from reduced penetrance gene pairs is statistically different to the ratio found for controls. If a significant difference was found, we could conclude that reduced penetrance pairs each have their own regulatory region/s which influence the presence of their respective expected phenotypes. It would then be possible to backtrack and identify regions that were labelled as intersects and investigate their function.

4.4. Results and Discussion

4.4.1. Gene Pair Choices

After postulating that there is a correlation between reduced penetrance gene pair choices and their 3D proximity in the cell nucleus, we found no evidence of such fragments being any more or less represented within the top interactions from 100 kb *in situ* Hi-C data. The occurrences of IF ranks, r, being higher than our variable IF rank thresholds was not significantly different between cases and controls for both of our datasets (table 4.1).

Table 4.1 Fisher's exact test for proportions of reduced penetrance gene pairs being either local or remote (within the 3D cell nucleus) against controls. Local is defined by a pair's IF rank being higher than a threshold (i.e. having an IF within the top 10 ranks, for example) and remote would therefore be cases where IF ranks are outside of a threshold. Dataset 1 contains inter-chromosomal primary and modifier gene pairs; dataset 2 contains inter-chromosomal digenic mutation pairs.

Threshold	Local pairs (dataset 1)	P-value	Local pairs (dataset 2)	P-value
10	4	0.330164	12	0.919297
11	5	0.304127	19	0.651819
12	6	0.337225	21	0.847624
13	7	0.425228	26	0.915969
14	9	0.411347	34	0.710361
15	11	0.415691	43	0.407925
16	11	0.835090	47	0.612769
17	14	0.630172	52	0.563331
18	14	0.860170	56	0.513406
19	15	0.738332	59	0.493136
20	16	0.621349	62	0.517863

Despite our findings leading us to accept that the proportions of local and remote gene pairs for cases against controls are the same, there are some important points to consider. Firstly, as a result of this analysis, we could conclude that the choice of gene pairs associated with reduced penetrance comes down to much more than just the influence of 3D proximity. Hence, isolating this variable for inspection seems to be an oversimplification. Secondly, and perhaps more importantly, we could suggest that although the gene pairs themselves are not neighbours, there could be other undiscovered genomic fragments that act as a regulatory bridge, thus activating or suppressing the expected phenotype. This idea led us to pursue the proceeding analyses of global and local penetrance modulators.

4.4.2. Penetrance-Associated Regulators

We identified the most prevalent 1 Mb fragments from our analysis of each gene associated with inter-chromosomal digenic activity. The more frequently a particular fragment was found to be in the top ten interactions with these genes, the higher the calculated prevalence. We hypothesise that high prevalence is an indicator for the region in question being influential in penetrance regulation for any phenotype. Table 4.2 shows genomic regions that had the highest prevalence values. These were also compared against the fragment's prevalence in control regions in order to make an informed decision about whether the fragment is specifically interacting with penetrance-associated genes or if this behaviour is ubiquitous.

Table 4.2 Fragments which commonly occur in the top ten interactions with digenic mutation genes. Prevalence values underlined indicate overrepresentation compared to controls, and those that are **bold** indicate statistical significance of $p<0.05$. P-values are generated using Fisher's exact test against controls.

Fragment	Gene 1 list prevalence % (and p-value)	Gene 2 list prevalence % (and p-value)	Control prevalence %
chr10:41,000,001-42,000,000	98.41 (0.1510)	**100 (0.0041)**	94.72
chr2:91,000,001-92,000,000	93.65 (0.4209)	94.57 (0.2436)	91.94
chr17:22,000,001-23,000,000	35.38	45.45	56.50
chr1:121,000,001-122,000,000	25.00	22.22	30.41
chr3:198,000,001-199,000,000	**20.97 (0.0272)**	12.24 (0.4806)	11.68
chr18:16,000,001-17,000,000	19.40	21.78 (0.4167)	20.50
chr9:66,000,001-67,000,000	14.71 (0.1990)	10.00	10.79
chr1:1-1,000,000	11.67	12.22	13.11
chr4:48,000,001-49,000,000	10.77 (0.4836)	12.50 (0.2600)	10.03
chr7:61,000,001-62,000,000	9.23	16.49	25.40
chr16:33,000,001-34,000,000	9.09	7.07	9.82
chr16:69,000,001-70,000,000	**9.09 (0.0159)**	**7.07 (0.0332)**	2.99
chr1:141,000,001-142,000,000	6.67	3.33	10.09

At the top of our prevalence list, we encountered fragments chr10:41,000,001-42,000,000 and chr2:91,000,001-92,000,000 exhibiting almost complete ubiquity amongst our digenic gene high interaction lists. However, we saw that for controls, i.e. regions not necessarily associated with any reduced penetrance characteristic, there was also an extremely high prevalence rate (94.72% and 91.94% respectively). A high prevalence in controls could imply a bias with our dilution Hi-C data; a single fragment cannot be physically close to all other genomic regions at any one time, although there could be other explanations. Since our Hi-C data is a snapshot of a population of cells, it is possible that these highly prevalent regions act as a hub of activity during different stages of a cell's life. This therefore makes it possible that a single region can exhibit a high interaction frequency with many other genomic regions, albeit at different moments in time.

Looking back at table 4.2, we encountered one megabase fragment on chromosome 16, positions 69,000,001-70,000,000 which exhibited a significantly higher prevalence for both gene lists compared to controls (Fisher's exact test; respective p-values of 0.0159 and 0.0332). This region, containing a total of 20 genes (table 4.3), could therefore be seen as an indiscriminate regulator of penetrance due to its overrepresentation in digenic mutation gene pairs and its enrichment in genes.

Table 4.3 List of genes contained within the region on chromosome 16, positions 69,000,001-70,000,000.

Genes			
TANGO6	*HAS3*	*CHTF8*	*CIRH1A*
SNTB2	*VPS4A*	*PDF*	*COG8*
NIP7	*TMED6*	*TERF2*	*CYB5B*
MIR1538	*NFAT5*	*NQO1*	*NOB1*
WWP2	*MIR140*	*CLEC18A*	*CLEC18C*

Two other regions with a significant presence in digenic interactions: chromosome 10, positions 41,000,001-42,000,000 and chromosome 3, positions 198,000,001-199,000,000, are gene-poor.

4.4.3. Third-Party Regulators of Reduced Penetrance Gene Pairs

Even for the smallest choice of our threshold, h, we found intersect fragments from the case datasets. Our controls, however, tended to exhibit a statistically similar number of intersects at any given threshold choice (table 4.4).

Table 4.4 Number of intersects between reduced penetrance pairs, given a threshold choice, h. P-values are obtained by performing Fisher's exact test against control pairs. Dataset 1 contains inter-chromosomal primary and modifier gene pairs; dataset 2 contains inter-chromosomal digenic mutation pairs.

Threshold	Intersects (and p-value); dataset 1	Intersects (and p-value); dataset 2
1	12 (0.5)	37 (0.3889)
2	14 (0.5)	48 (0.5522)
3	15 (0.5988)	56 (0.3498)
4	15 (0.5)	66 (0.5)
5	17 (0.4007)	73 (0.5522)
6	23 (0.3939)	85 (0.4446)
7	25 (0.5)	93 (0.4402)
8	30 (0.1283)	100 (0.4339)
9	31 (0.0980)	109 (0.1648)
10	32 (0.0566)	114 (0.0579)

Despite obtaining smaller p-values as we relaxed our threshold further, we could not reach the point of statistically significant difference between cases and controls. Furthermore, increasing our threshold beyond $h = 10$ would not see a continued decrease in p-values, since this is the stage where almost all gene pairs exhibit at least one intersect (32 out of 32 for primary and modifier pairs, 114 out of 121 for digenic mutation pairs). Therefore, any subsequent increase in intersect counts would only be seen in controls, resulting in a return to statistical similarity.

From this analysis, we could not conclude to 95% confidence that partner genes shared common third-party interacting fragments, although there were things we could learn from the methodology and results. Firstly, our method focused on the number of intersects rather than their identity (genomic position). Perhaps the number of third-party regions found between cases and controls should not have been the point of focus, but rather the identities of fragments – particularly those found in cases and not controls. Following from this observation, fragments satisfying a presence in cases and absence in controls could have a variable influence on the final function of the target genes. With the data available, we were not able to conduct experiments to investigate the level of influence of such fragments, but perhaps this is an area of future wet-lab research if potential regions were identified.

It is possible that this analysis (and preceding analyses of reduced penetrance pairs) has fallen victim to small sample sizes; we worked with a total of 170 inter- and inter-chromosomal gene pairs between two datasets. Variable penetrance is in its infancy in terms of our understanding, therefore sourcing a reliable and comprehensive dataset of genes known to exhibit reduced penetrance traits is difficult. With the possibility of more discoveries in this field comes the opportunity to increase our database and consequently our sample size, which may yield improved results if our analysis is repeated in the future.

Chapter V

Gene Fusion Events

The separation and relocation of a genomic fragment is known as chromosomal translocation. These can occur both inter- and intra-chromosomally and if genes are affected at the separation sites, this type of rearrangement can result in so-called fusion genes. It is known that cases of gene fusion events cause diverse types of genetic disease (Kim et al., 2009); a prominent example is the *BCR-ABL1* gene fusion causing chronic myeloid leukaemia (CML; Hunter, 2007). Gene fusions account for 20% of human cancer morbidity (reviewed in Mitelman et al., 2007) and so an improvement in understanding the mechanisms involved would go a long way for prevention and treatment.

Our general hypothesis in this chapter was that any two regions which harboured one half of a fusion gene pair would also be neighbours in 3D space. This is not to be confused with the proximity between all fusion genes; rather the initial locations of a pair of genes are known to commonly fuse together. We investigated this hypothesis using two approaches. Firstly, we proposed that any region harbouring a fusion gene would show an enrichment of interactions from our 3D interaction data (Hi-C; Lieberman-Aiden et al., 2009; Rao et al., 2014) and consequently a closeness to other fragments within the cell nucleus. Secondly, we hypothesised that the choice of gene fusion pairs is not random. That is, one half of a known fusion gene pair would consistently favour a particular gene as a fusion candidate. We understand that the 3D proximity of gene pairs is not the only criterion for a fusion event – the nucleotide sequence of the potential fusion site must be acceptable for example – but we suggest that the 3D structure of the human genome contributes to the choice of partners.

5.1. Basic Definitions

There are three types of chromosomal rearrangement that can lead to a fusion gene. For the inter-chromosomal case, a reciprocal *translocation* can occur, whereby two fragments on separate chromosomes can detach and simply swap places (figure 5.1). This is the type of rearrangement that occurs for the aforementioned *BCR-ABL1* case, where the *ABL1* gene on chromosome 9 fuses to the *BCR* gene on chromosome 22, forming the *Philadelphia chromosome*.

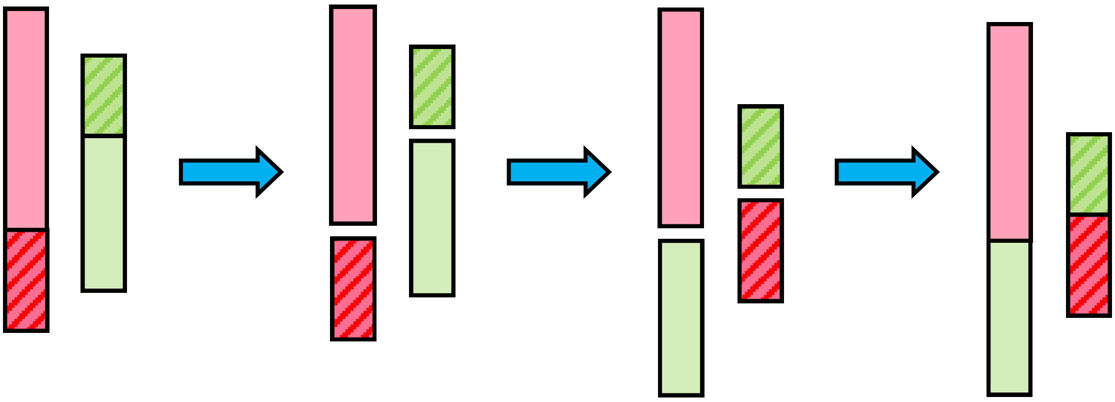


Figure 5.1 Diagram describing an inter-chromosomal translocation.

One intra-chromosomal example of rearrangement that leads to a gene fusion event is a *deletion*. This is where a fragment is removed from the chromosome and the remaining fragments join (figure 5.2).

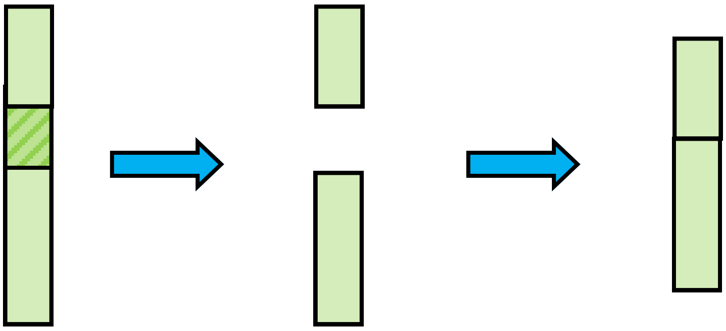


Figure 5.2 Diagram describing an intra-chromosomal deletion and subsequent fusion.

The third example of rearrangement leading to a gene fusion is an intra-chromosomal *inversion* (figure 5.3). Here, a fragment separates from the chromosome, rotates and reunites such that the fragment fuses to the opposite separation points.

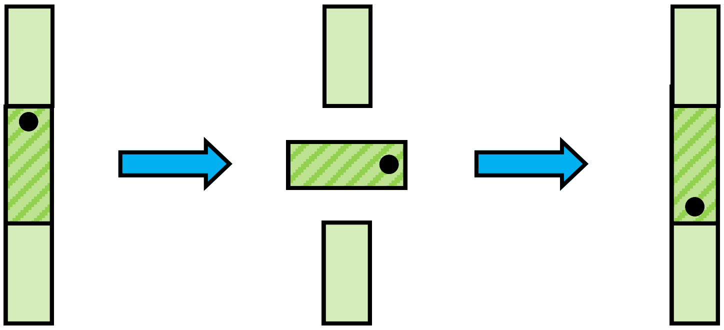


Figure 5.3 Diagram describing an intra-chromosomal inversion.

Gene fusion events are clearly a symptom of genomic reorganisation. Fragments containing genes that are known to be associated with fusion events are important to analyse by way of ascertaining their folding behaviours in healthy cell lines. In other words, it is necessary to determine whether these types of reorganisations occur within a small/precise 3D neighbourhood or rather a much larger space. The former seems likely; it is more plausible that a loose DNA fragment would rebind to something nearby rather than making a journey to a remote region. This idea is reinforced when considering that there are relatively common gene fusion events within the population – the folding principles of the human genome are non-random (Lieberman-Aiden et al., 2009), therefore fusion gene-containing fragments will often be found within the same 3D neighbourhood, which provides ideal conditions for chromosomal rearrangements to occur.

5.2. Materials

5.2.1. Dataset of Gene Fusion Pairs

The ChimerDB 2.0 database (Kim et al., 2009), containing 11,747 fusion gene pairs, was used for our analysis. These were labelled *head* and *tail* genes in order to show the direction of the fusion event. Each fusion was sourced from various repositories: the Mitelman's database (Mitelman et al., 2007), the Sanger Cancer Genome Project (Futreal et al., 2004), the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2008) and a range of PubMed searches. Entries were categorised into one of three classes: A, B or C, depending on the confidence of the pair representing a genuine fusion. From the total 11,747 fusion pairs available, 1,643 were chosen. This represented the complete set of the most reliable, class A fusion gene pairs, 76% of which were inter-chromosomal fusion events (1,247 total). The remaining 396 intra-chromosomal pairs were cases where the head and tail genes were separated by a minimum of 1 Mb and were therefore non-adjacent.

5.3. Methods

5.3.1. Generating Control Datasets

In our gene fusion analysis, we employed a mixture of methods, which came under two main branches. The first branch involved the interrogation of head and tail genes independently; the second branch analysed head and tail genes as a dependent pair. Hence, our generation of control datasets needed to reflect each branch of analysis appropriately.

Controls for Independent Genes

For descriptive purposes, we define a binned genomic region in Hi-C data as a pair of co-ordinates (chromosome, bin number). Our controls corresponding to a list of independent head and/or tail genes would mimic the size of the list as well as the distribution of chromosomes. For example, the number of control fragments on chromosome 6 would match the number of genes found on chromosome 6 in our test list. A description of the algorithm for generating controls is shown in figure 5.4.

```
I1. [Retain chromosome.] Given a gene fusion region $(\alpha, i)$ and desired control region $(\alpha^*, i^*)$, set $\alpha^* = \alpha$.
I2. [Select random bin.] Randomly generate a bin, $i^*$, within the bin range appropriate for chromosome $\alpha$.
I3. [Is region associated with gene fusions?] If the generated control region $(\alpha^*, i^*)$ does not exist as a head/tail gene in the ChimerDB database, our algorithm terminates; $(\alpha^*, i^*)$ is our control region. If the region does exist in ChimerDB, repeat step I2.
```

Figure 5.4 Algorithm I: pseudocode generating control fragments to test against independent head/tail genes.

The motivation behind retaining the chromosome choice was to match cases and controls in terms of genomic location. This would avoid a chromosome bias that could be introduced by selecting any region.

Controls for Dependent Gene Pairs

The generation of controls for dependent fusion gene pairs followed the same principles as seen for independent genes, although this procedure was modified slightly to account for a pair of regions as opposed to one (figure 5.5).

D1.	**[Retain entire head gene and chromosome of tail gene.]** Given a gene fusion pair defined by head and tail regions (α,i) and (β, j), respectively and corresponding desired control regions (α^*,i^*) and (β^*, j^*), respectively, set $\alpha^*=\alpha$, $i^*=i$ and $\beta^*=\beta$.
D2.	**[Select random tail gene bin.]** Randomly generate a bin, j^*, within the bin range appropriate for chromosome β.
D3.	**[Is tail gene region associated with gene fusions?]** If the generated control region (β^*, j^*) does not exist as a head/tail gene in the ChimerDB database, our algorithm terminates; (β^*, j^*) is our control region. If the region does exist in ChimerDB, repeat step D2.

Figure 5.5 Algorithm D: pseudocode generating control fragment pairs to test against gene fusion pairs.

Selecting a brand new pair of regions would not have represented a fair comparison between cases and controls. For our analysis of pairs, we wished to investigate the 3D proximity of head and tail genes in comparison to the proximity of a head gene and an unrelated region. Hence, we retained the head gene for our controls and selected a random tail region on the original tail gene chromosome.

5.3.2. Global Interaction Profiles of Fusion Genes

Our first hypothesis was that regions harbouring either head or tail genes were enriched in interactions measured by various Hi-C methods. In other words, we proposed that a region containing a fusion gene was likely to be on average less isolated in 3D than a randomly selected control. If true, fusion gene regions could be viewed as hubs, with various regulatory fragments being nearby in 3D space. We used both dilution Hi-C (Lieberman-Aiden et al., 2009) and *in situ* Hi-C data (Rao et al., 2014) at resolutions of 1 Mb and 100 kb for this analysis.

To test this hypothesis, we first obtained the mean interaction frequency (IF) for each fusion gene region. Since each index of a Hi-C matrix, M, corresponds to a given binned region, we were able to calculate both the intra- and inter-chromosomal IF means (denoted by $\bar{x}_i$ for bin i) by taking the sum of the corresponding row and dividing by the number of non-zero entries in that row. Given that $M_f = [m_{f1}, m_{f2}, \ldots, m_{fk}]$, where m_{fl} is the interaction frequency between region f and any other region, l, let $M''_f = [m''_{f1}, m''_{f2}, \ldots, m''_{f(k-z)}]$ be a subset of M_f containing only non-zero entries (hence, a total of z zero entries are removed). Thus, $\bar{x}_i$ is given by

$$\bar{x}_i = \sum_{j=1}^{k-z} \frac{m''_{fj}}{k-z}.$$

For each tested region, controls were generated 1,000 times and the mean, $\bar{x}_i$, was obtained by taking an average of the 1,000 control values. Thus, we had two sets of IF means: let $\bar{F}$ be the vector of mean IF values for fusion genes and let $\bar{C}$ be the vector of mean IF values for corresponding controls. The ordering of $\bar{F}$ and $\bar{C}$ is important, since entries are in matched pairs.

In explicit terms, our null and alternative hypotheses were as follows:

$$\begin{aligned} H_0 &: \quad \bar{F} - \bar{C} = 0, \\ H_1 &: \quad \bar{F} - \bar{C} > 0. \end{aligned}$$

We tested for differences in means using the two sample t-test; our alternative hypothesis was one-tailed. That is, we specified that the means for our fusion dataset would be greater than corresponding controls.

5.3.3. Proximity of Fusion Gene Pairs

We also hypothesised that regions harbouring gene fusion pairs had high interaction frequencies in Hi-C data relative to control pairs. If true, this would consequently imply that the choice of gene fusion partner is non-random and the 3D architecture of the human genome plays a significant role in the choice.

We used the counter vector method seen in our reduced penetrance analysis with gene fusion pairs (figure 5.6). Regions harbouring fusion genes were described by a pair of co-ordinates representing chromosome and bin. We were therefore able to determine the IF rank of both cases and controls, and we could consequently obtain frequencies of pairs that satisfied a chosen rank threshold, h. Corresponding controls were generated 1,000 times and the total frequencies were averaged, thus giving non-discrete counts in our tables.

```
F1. [Construct counter vector.] Let $f$ be the index of the genome-wide, $k \times k$ contact matrix, $M$, that corresponds to chromosome $\alpha$, bin $i$. It follows that $M_f = [m_{f1}, m_{f2}, \ldots, m_{fk}]$, where $m_{fl}$ is the interaction frequency between region $f$ and any other region, $l$. We construct our counter vector, $v = [v_1, v_2, \ldots, v_n]$, where $v_1$ is a counter for the number of interactions (or entries of $M_f$) with frequency 1, $v_2$ with frequency 2, and so on. The last entry, $v_n$, corresponds to the maximum interaction frequency recorded for fragment $f$.
F2. [Find IF of pair.] Find $m_{fg}$, where $g$ is the index of $M$ corresponding to chromosome $\beta$, bin $j$.
F3. [Find IF rank.] If $c = m_{fg}$, then the interaction frequency rank, $r$, is found by counting the number of non-zero entries of $v$ between $v_c$ and $v_n$. If $c = 0$, then $r = n + 1$.
```

Figure 5.6 Algorithm F: pseudocode ranking fragment pairs according to their interaction frequency relative to other interacting fragments. A higher rank implies the tested pair of regions are relative neighbours in 3D space.

After tabulation of frequencies, we performed one-tailed Fisher's exact proportion tests to measure the statistical difference in the proportions of neighbouring regions between gene fusion pairs and controls. Our hypotheses were as follows:

H_0: The proportions of fusion events within the top h interacting fragments are the same for both the dataset of fusion genes and controls.

H_1: The proportions of fusion events within the top h interacting fragments are greater for our fusion genes dataset compared to controls.

Resulting p-values that satisfied $p < 0.05$ would lead us to reject our null hypothesis, in favour of concluding that the proportion of head and tail gene partners within highly interacting fragments is greater than for controls.

5.4. Results and Discussion

5.4.1. Interaction Enrichment of Fusion Genes

Interrogation of regions harbouring fusion genes by using both dilution and *in situ* Hi-C data at various resolutions revealed significant differences of interaction frequencies between cases and controls (table 5.1).

Table 5.1 Two-sample t-testing of the interaction frequencies of regions harbouring fusion genes against controls. Small p-values indicate fusion regions having an enrichment of interactions from Hi-C data.

		Head Gene Regions	**Tail Gene Regions**
Dilution Hi-C	**Intra 1 Mb**	5.95×10^{-3}	3.74×10^{-3}
	Intra 100 kb	6.47×10^{-12}	1.15×10^{-10}
	Inter 1 Mb	5.95×10^{-63}	9.49×10^{-67}
***In situ* Hi-C**	**Intra 1 Mb**	4.45×10^{-3}	1.34×10^{-3}
	Intra 100 kb	0.090739	0.018708
	Inter 100 kb	4.12×10^{-154}	1.96×10^{-157}

We see the most significant differences between fusion regions and controls for inter-chromosomal tests, particularly in comparison to intra-chromosomal t-tests. This difference in significance may occur as a result of two factors. Firstly, our intra-chromosomal dataset is almost four times smaller than our inter-chromosomal dataset. Therefore, testing with a smaller n has an effect on the relative statistical power. Secondly, intra-chromosomal results could be affected by chromosome territory bias. That is, interaction frequencies for intra-chromosomal pairs are much higher than their inter-chromosomal counterparts. This leads to a tiling property seen in heat maps of Hi-C data (figure 5.7).

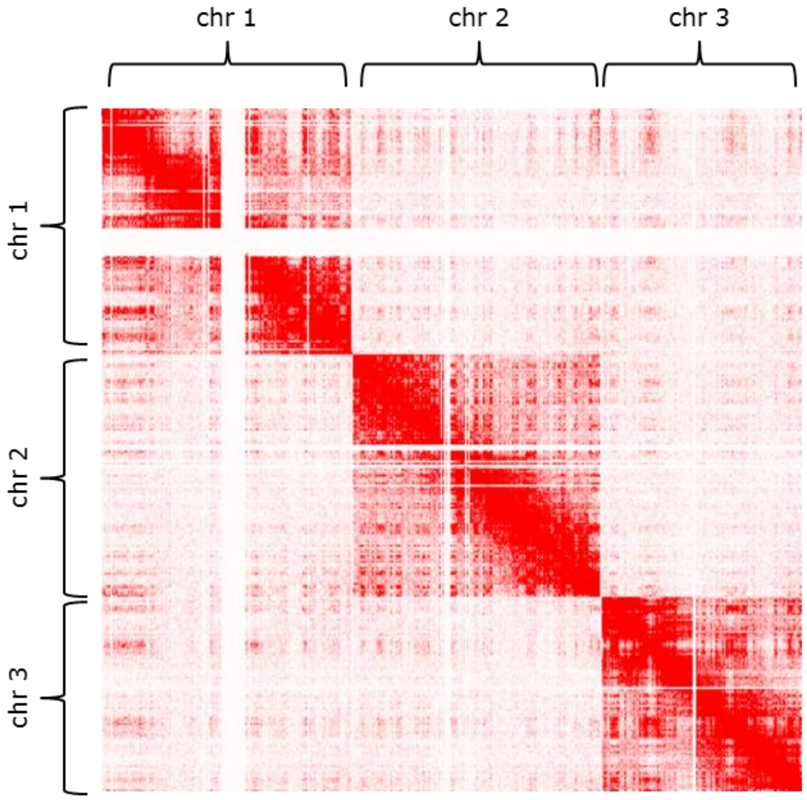


Figure 5.7 Interaction heat map displaying the intra-chromosomal tiling feature which exists as a result of chromosome territory bias.

The reason this feature exists is because regions on the same chromosome are more likely to be 3D neighbours than those on different chromosomes. In essence, we are saying that the Hi-C methodology is prone to a one-dimensional proximity bias. Therefore, a lesser difference is observed between intra-chromosomal cases and controls because many more interaction frequencies are relatively high.

5.4.2. Fusion Gene Pair Choices

Pairs of regions which harboured head and tail fusion genes were found more frequently within the highest IF ranks than corresponding controls. Table 5.2 shows results from a variety of threshold choices for Hi-C data at 1 Mb resolution. The more the IF rank threshold was relaxed, the higher the observed statistically significant difference in proportions between gene fusion regions and controls. For 1 Mb data, a threshold choice of $h = 10$ represents approximately the top 8% of intra-chromosomal interactions and the top 0.3% of inter-chromosomal interactions. We were able to conclude a significant difference in proportions at this choice for all three 1 Mb tests (inter-/intra-chromosomal dilution Hi-C and intra-chromosomal *in situ* Hi-C). This implies that the choice of gene fusion pairs is at least partially dependent on where these regions are in 3D space.

Table 5.2 Fisher's exact test for proportions of gene fusion pairs being either local or remote (within the 3D cell nucleus) against controls. Local is defined by a pair's IF rank being higher than a threshold (i.e. having an IF within the top 10 ranks, for example) and remote would therefore be cases where IF ranks are outside of a threshold. Results from this table were all calculated from 1 Mb Hi-C data.

Threshold	Local Pairs (intra dilution Hi-C)	P-value	Local pairs (intra *in situ* Hi-C)	P-value	Local pairs (inter dilution Hi-C)	P-value
1	0	1	8	1	0	1
2	13	2.60×10^{-3}	17	0.143123	0	1
3	22	1.14×10^{-4}	22	0.052604	1	1
4	43	7.76×10^{-9}	37	3.66×10^{-4}	3	0.943798
5	55	4.36×10^{-11}	43	6.68×10^{-5}	5	0.595015
6	61	7.94×10^{-12}	49	1.60×10^{-6}	8	0.243799
7	70	2.48×10^{-13}	55	2.97×10^{-6}	14	0.027174
8	81	1.55×10^{-15}	62	3.36×10^{-7}	16	0.015058
9	86	4.44×10^{-16}	64	5.02×10^{-7}	17	0.013782
10	90	3.33×10^{-16}	70	9.05×10^{-8}	21	3.62×10^{-3}

We repeated this experiment for a higher resolution of *in situ* Hi-C data (100 kb; table 5.3). Once again, significant differences in proportions were found given the right threshold choices. For inter-chromosomal tests, we found significant proportion differences at a threshold choice of $h = 8$ (at a 95% confidence level). This threshold represents approximately the top 0.03% of interactions, and is therefore still very stringent. Similarly, we found significance at a threshold choice of

$h = 29$ (corresponding to approximately the top 2.2% of interactions) for intra-chromosomal tests.

Table 5.3 Fisher's exact test for proportions of gene fusion pairs being either local or remote (within the 3D cell nucleus) against controls. Results from this table were all calculated from 100 kb *in situ* Hi-C data.

Threshold	Local pairs (inter)	P-value	Threshold	Local pairs (intra)	P-value
5	13	0.223215	25	24	0.325254
6	15	0.177400	26	26	0.214532
7	17	0.367634	27	29	0.104971
8	36	5.06×10^{-3}	28	31	0.064102
9	58	1.35×10^{-3}	29	33	0.039045
10	96	2.15×10^{-5}	30	36	0.017635
11	143	8.95×10^{-8}	31	39	8.38×10^{-3}
12	200	4.41×10^{-10}	32	40	8.83×10^{-3}
13	260	2.13×10^{-11}	33	43	4.37×10^{-3}
14	322	2.00×10^{-14}	34	45	3.59×10^{-3}
15	382	7.47×10^{-16}	35	46	2.75×10^{-3}

Particular attention should be given to table 5.3, since the 100 kb resolution of this data is perfect for encompassing genes within one or two bins, whilst fragments remain small enough to make precise distinctions between immediate neighbours. For example, suppose a head gene was wholly contained within one 100 kb bin and its associated regulatory elements were located on a nearby, separate bin. With this resolution, it is possible to distinguish between the head gene itself being a 3D neighbour of a tail gene, or alternatively an associated promoter/enhancer being closest. This is not as plausible with 1 Mb data, since the neighbourhood of a gene and associated regulatory elements are usually within a region of at most 1 Mb in length (Symmons & Spitz, 2013) and are therefore typically contained within the same bin. In uncommon cases where the distance between genes and regulatory elements is larger than 1 Mb, Hi-C data at 100 kb resolution remains the more precise means of distinguishing the closest fragments.

The findings from our analysis all support the hypothesis that the 3D structure of the human genome is highly influential in fusion gene pair formations. At the very least, we know there is a correlation. The proximity of head and tail genes in healthy cell lines is proof that either (1) being in a 3D neighbourhood is a driving factor for gene fusion events, or (2) if 3D proximity is not the underlying cause, it certainly provides ideal conditions for fusion events to occur.

Indeed, previous studies have found that DNA fragments that are loose as a result of double-stranded breaks have a limited mobility (reviewed in Wijchers & de Laat, 2011). This lack of mobility was found by recreating a 3D spatial closeness of known head and tail fusion genes, and observing the fusion efficacy at given distances. For example, in prostate cancer cells, the head gene *TMPRSS2* (on chromosome 21) was found to fuse to tail genes *ERG* (also on chromosome 21) and *ETV1* (on chromosome 7) when the 3D proximity between head and tail genes was sufficiently small (Lin et al., 2009; Mani et al., 2009). The fusion preferences of head and tail genes was also shown to be conserved between cells belonging to hosts of various ages. The head gene *RET* was found to fuse with tail regions containing *NCOA4* and *H4*, causing thyroid cancer (Gandhi et al., 2006). These events occur despite such regions being one-dimensionally remote on chromosome 10: *RET* and *NCOA4* are separated by almost eight million base pairs, for example. Furthermore, three gene families involved in recurrent translocations causing cancer, *MYC*, *BCL* and *IGH*, were found to preferentially fuse. Interestingly, genes from the three families described above were all commonly found at the interior of the nucleus, implying a small 3D genomic distance between one another (Roix et al., 2003).

An analysis of almost 9,000 non-pathogenic deletion breakpoints from over 1,000 human samples showed a strong association between non-allelic breakpoint sites with a high sequence similarity and open chromatin marks, identified by eigenvector transformations of Hi-C matrices (Abyzov et al., 2015). Whilst these findings confirm that data describing the 3D structure of the human genome is influential, its focus is primarily on the chromatin state. Hence, our results align with this study, but further analysis is needed in order to ascertain whether our regions have an open or closed chromatin signature.

Markers for fusion events can also transcend human cell lines. Additionally, both inter- and intra-chromosomal translocation frequencies were found to correlate with the 3D organisation of the mouse genome (Zhang et al., 2012). The study induced breakpoints across whole chromosomes and observed the frequency of translocation events at each breakpoint, whilst recording the initial distances between translocation pairs. Translocations in *cis* occurred between regions separated by at least 1 Mb, and there was also a correlation between inter-chromosomal Hi-C interaction frequencies at 5 Mb resolution and translocation frequencies at corresponding 3D locations. Whilst these results align with our own analysis for gene fusion events, it is important to note that the Hi-C data used was from a different species and, perhaps most interestingly, our results were obtained with up to a 50-fold increase of Hi-C resolution.

The next natural step for our gene fusion analysis is to incorporate various network measures, described in chapter 3 of this thesis, into analysis of networks based on published Hi-C data. In a previous study, it was suggested to use node centrality measures to identify drivers of gene fusion events (Wu et al., 2013). The key difference was that networks for this study were constructed by assigning head/tail genes as nodes, and edges connected a pair of nodes if they participated in a gene fusion event. Nodes with a high centrality, therefore acting as hubs in so-called *fusion networks*, were found to be the most prolific in tumour formation.

Our proposal would be to identify fusion drivers for nominated disorders by constructing 3D interaction networks from cell-specific Hi-C data and using more comprehensive datasets of fusion events. Whilst recent years have seen an increase in publications of cell-specific Hi-C data (such as for brain tissues; Won et al., 2016), at the time of study, we had a limited choice of reliable Hi-C libraries. Therefore, this was beyond the scope of the thesis and is a suggested direction for future work. One would expect to find that head/tail fusion genes exhibit an enrichment of connections in networks describing the 3D architecture of the human genome. Consequently, we would expect these fusion genes to be hubs in such networks, displaying high scores from various centrality measures. Furthermore, these hubs will be especially easy to identify when future studies increase the size and reliability of fusion event datasets.

Chapter VI

Schizophrenia

Schizophrenia is a chronic, severe and disabling brain disorder that has a lifetime prevalence of 1% (reviewed in Rees et al., 2015), with common symptoms including hallucinations, delusions and changes in behaviour. The causes of schizophrenia are known to be a combination of environment, brain structure/chemistry and perhaps most importantly, genotype (Eaton et al., 2008). The role of the genotype is evidenced by the fact that the occurrence of schizophrenia is 10% for those that have a first-degree relative with the disorder, and 40-65% for a monozygotic twin (Cardno & Gottesman, 2000). Furthermore, heritability in liability to this disease is reported at 70-80% (Keller et al., 2012).

Previous studies have identified genetic variants, such as single nucleotide polymorphisms (SNPs), that are known to play a role in the development of schizophrenia (Ripke et al., 2011). These variants have an association with the disorder measured by statistical p-values obtained from genome-wide association studies (GWAS). The cut-off for genome-wide significance of SNPs having a disease association is $p < 5 \times 10^{-8}$. Whilst many SNPs do not meet this cut-off, it is commonly accepted that many genetic variants may each make a small contribution to a disorder through interactions with each other (Jia et al., 2010). Thus, individually, SNPs may not seem to have a significant effect on disease, but when treated as a group, one can investigate the combined effect this may have. This is especially important since the majority of SNPs are found in intergenic (non-coding) DNA regions, which account for approximately 98% of the human genome (Elgar & Vavouri, 2008). Although relatively little is known about these regions, recent research suggests that the non-coding regions in the human genome are enriched in functional elements (The ENCODE Project Consortium, 2012), therefore creating a regulatory map of genetic variants found in these regions could shed light on the functional importance of SNPs in schizophrenia due to their role in influencing gene expression.

In this study, we hypothesised that both common and individually rare variants found by GWAS exert their influence on target genes, either directly via long-range looping interactions between fragments that harbour SNPs and target genes/promoters, or are propagated via the network of interactions governed by the 3D architecture of the human genome. Hi-C data was used to quantify 3D distances between chromosomal regions and various network measures were calculated in order to achieve our aim of producing a novel library of genomic regions most associated with schizophrenia. Subsequent techniques were also used as a means of interrogating such regions. We used DAVID software (Huang et al., 2009) to classify the genes found within our regions into functionally similar groups. Additionally, we utilised expression quantitative trait loci (eQTL) analysis (Shabalin, 2012) to identify gene expression within distinct regions of the organ which is primarily responsible for schizophrenia development – the brain.

6.1. Background

6.1.1. Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNPs) are individual base pair changes in a genome and are the most frequent form of genetic variation in the human genome. Synonymous (or silent) SNPs are considered to have no effect on the proteins produced from a nucleotide sequence, whereas non-synonymous SNPs result in a change of the amino acid sequence via missense or nonsense polymorphisms. Whilst a missense SNP will alter the amino acid sequence, a nonsense SNP induces a premature stop codon, thus halting the function of proteins. The relative locations of SNPs are also indicators of functional change; they can occur in non-coding regions of the genome, such as promoters or enhancers, which in turn can affect gene expression. The locations and densities of such SNPs across different populations have been catalogued thanks to the International HapMap project (International HapMap Consortium, 2005). This project describes the common patterns of human genetic variation and is invaluable in finding genetic SNPs associated with disease via genome-wide association studies.

6.1.2. Genome-Wide Association Studies

Genome-wide association studies (GWAS) measure and analyse SNPs from across the human genome in an effort to identify genetic risk factors for diseases that are common in the population (Bush & Moore, 2012). For diseases that frequently occur within the population, the common disease/common variant (CD/CV) hypothesis was developed as a way of identifying disease associated SNPs (Reich & Lander, 2001). The hypothesis states that a genetic variant in the population that is equally as prevalent as a disorder may share a direct cause-effect relationship. This does not ignore the idea that the SNPs found can have a variable effect on the disease in question. GWAS have identified common variants with a small influence on a disease, such as variants in the *LMTK2* gene for prostate cancer (Eeles et al., 2008) and also common variants which have a large influence, such as SNPs found in the *APOE4* gene for Alzheimer’s disease (Corder et al., 1993). Typically, both SNP frequency and effect size, the latter measured by odds-ratios and statistical p-values, can lie anywhere on a spectrum from small to large (effect size; odds-ratio) or rare to common (SNP frequency). The odds-ratio is a measure of the likelihood of disease in an exposed subject compared to non-exposed subjects (Clarke et al., 2011), where exposure can be interpreted as either having an allelic trait (i.e. allele a or A) or a genotypic trait (i.e. genotype AA, aa or aA). Corresponding p-values were initially calculated using a simple chi-square test of a 2-by-2 contingency table describing the incidence between cases and controls.

Nowadays, logistic regression is used (described in section 3.2.1 and reviewed in Bush & Moore, 2012). This is a modification of linear regression, where predictors can be both continuous and categorical, whilst the response is transformed from continuous into binary form (i.e. phenotype presence/absence). For GWAS, the set of independent predictive variables, $X = \{X_1, \ldots, X_n\}$, always includes a genotype variable describing SNPs (for example, a normal genotype '…AATT…' is set to $X_i = 0$ and a SNP-affected genotype '…AAGT…' is set to $X_i = 1$), and can optionally incorporate variables such as age, sex, ethnicity, blood type, etc. to give the model additional complexity.

Measures of effect size are commonplace in GWAS and form an integral part of the summary statistics in such studies. Such statistics, as well as SNP frequency, are important considerations when testing the CD/CV hypothesis. The CD/CV method can identify common SNPs associated with a disease, but it does not necessarily show a full picture, with potential rare variants being overlooked from this type of study.

Thanks to HapMap genotype data, SNPs identified by GWAS can also identified as being in linkage equilibrium (LE) or linkage disequilibrium (LD). LD describes the co-inheritance of SNPs within a population over time. Suppose we have two variants (alleles), $\{A, a\}$, at locus 1 and two alleles, $\{B, b\}$, at locus 2 on a section of chromosome (haplotype). The chance of inheritance for each allele can therefore be expressed as

$$\begin{aligned} P(A) &= p, \\ P(a) &= (1-p), \\ P(B) &= q, \\ P(b) &= (1-q). \end{aligned}$$

Assuming independence, one would expect the following haplotype frequencies:

$$\begin{aligned} P(AB) &= pq, \\ P(Ab) &= p(1-q), \\ P(aB) &= q(1-p), \\ P(ab) &= (1-p)(1-q). \end{aligned}$$

This set of frequencies describes perfect LE. That is, there is a distinct ratio of haplotype frequencies occurring through generations. It is worth noting that LE/LD is still applicable with more than two alleles. In cases where we do not see the expected haplotype frequencies, we have LD, which can be quantified using the measure D (Lewontin & Kojima, 1960):

$$D = P(AB)P(ab) - P(Ab)P(aB).$$

Clearly, for a case of LE, as seen above, the value of D is zero. A higher than expected frequency of our homozygotic pairs would satisfy $D > 0$, whereas a higher frequency of heterozygotic pairs would satisfy $D < 0$.

The model after one generation can be expressed in a similar way:

$$D' = P'(AB)P'(ab) - P'(Ab)P'(aB),$$

where

$$\begin{aligned} P'(AB) &= P(AB) - rD, \\ P'(Ab) &= P(Ab) + rD, \\ P'(aB) &= P(aB) + rD, \\ P'(ab) &= P(ab) - rD, \end{aligned}$$

and r is the rate of recombination, which is a process where chromosomes can overlap and exchange genetic material during meiosis. This rate satisfies $0 \le r \le 0.5$. Recombination can therefore contribute to a decay of LD over time, given by

$D' = D - rD = D(1-r)$. Since D' is directly proportional to r, SNPs or markers which have a low recombination rate stay in linkage through generations within a population and are therefore said to be in LD.

Cases of LD are now also measured by using R^2. This is an output obtained from logistic regression which measures how well one SNP can act as a proxy for another (The International HapMap Consortium, 2005). Values for R^2 can lie anywhere within the boundaries of 0 and 1, with a score of 1 indicating perfect LD between SNPs (undisrupted by recombination).

These instances are cases of interest for GWAS because without this property, association studies would have to independently analyse each SNP. That is, co-inheritance of SNPs on a population scale provides us with a cost-effective way of performing GWAS, as continuous stretches of DNA containing particular SNPs are shared among a vast group of people.

There are two main classifications of phenotypes in GWAS: categorical and quantitative. The categorical class is often *case versus control*, and arguably seems an oversimplified method, given the complex nature of many disorders. The most common method of statistical analysis for categorical data is a contingency table method – using either the chi-square test or Fisher's exact test (a variant of chi-square). The null hypothesis would be that there is no association between the phenotype and genotype classes, and a low p-value would indicate the chance of seeing no association is small. Genome-wide significance is set at $p < 5 \times 10^{-8}$ due to allowances being made for multiple testing – 10^6 comparisons are needed, resulting in a Bonferroni correction of $p < 0.05/10^6 = 5 \times 10^{-8}$. The quantitative class is often analysed using Analysis of Variance (ANOVA), with the null hypothesis being that there is no difference between the trait means of any genotype group. Despite a preference for quantitative measures, both classifications are used in GWAS and both yield successful results, although one should be wary of factors which may influence a particular trait, such as gender and age. These factors cause an increase in degrees of freedom and therefore a loss of statistical robustness.

6.2. Materials

6.2.1. Dataset of Schizophrenia Genes

We obtained a list of 347 genes that are known to have an association with schizophrenia. These SNP-containing genes were identified on a sample of 21,856 individuals of European ancestry and replicated on a sample of 29,839 independent subjects (table A6.1; Ripke et al., 2011). The genomic positions of each gene were lifted over to the hg19 reference genome assembly using the Lift Genome Annotation program available at https://genome.ucsc.edu/cgi-bin/hgLiftOver (Kent et al., 2002). The purpose of lifting over to the hg19 reference genome is to align gene positions with both *in situ* Hi-C and Capture Hi-C data.

6.2.2. Dataset of SNPs

A total of 1,252,901 SNPs spanning 22 chromosomes have been catalogued and scored according to their association with schizophrenia via p-values (Ripke et al., 2011). These scores lie within the range $4.3\times10^{-11} < p < 0.998$, with only 136 SNPs reaching the genome-wide significance threshold of $p < 5\times10^{-8}$. Each SNP in this dataset is labelled with an exact genomic location relative to the hg19 reference genome, in order to align with *in situ*/Capture Hi-C data.

6.2.3. Database of Gene Expression Regulators

We utilised data from the Functional Annotation of the Mammalian Genome project (FANTOM5; The FANTOM Consortium, 2014) in order to locate transcription start sites (TSSs), and therefore promoters, from the Cap Analysis of Gene Expression (CAGE) method (Kanamori-Katayama et al., 2011). The CAGE method reads short nucleotide sequences (known as tags) and identifies the origin of the tag using a reference genome. CAGE peaks are genomic regions that show a high concentration or clustering of tags, which are quantified by normalised tags per million (TPM) values. Typically, CAGE peaks with a high TPM count correlate with high expression changes.

Although the FANTOM5 project investigates expression across 975 human and 399 mouse samples, we are interested specifically in human brain tissues. The database called Semantic catalogue of Samples, Transcription initiation And Regulators (SSTAR; Abugessaisa et al., 2016) provided us with the data for the brain tissue samples from three sources: an adult sample from both genders with ages 77, 79 and 81 years, another adult sample from a male aged 18 years, and finally a fetal sample (both genders) of age 20-33 weeks. We used SSTAR to search for transcription factors (TFs) with enriched expression in these tissue samples. The data itself contains the CAGE peak names (labelled with the corresponding TF and in descending order of tag support, i.e. p1 is the promoter with the highest tag support, followed by p2, etc.), TPM values and normalised expression values. For each sample, we have access to the top 1000 TFs in terms of their expression enrichment.

6.2.4. Gene Expression in Brain Regions

In order to identify SNPs which are significantly associated with the change in expression of known genes, we use expression quantitative trait loci (eQTL) analysis (reviewed in Shabalin, 2012). Specifically, we use an exon-specific database which describes expression in ten different brain regions: cerebellar cortex (CRBL), frontal cortex (FCTX), hippocampus (HIPP), inferior olivary nucleus (sub-dissected from the medulla; MEDU), occipital cortex (OCTX), putamen (PUTM), substantia nigra (SNIG), temporal cortex (TCTX), thalamus (THAL) and intralobular white matter (WHMT). This database is known as the Brain eQTL Almanac (Braineac; Ramasamy et al., 2014). Ramasamy et al. (2014) reported that *trans*-eQTL signals had a high false positive rate, therefore only *cis*-eQTL signals (up to 1 Mb in distance from a given marker, hence intra-chromosomal) were used for our analysis.

By using the Braineac web tool available at www.braineac.org, we are able to input SNPs (by their genomic location or reference SNP cluster ID (rsID)) and obtain an expression profile of the most affected genes in all ten brain regions, including an expression average across all regions (aveALL). The relative expressions are calculated by modelling the effect of genotype using a least squares or ANOVA (analysis of variance) model and resulting p-values correlate with the strength of expression change of the affected gene.

6.2.5. Gene Functional Classification

The Database for Annotation, Visualisation and Integrated Discovery (DAVID) is a web tool that is able to classify groups of genes with similar function (Huang et al., 2009). There are two motivations behind utilising this web tool; firstly, we wish to save ourselves from investigating biological functions on a gene-by-gene basis for large lists and secondly, we want to be able to understand the function of a gene (or group of genes) in order to make conclusions about possible links to schizophrenia development.

Our first motivation is achieved by using the Gene Functional Classification tool. The user inputs a list of gene names as well as a choice of classification stringency (default is medium), which decides the threshold for strength of functional association between genes within a group. Genes are consequently clustered into groups if they meet the association criteria and the overall similarity strength between grouped genes is described via an enrichment score. The higher the enrichment score, the tighter the biological relationship between genes. This score is logarithmically correlated with the calculation of an EASE (Expression Analysis Systematic Explorer) score, which is a one-tailed modification of Fisher's exact test used to measure gene enrichment in annotation terms (Huang et al., 2009). An enrichment score greater than 1.3 corresponds to an EASE score of $p = 0.05$, therefore scores exceeding 1.3 are interesting cases.

The Gene Name Batch Viewer tool is used to ascertain the biological function of genes on a one-by-one basis. Once a functional group of genes is found, we can explore the properties of each gene. As well as the biological function, we are able to find other related features, such as tissue specificity, known disease associations and also extra information via links to publications that cite the gene in their studies.

6.3. Methods

In general, our approach for schizophrenia was two-fold. Firstly, we constructed 3D interaction networks from Hi-C data and performed unbiased analysis in what we called our *top-down* (or *global*) approach. This approach consisted of identifying the most important nodes by: (1) creating sub-networks containing only the highest Hi-C interaction frequencies, (2) calculating various network measures from raw and normalised Hi-C networks and (3) identifying connected components at given interaction frequency thresholds. The aim of this approach was to indiscriminately find nodes corresponding to genomic regions that could be associated with schizophrenia. Once we had found such regions, we termed them *candidate regions*.

We then introduced our *bottom-up* (or *targeted*) approach, in which we investigated our candidate regions more closely. This was achieved by going through a checklist of features that could be signposts for schizophrenia association, such as an overrepresentation of SNPs, functionally relevant genes and enrichment of gene expression in brain tissues. Some methods from our global approach were also utilised in our targeted approach, such as examining the centrality measures of particular nodes and comparing their scores relative to the rest of the network. Candidate regions exhibiting many or all of these features were therefore identified as being the most likely contributors to schizophrenia development.

6.3.1. Extended Gene Regions

Using Capture Hi-C data (for a chosen cell line – see justification for the choice of cell line in section 6.3.5) which identifies significant interactions between promoters and genes/enhancers, we were able to pair each gene associated with schizophrenia to its corresponding promoter by concatenating regions recorded as baits, thus forming continuous gene-promoter regions. From here, we employed Capture Hi-C data from the same cell line which identifies promoter-other interactions ('other' consisting of regulatory elements such as enhancers) in order to find associated enhancers both upstream and downstream of any given gene-promoter region. Once we had found these, we included the genomic locations of the enhancers to our existing gene-promoter regions to form a larger amalgamated, continuous region which we term as an extended gene region (EGR; figure 6.1). In most cases, these EGRs were labelled with the name of the corresponding gene. However, there are also two special cases to consider. In instances where one or more of the original 347 EGRs were wholly contained within another EGR, we would amalgamate these into a single EGR. For example, if *G1* had *G2* within its genomic boundaries, we would label the new EGR as *G1{G2}*. Our second special case involves EGRs which are not wholly contained, but slightly overlap, i.e. there is some shared genomic space between two or more EGRs. Assuming *G3* and *G4* overlapped, this would be labelled *G3-G4*. The resulting EGR list for our analysis is reported in table A6.2.

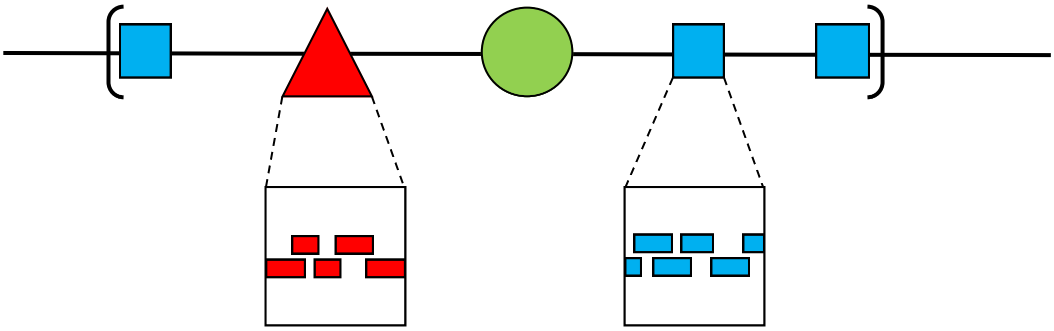


Figure 6.1 Schematic representation of the creation of extended gene regions (EGRs). Genes (green circles), promoters (red triangles) and enhancers (blue squares) are depicted here in 1D space. An EGR is the concatenation of regulatory elements associated with a gene. Boundaries of EGRs are represented by the brackets in this figure. Promoters and enhancers are identified by the presence of baits and fragments interacting with these baits from the Capture Hi-C procedure. These fragments are shown as coloured rectangles upon closer inspection of a regulatory element. These typically overlap, and so the length of our regulatory elements are decided by the length of the amalgamated baits.

6.3.2. Network Analysis

Sub-networks of Regions Connected with High Interaction Frequencies

For the first part of our global approach, our initial set of all nodes, V, was partitioned to create two new, distinct subsets, U and V, where $U \bigcup W = V$. This resulted in a bipartite network representation between subset $U = \{u_1, u_2, \ldots\}$, containing all unique EGRs, and subset $W = \{w_1, w_2, \ldots\}$, containing genomic regions which interact with one or more element in U with a suitably high interaction frequency (IF) from inter-chromosomal *in situ* Hi-C data. Nodes in U and W were connected by an edge if the IF between the regions described by each node exceeded a threshold which we defined. The threshold was chosen such that our networks described <1% of all interaction frequencies: only the very highest remained. Typically, resulting networks contained between 100 and 1,000 nodes, where nodes in U could only be connected to nodes in W and vice versa (i.e. the bipartite property, described above).

We were therefore left with a network describing the regions of the human genome that interacted most frequently with our EGRs. The network was described by using a weighted adjacency matrix, in which the weights represent the IF between regions. This was used in order to identify the most important connections and to be able to distinguish those particular edges from the rest of the network. The edges were assigned weights according to the number of connections between nodes u_i and w_j, which was derived from the relevant *in situ* Hi-C dataset (figure 6.2).

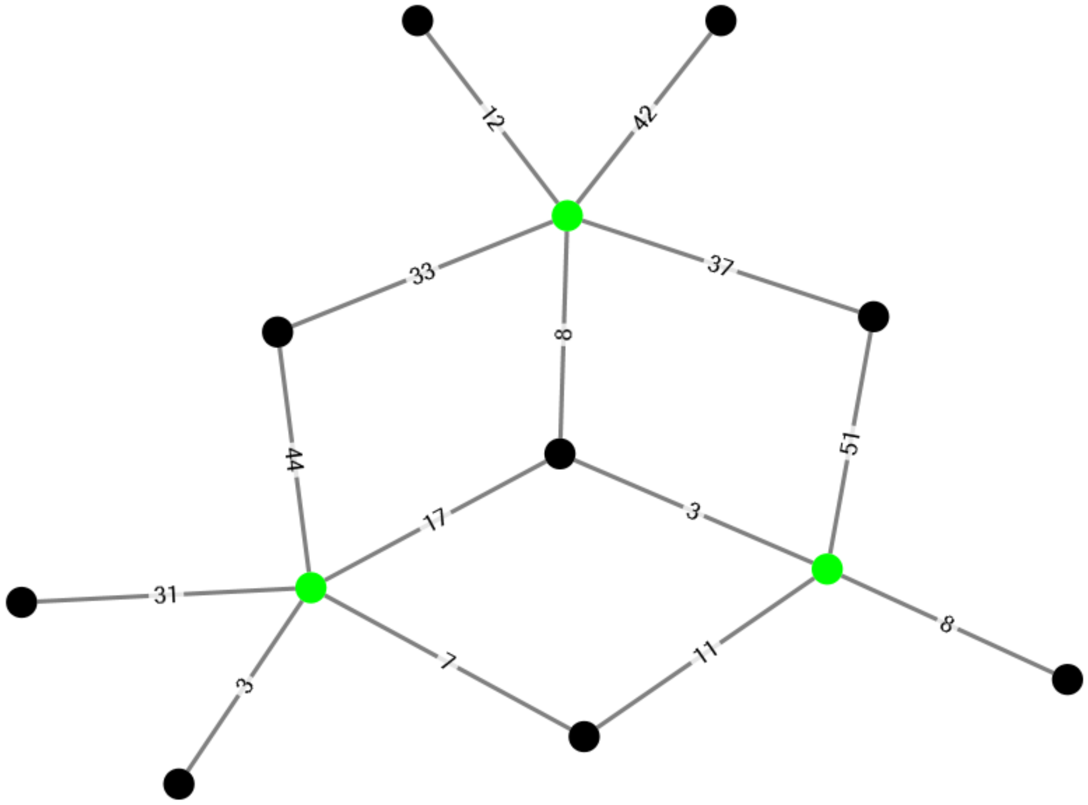


Figure 6.2 Example bipartite 3D interaction network between binned EGRs (regions from U ; green nodes) and binned regions found to have a Hi-C interaction frequency above our chosen threshold (regions from W ; black nodes).

Multiple regions from W were merged into a single node if and only if $\{w_j, w_{j+1}, \ldots\}$ formed a continuous region and were all connected to a distinct EGR, v_i. We made this alteration because consecutive genomic fragments that interact with the same regions are effectively a single longer fragment with one cumulative connection. The network consequently reduces in complexity without any loss of information. Also, if an interacting region from the network (in set W) was found to share some common genomic space with a node in set U, we would change the label of node w_j to the corresponding node in U, thus introducing the possibility of connections within set U. This second alteration was performed in order to avoid distinct nodes in our network describing identical stretches of DNA. Our final network would therefore no longer necessarily have bipartite properties, although nodes in W would still not share connections.

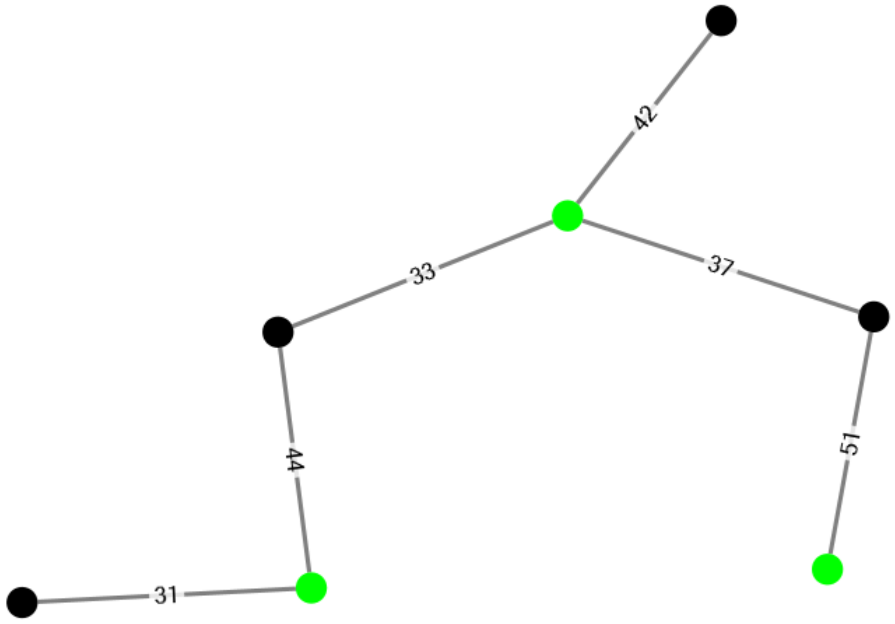


Figure 6.3 Pruned 3D interaction network. Weighted connections of less than 25 have been removed. Remaining black nodes are potential candidates for schizophrenia association.

After constructing networks for both raw and seven other normalised *in situ* Hi-C datasets (described in Chapter 2), we pruned them such that only the edges with the highest weights remained (figure 6.3). This thresholding was chosen because of the population-based nature of Hi-C data: there is no one fixed moment during the cell cycle or a single cell/specific tissue where proximities are being captured. Therefore, since the genome is not static within the cell nucleus during its lifetime, it is possible for proximities to be captured in passing. We assumed these passing cases were described by low interaction frequencies and were therefore considered not to be representative of sustained contacts between two fragments. Hence, passing cases were removed through the pruning procedure. This typically left us with a network of approximately 10 nodes from W, as well as a number of EGRs. The regions from W were listed and ranked according to the edge with the highest weight that was incident to w_j. Since we performed these steps with eight raw/normalised variants of *in situ* Hi-C data (described in chapter 2), we could also investigate the prevalence of nodes from W across all eight lists (i.e. how often a particular node appears in the pruned networks). Hence, we constructed a master list, which sorted all remaining regions via two criteria: *individual list rank* and *list prevalence*. For list ranking, the node with the largest single edge weight was assigned rank 1, and subsequent nodes were ranked with increasing numbers. Nodes with equal edge weights were ranked identically, with the next node being given rank $n+1$, where n is the number of nodes which have already been assigned a rank. List prevalence was assigned to nodes according to the number of lists a given node would appear in. This left us with our desired output – the nodes that were sorted to the top of our master list would represent the regions of the genome which hypothetically had the most profound influence on schizophrenia via long-range interactions. Figure 6.4 summarises these steps algorithmically.

N1.	**[Filter Hi-C matrix.]** Set entries of the Hi-C interaction matrix to zero if they do not meet our chosen threshold, t. If a pair of nodes, $\{v_i, v_j\}$, are connected by an edge and neither v_i or v_j are an EGR, remove the edge connecting them. That is, M is a submatrix containing indices only applicable to EGRs and $(M_{v_i,v_j} < t) = 0$.
N2.	**[Construct node subsets.]** Partition the node set $V = \{v_1, v_2, \ldots\}$ into subsets U and W, where nodes from U are EGRs and nodes from W are all interacting regions. Any nodes which overlap in position are amalgamated into single nodes.
N3.	**[Create network.]** Construct a weighted adjacency matrix, A, which describes the connections between nodes in U and W. Edges are weighted according to the number of connections fragments in u_i have with fragments in w_j.
N4.	**[Prune network.]** If entries of A do not meet our edge weight threshold, set them to zero.
N5.	**[Identify remaining w_j nodes.]** Any remaining nodes from W are listed and sorted in descending order of incident edge weights.
N6.	**[Compile master list.]** Steps N1-N5 are repeated for each normalised *in situ* Hi-C data. Then, from all 8 independent lists, we compile a master list of all remaining nodes and sort them according to the rank within their individual list and their list prevalence (i.e. the number of lists they appear in). Nodes do not have to appear in all lists to become an entry of our master list, but those which appear in most are likely to be ranked higher (a highly ranked candidate would typically appear in ≥ 4 lists). Nodes at the top of this list are therefore identified as possible schizophrenia-related candidate regions.

Figure 6.4 Algorithm N: pseudocode identifying the regions of the genome which interact most frequently with schizophrenia EGRs according to *in situ* Hi-C data.

Node Centrality Measures

Given that the overarching aim of this chapter is to identify genomic regions which have an association with schizophrenia through analysis of 3D interaction networks, intuition tells us that this can be achieved by identifying the most important nodes. Centrality is a de facto way of finding nodes of influence. Since the relative influence of any node can be explicitly measured, it follows that nodes can be easily compared and contrasted, i.e. node i is seen as more important than node j simply because of its higher centrality score.

For global 3D interaction networks, we used four centrality measures: betweenness, degree, eigenvector and PageRank (descriptions and motivations for use are given in chapter 3) to identify important nodes which consequently correspond to candidate regions. Intra- and inter-chromosomal *in situ* Hi-C data was used at 100 kb resolution to construct these networks, which were converted from weighted to unweighted by setting an interaction frequency threshold. Our first inter-chromosomal threshold was set so that any entry exceeding the median of non-zero entries from the Hi-C adjacency matrix was considered as an interaction and our second inter-chromosomal threshold was set as the third quartile of non-zero entries of the Hi-C matrix. We set these thresholds because the distribution of IFs in Hi-C data is heavily left-skewed. That is, there is a large number of low IFs, therefore we set the thresholds to ignore these counts in favour of genomic fragments that are more frequently crosslinked (hence removing potential false positive cases). For intra-chromosomal data, we set the threshold at $F > 0$. That is, any non-zero entry of the adjacency matrix was considered an interaction. The motivation for our intra-chromosomal threshold was that setting anything higher than zero would increase the chance of removing remote (3D) interactions in favour of interacting regions that were one-dimensionally close. We chose more than one threshold for inter-chromosomal data in order to examine the behaviour of nodes at varying stringencies, i.e. answering the question, “Would the most central nodes change or be preserved at different thresholds?”

From our binary Hi-C networks, we sorted the centrality scores in descending order, identifying those at the top as being our candidates ready for targeted analysis. We put additional emphasis on nodes that appeared near the top of more than one ranked centrality list and also nodes that were present between different threshold choices, as these are regions that communicate extremely well with other nodes in the network. We also recognised that some centrality measures are highly correlated compared to others (an example being eigenvector and PageRank centrality). We therefore emphasised the importance of nodes that rank highly in lists of uncorrelated measures (such as betweenness centrality) more so than those ranking highly between similar measures.

Local Clustering Coefficients

We were able to identify communities of nodes that preferentially interacted with one another (i.e. clustered) via the local clustering coefficient. That is, clustering is not a measure that identifies a node's importance relative to the whole network, rather it acts as a means of finding small areas of close communication – in our case, this means finding genomic regions that consistently share a 3D neighbourhood.

Our Hi-C networks were constructed exactly as seen above (identical to those used for centrality measures) and we identified nodes with the highest local clustering coefficients as regions of interest that consistently participated in 3D interactions with other distinct regions. Clusters of most interest were those that contained genes that have an association with schizophrenia. Subsequent targeted approaches would also reveal the potential functional relevance of the novel regions found within particular clusters – perhaps clusters indicate regions of functional similarity, for example.

Connected Component Analysis

Our final top-down (global) approach aimed to identify candidate regions by examining the components of our 3D interaction networks. Initially, we had our raw and normalised inter-chromosomal, weighted adjacency matrices, describing the 3D proximities between pairs of genomic regions between chromosomes. We treated this data by first setting an interaction frequency threshold, which allowed us to distinguish between genomic pairs that were sufficiently close in a 3D setting and those that were not 3D neighbours. Hence, our Hi-C adjacency matrices had binary properties and were therefore unweighted. After choosing a suitable interaction frequency threshold for raw Hi-C data, we ensured that an equivalence of thresholds was struck between raw and normalised matrices by examining the percentile of the raw threshold choice, and consequently mimicking that percentile with other data. For example, let $F = 50$ be our chosen interaction frequency threshold for raw, *in situ* Hi-C data. Suppose we found $F = 50$ to be the 81st percentile of interaction frequencies in the raw data, we would then seek the 81st percentile within a normalised matrix and set the threshold accordingly (e.g. if the 81st percentile for the Knight & Ruiz normalised Hi-C matrix was $F = 32$, then our threshold for the KR matrix would be 32). Note that for cases where an irrational F was found, F was rounded to the nearest integer. In short, we threshold so that the same percentage of interactions are present amongst different Hi-C datasets.

After completion of the thresholding procedure, we were able to easily identify connected components by finding sub-networks which satisfied the condition that any pair of nodes were connected by at least one path, but did not connect to any node from elsewhere in the network (figure 6.5).

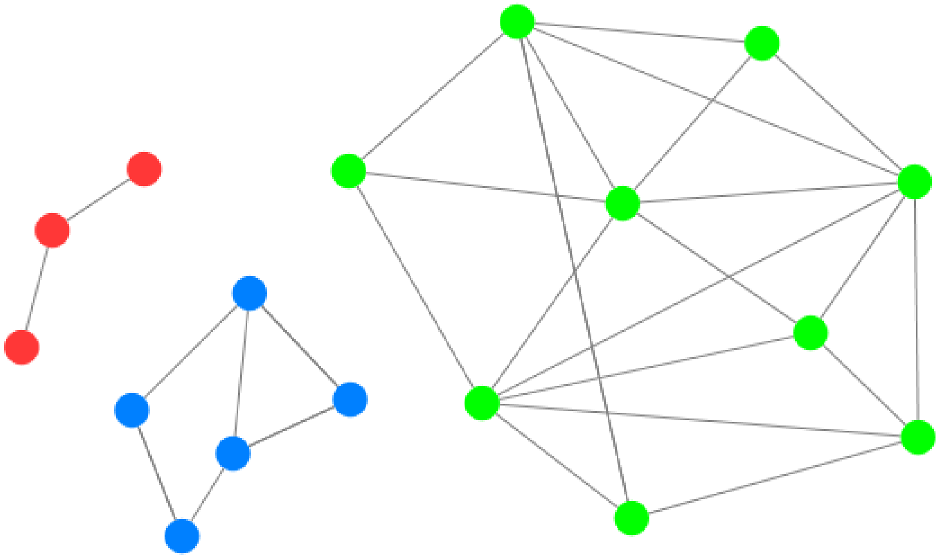


Figure 6.5 A network with three connected components. Nodes of identical colour are contained within the same component.

Our next step of the search for candidate regions led us to look for families of nodes which consistently interacted within the same component – particularly if the component had a large presence of genes that have known associations with schizophrenia. Nodes of most interest were those that (1) connected with many nodes locally (i.e. within the component) and globally (i.e. within the whole network, given that the interaction frequency threshold was relaxed to allow for components to be more connected), or (2) did not necessarily have a high degree, but did connect important components. In other words, we used our centrality measures as a tool to identify and verify important nodes within components.

Crucially, our analysis only focused on influential nodes that share connections/components with schizophrenia-associated genes. Otherwise, our findings would be non-specific and we would therefore fail to identify candidate regions for schizophrenia association. In addition, we implemented our targeted approach within our component analysis by identifying the most influential genes from our pool of existing schizophrenia associated genes. We identified important genes by their presence in components of all raw/normalised data. Since raw, inter-chromosomal vanilla coverage (INTERVC) and inter-chromosomal Knight & Ruiz (INTERKR) thresholded networks were used, if a schizophrenia-associated gene appeared as a member of a connected component (with other genes/regions) in all three Hi-C networks, we identified that gene as having the most overlaps and perhaps therefore being more influential.

6.3.3. SNP Analysis

After locating schizophrenia candidate regions through various indiscriminate network approaches, we wished to investigate these genomic regions in more detail. In particular, the occurrence of SNPs is perhaps one of the most important features to assist in identifying important regions. Features of SNPs across a bounded region that we were interested in included the SNP frequency, the spread or distribution of said SNPs, and the strength of association with schizophrenia (obtained by p-values from GWAS).

Since candidates forming the set W can vary in size (from 100 kb to several Mb), the raw frequency of SNPs may introduce a bias towards the larger regions. Hence, we formulated a SNP density score, which calculates the average number of SNPs per 100 kb fragment. This, together with observing the range of SNP p-values each region contains, will give us an insight as to which regions from W in particular may have the most influence on schizophrenia.

6.3.4. Analysis of Expression Regulators

Using FANTOM5 brain tissue data, we explored the association between gene-containing candidate regions for schizophrenia and the presence of TSS enrichment, which indicates a promoter region. Both the raw count of CAGE peaks and the relative expression values were used as a measure to confirm/deny whether the regions we found from our network analysis were likely to play a role in schizophrenia development.

We also used age as a variable due to its difference in our three brain tissue samples. Similarities/differences in the frequency of CAGE peaks between all three samples were explored, as well as the strength of expressions and TPM counts. This analysis in particular could further explain properties of the regions our network algorithm finds, hypothetically indicating at what stage in life a functional element could contribute to schizophrenia formation, and not just where.

6.3.5. Choice of Hi-C Cell Line to Emulate 3D Interactions in Brain Tissues

We sought to validate the use of *in situ* Hi-C, which was obtained from various non-brain cell lines, for the study of interactions in brain tissues. Our first conundrum was to choose the most suitable data from the available *in situ* cell lines: GM12878 (lymphoblastoid cell line), HMEC (mammary epithelial) or IMR90 (lung fibroblast). A cell line with the highest percentage of brain-related eQTL pairs residing within intra-chromosomally interacting regions was considered suitable for further identification of SNPs and their target genes. The eQTL pairs used were the *average of all* (aveALL) brain regions from Braineac data (Ramasamy et al., 2014). We first created three groups of *cis*-eQTL pairs based on their relative one-dimensional locations. Pairs separated by >10 kb, >100 kb or >1 Mb were placed into respective groups and any remaining pairs that were within these distance thresholds were deleted. That is, if the SNP and TSS positions of an eQTL pair were within the same binned region, the eQTL pair was not considered. We also only used one SNP for any given bin – this was selected by using the SNP causing the most expression change per TSS. For example, consider three SNPs: S_1, S_2 and S_3, that are all located on the same binned genomic fragment. SNPs S_1 and S_2 both pair with TSS T_1, and S_3 pairs with T_2 (note that T_1 and T_2 are located on a different genomic fragment to the SNPs). The expression change p-values corresponding to S_1, S_2 and S_3 are 0.03, 0.004 and 0.007, respectively (figure 6.6). In this scenario, we would use only the pair corresponding to S_2 for our analysis (i.e. S_2-T_1), as it is responsible for the most gene expression change from the available options. The cell line exhibiting the most overlap between eQTL pairs and interactions in Hi-C data would then be chosen as the preferred library to emulate 3D interactions in brain tissues.

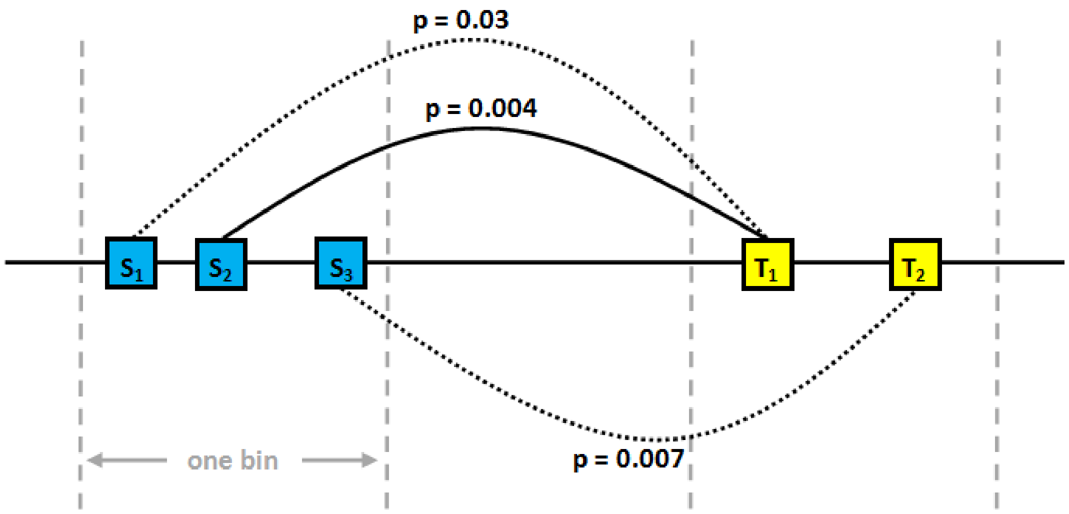


Figure 6.6 Example of three SNP and TSS (eQTL) pairs located on the same bins. The pair with the most expression change (hence smallest p-value; bold line) remains, whereas other cases are deleted (dotted lines).

6.3.6. Brain eQTL Analysis

We then used eQTL data to further test at a more stringent level for an association between the presence of *cis*-eQTL pairs and their corresponding 3D proximity from our chosen *in situ* Hi-C cell line. This was achieved by constructing a 2-by-2 contingency table of true/false counts corresponding to the presence of *cis*-eQTL pairs and 3D proximity of SNP and gene regions (figure 6.7). Fisher's exact test was then employed to consider association between these variables across data for all ten brain regions, as well as for the average of all regions (aveALL). Note that p-values correlate with a strong dependency amongst SNP-TSS pairs of 3D proximity and expression change.

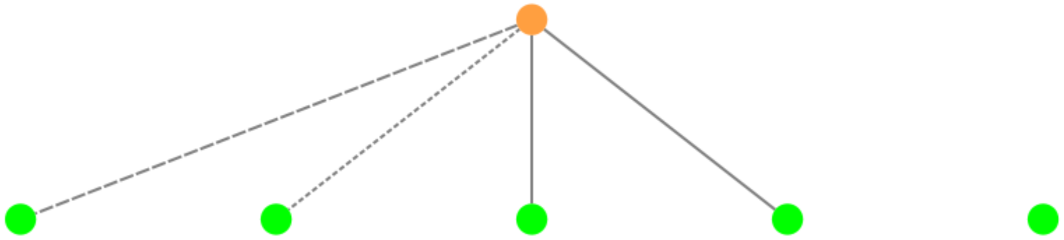


Figure 6.7 Network interpretation of the construction of contingency tables for Hi-C and eQTL association. There are four scenarios to consider: (1) if a SNP (orange node) and gene (green node) have a high interaction frequency in Hi-C data, they are connected by a dashed edge (first connection, from left to right); (2) if a SNP and gene form a *cis*-eQTL pair, they are connected by a dotted edge (second connection); (3) if both conditions (1 & 2) are met, nodes are connected by a solid edge (third and fourth connection); (4) if neither of the first two conditions are met, there is no connection between nodes (furthest right green node). Frequencies of each type of connection are counted and subsequently implemented into a 2-by-2 contingency table (table 6.1).

Table 6.1 Example 2-by-2 contingency table corresponding to information from figure 6.7.

	***Cis*-eQTL pair**	**No *cis*-eQTL pair**
Hi-C interaction	2	1
No Hi-C interaction	1	1

A SNP and corresponding gene are said to be neighbouring in 3D space according to the IF threshold we set for them. Since we employed this method for three resolutions of *in situ* Hi-C data (5 kb, 10 kb and 25 kb), our choice of IF threshold needed to reflect the chosen bin size appropriately. Therefore, we said that a pair of genomic regions are sufficiently close in 3D space if the IF was greater than the median of non-zero counts in our intra-chromosomal Hi-C matrices. With respect to our contingency tables, anything that exceeded this threshold would be counted as true (condition 1, and in some cases condition 3; figure 6.7).

The strength of gene expression change that a SNP had on a gene was measured by its p-value from *cis*-eQTL analysis. Small p-values indicated a significant change of expression in the affected gene, therefore we also set a threshold to decipher pairs that did not experience an expression change to those that did. Our threshold was set such that entries of our contingency table were counted as true for the presence of *cis*-eQTL pairs if $p < 0.05$ (conditions 2 and 3; figure 6.7).

Assuming there was a link between eQTL pairs and their proximity in 3D space for our chosen cell line, candidate regions (from W) found from our network analysis were consequently investigated in our targeted approach. All SNPs occurring within any genomic region from W were entered into the Braineac web tool, which subsequently returned a comprehensive list of all affected genes with corresponding expression p-values across all brain regions.

Our aim with this data was two-fold – firstly, we wished to examine the expression profile of regions that we had found to be schizophrenia candidates on a macro scale. That is, given a region of the genome spanning one or more 100 kb bins, we wished to find the brain region(s) which exhibited the most expression change according to Braineac data. Secondly, from our highly expressed, large genomic regions we sought to identify clusters of SNPs spanning smaller genomic regions which had the most profound impact on gene expression. Both of these objectives were best visualised by producing heat maps of expression in genomic regions across our ten distinct brain regions. We converted our p-values from Braineac data into expression scores by taking a $-\log_{10}$ transform. Areas of high expression and SNP clusters were therefore easily identifiable from the resulting heat maps. On a macro scale, we took the mean expression score per genomic fragment, and on a more refined scale, we were able to take mean expression score for each SNP by averaging the best ten markers. We focused on the best ten markers because averaging all markers per SNP results in homogeneity due to the vast number of markers with low expression scores.

6.4. Results and Discussion

6.4.1. Justification of Using GM12878 Hi-C Data for Emulating the Brain

Using *in situ* Hi-C data at corresponding resolutions (10 kb, 100 kb and 1 Mb), we set the interaction frequency threshold at 80 as a means of defining that a pair of intra-chromosomal regions were reasonably nearby in 3D. We then sought the binned locations of each *cis*-eQTL pair with p-value, corresponding to the expression change, being less than 10^{-6} and determined whether *cis*-eQTL harbouring fragments had an interaction frequency greater than our threshold, hence constituting a significant interaction. Almost all (95%) *cis*-eQTL pairs were found on interacting fragments in all cell lines used (GM12878, HMEC, IMR90) across all bin sizes combined. However, the proportion of *cis*-eQTL pairs residing within strongly (with frequency of interactions ≥80) interacting fragments was the highest in the GM12878 cell line (73%). These proportions were much lower for IMR90 (23.8%) and HMEC (6.6%) cell lines. Using Fisher's exact test to compare the proportions of strongly interacting fragments harbouring *cis*-eQTL pair in different cell lines, we found that GM12878 had a significantly higher proportion as compared to IMR90 or HMEC ($p < 10^{-398}$). With only 27% of *cis*-eQTL pairs not forming a strong looping interaction in GM12878 data, we concluded that this cell line was the most suitable for emulating the 3D architecture of the human genome in brain tissues. Henceforth, all mention of Hi-C data corresponds to *in situ* Hi-C for the GM12878 cell line, unless otherwise stated.

Using our preferred cell line from *in situ* Hi-C data (blood; GM12878), we have shown that the presence of eQTL pairs in ten brain regions and corresponding proximities in 3D space are not independent (Fisher's exact test; table 6.2). Entries in table 6.2 have been corrected for multiple testing by multiplying p-values by ten, which is the total number of brain regions in this study. There is some variability in association strength between brain regions and resolution choice, and perhaps what is most striking is the consistently small p-values found with smaller bin sizes (particularly 5 kb resolution). This is encouraging, since we can conclude that as region specificity improves, so does the association between gene expression and 3D proximity. This gives further validity to our use of *in situ* Hi-C data from the blood cell line in studying brain interactions.

Table 6.2 Results from Fisher's exact tests between *cis*-eQTL pairs and 3D proximity according to Braineac and *in situ* Hi-C data respectively (see section 6.3.6 for method).

	5 kb resolution	**10 kb resolution**	**25 kb resolution**
aveALL	2.54×10^{-37}	2.44×10^{-20}	3.31×10^{-2}
CRBL	5.73×10^{-25}	4.75×10^{-12}	2.20×10^{-8}
FCTX	5.79×10^{-15}	5.94×10^{-10}	1.07×10^{-8}
HIPP	1.31×10^{-34}	1.03×10^{-14}	1.13×10^{-5}
MEDU	1.10×10^{-21}	2.59×10^{-12}	1.90×10^{-5}
OCTX	8.31×10^{-19}	1.53×10^{-8}	1.77×10^{-4}
PUTM	6.42×10^{-15}	6.64×10^{-9}	1
SNIG	6.16×10^{-23}	0.101	0.311
TCTX	1.70×10^{-27}	5.85×10^{-11}	5.51×10^{-4}
THAL	1.28×10^{-28}	2.36×10^{-19}	0.192
WHMT	1.38×10^{-10}	1.27×10^{-14}	1

6.4.2. Global Approach: Identifying Candidate Regions

3D Interaction Sub-networks

We found a number of high ranking regions from W that contained genes which are known to have an association with schizophrenia (table 6.3). From a total of 54 nodes (covering 35 unique continuous regions) in W across all raw and normalised *in situ* Hi-C data, 28 of these were found to contain at least one gene not present in our original dataset. There were four particular nodes of interest which harbour genes with a known association with schizophrenia, reported by various authors.

Table 6.3 Novel nodes identified by our network analysis that contain schizophrenia-associated genes.

Node (hg19 position)	Dataset used	Schizophrenia-associated genes	Reference(s)
chr6:200,000-400,000	RAW, SQRTVC	*IRF4*	Moskvina et al., 2010
chr6:26,000,000-26,500,000	RAW, INTERKR	*HIST1H2BC, HIST1H2BD, HIST1H2BG, HIST1H2BH*	Sanders et al., 2013
chr6:26,000,000-26,500,000	RAW, INTERKR	*HIST1H1E*	Föcking et al., 2014
chr6:26,000,000-26,500,000	RAW, INTERKR	*BTN3A3*	Chen et al., 2014
chr11:100,000-800,000	RAW	*IFITM1, IFITM2, IFITM3*	Hwang et al., 2013; Saetre et al., 2007; Arion et al., 2007
chr11:100,000-800,000	RAW	*HRAS*	Goriely & Wilkie, 2012; Comings et al., 1996
chr11:100,000-800,000	RAW	*IRF7*	Lukasz et al., 2013
chr11:100,000-800,000	RAW	*DRD4*	Cheng et al., 2014
chr19:300,000-1,400,000	RAW	*FGF22*	Terauchi et al., 2010
chr19:300,000-1,400,000	RAW	*GRIN3B*	Lin et al., 2014

RAW:	Raw interaction counts.
KR:	Knight & Ruiz matrix balancing.
INTERKR:	Inter-chromosomal Knight & Ruiz matrix balancing.
GWKR:	Genome-wide Knight & Ruiz matrix balancing.
VC:	Vanilla coverage normalisation.
INTERVC:	Inter-chromosomal vanilla coverage normalisation.
GWVC:	Genome-wide vanilla coverage normalisation.
SQRTVC:	Square-rooted vanilla coverage normalisation.

Figure 6.8 Glossary for *in situ* Hi-C data types.

Of the raw and seven normalised sets of *in situ* Hi-C data used in this study (figure 6.8 serves as a reference for identifying the type of *in situ* Hi-C data from their respective acronyms/abbreviations), we found that regions from W containing genes associated with schizophrenia were all found in the raw data (15 out of 15 associated genes). A region on chromosome 6 between positions 26,000,000-26,500,000 (hg19 assembly) containing genes *HIST1H2BC*, *HIST1H2BD*, *HIST1H2BG*, *HIST1H2BH*, *HIST1H1E* and *BTN3A3*, was additionally found from the INTERKR data. Five of these genes belong to the histone gene family, which are responsible for the compaction of chromatin into higher order structures and therefore play a role in the 3D organisation of the human genome. Furthermore, a region on chromosome 6, positions 200,000-400,000, found from raw data and containing gene *IRF4*, was also found from the normalised, SQRTVC data. From genes listed in table 6.3, we found *FGF22* to have a particularly high comparative expression in brain tissues (second highest of 27 tested human tissues; Fagerberg et al., 2014). This gene belongs to the fibroblast growth factor family of genes and plays a role in many processes, such as embryonic development, tissue repair and cell growth.

Results obtained from this indiscriminate, global network analysis confirm the initial hypothesis that genes and genomic regions associated with schizophrenia are in close physical contact with each other, since we have found fragments that do not appear in our original dataset but are reported to have an association to schizophrenia in other publications. The raw *in situ* Hi-C data is perhaps the best predictor for finding regions which have significant disease associations, but we should not discard results from various normalised data as they could act as a validation technique in some cases, particularly since normalisation procedures serve to remove experimental bias from data.

Amongst entries of table 6.3, we have found gene-rich regions occurring on 11 different chromosomes which had no previous implication in schizophrenia at the time of analysis; the highest ranked from our sorted master list are shown in table 6.4. We identify these regions and the genes contained within them as having a potential influence on this disease according to our hypothesis, and therefore may warrant further inspection in future schizophrenia research.

Table 6.4 Genomic regions from network analysis that contain genes with no previous schizophrenia association. These are sorted according to step N6 of algorithm N (figure 6.4, section 6.3.2).

Node (hg19 position)	Dataset used	Genes in region
chr9:137,400,000-138,900,000	RAW	32
chr21:15,000,000-15,200,000	INTERVC, GWVC	*POTED, DQ590589, DQ591735*
chr9:39,800,000-40,300,000	INTERVC	*SPATA31A2, FAM74A1*
chr1:143,400,000-143,600,000	VC, SQRTVC, KR	*DQ587539, DQ590126, DQ596206, BC070106*
chr6:171,000,000-171,100,000	INTERVC, GWVC	*BC036251*
chr22:16,000,000-16,200,000	VC, INTERKR	*AK022914, LINC00516, BC017398, AK056135*
chr11:100,000-200,000	RAW, SQRTVC, KR, GWKR	*AL137655, LOC100133161, SCGB1C1, ODF3*
chr10:48,000,000-48,200,000	INTERVC	*CTSL1P2, BMS1P6, AGAP9*
chr16:100,000-900,000	RAW	57

For all gene-containing fragments in table 6.4, we used the DAVID Gene Functional Classification and Gene Name Batch Viewer web tools available at https://david.ncifcrf.gov/home.jsp (Huang et al., 2009) in order to find common gene functions. From a total of 110 genes tested, two groups of genes were found to cluster with enrichment scores greater than 1.3 (implying $p < 0.05$ for functional similarity). The first group contained four genes that were found to be functionally similar: *PAEP*, *LCN9*, *LCN1* and *OBP2A* (an enrichment score of 1.36). These genes are predominantly expressed in the lachrymal/salivary glands as well as the nasal cavity (*OBP2A*), the lungs (*OBP2A*) and the endometrium (uterus lining; *PAEP*). They are all part of the lypocalin family, which are proteins that transport hydrophobic molecules.

From the same clustering procedure, we found a group of five genes which shared strong functional similarities: *HBM*, *HBA2*, *HBZ*, *HBA1*, and *HBQ1* (an enrichment score of 4.45). All of these genes are part of the haemoglobin family and their function is to transport oxygen throughout the body via red blood cells. A possible link to schizophrenia here could be the efficacy of this group of genes in regulating oxygen transportation to the brain. This observation is supported by findings that signs of asphyxia at birth are associated with an increased risk of schizophrenia in adults (Dalman et al., 2001).

Table 6.4 contains regions with different properties compared to those in table 6.3. The main difference is which data these nodes have been identified from. Just three of nine from this list have been discovered from the raw *in situ* Hi-C data (with other regions being identified from other normalised Hi-C datasets), whereas all four regions in table 6.3 were identified using raw *in situ* Hi-C. There is also a wide range of region lengths in table 6.4, with the smallest and largest lengths being 100 kb and 1.5 Mb, respectively.

There are numerous ways we could interpret these results. Our algorithm has identified schizophrenia-associated regions almost entirely from raw *in situ* Hi-C data (table 6.3), so one could assume that regions found only in raw data are worth interrogating. This would leave us with three candidates: chr9:137,400,000-138,900,000, chr11:100,000-200,000 and chr16:100,000-900,000. Perhaps the most interesting case is the region on chromosome 11, as it is found not only in raw data, but is also discovered in SQRTVC, KR and GWKR data. Since we see a distinct difference between the outputs from investigations of raw and normalised data, it is reasonable to suggest that further lab-based functional ascertainment of candidate regions is required.

Centrality/Clustering Measures

For our inter-chromosomal Hi-C data thresholded at the median and third quartile (excluding zero entries), we created unweighted networks containing 30,367 nodes and 108,390,265 and 65,685,996 edges, respectively. From these 3D interaction networks, we found a number of regions that appeared more than once as one of the top five most central nodes (i.e. regions which dominate in the communication of a network; figures 6.9 and 6.10).

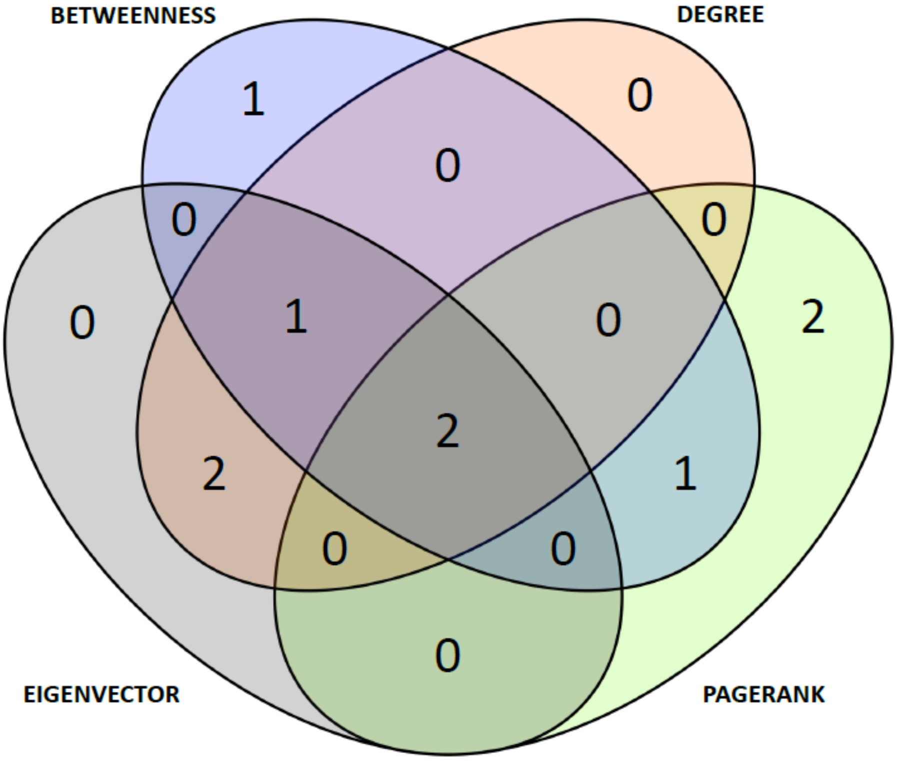


Figure 6.9 Overlap frequency of regions identified as one of the five most central nodes from the inter-chromosomal *in situ* Hi-C network thresholded at the median of non-zero interaction frequencies.

From our network thresholded at the median of non-zero interaction frequencies, we found six distinct regions that had a high centrality score for more than one type of measure (figure 6.9). Two of the six regions had a top five centrality score for all types of measure. Both were gene-poor regions: the first on chromosome 1, positions 121,400,001-121,500,000 and the second on chromosome 6, positions

58,700,001-58,800,000. Despite the lack of known functional genes in these regions, it is possible that either (1) there is additional information yet to be found from these regions, or (2) adjacent genomic fragments that may contain genes/SNPs could be the hidden drivers.

Since betweenness centrality is calculated in a different manner to all other measures (and is thus the least correlated), we were interested by the single region found in the top five for betweenness score and not found from other measures. This was a region on chromosome 6, positions 57,400,001-57,500,000, containing only one gene, *PRIM2*, that encodes for an enzyme that plays a key role in DNA replication.

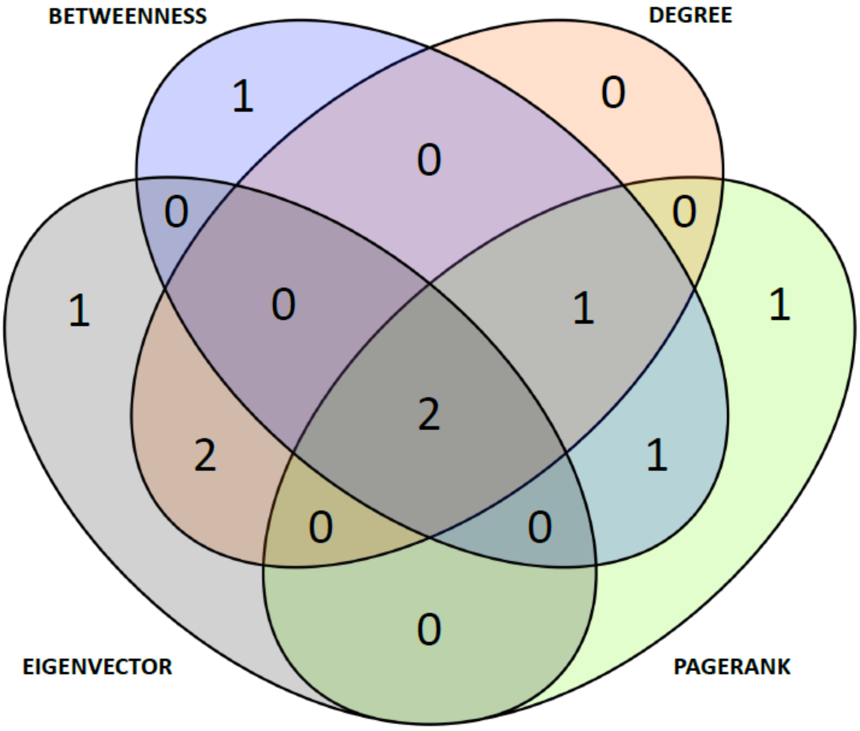


Figure 6.10 Overlap frequency of regions identified as one of the five most central nodes from the inter-chromosomal *in situ* Hi-C network thresholded at the third quartile of non-zero interaction frequencies.

Of the nine unique regions identified in our median thresholded network as being the most central, seven were identical for our network thresholded at the third quartile. Six regions appeared in more than one top five list (figure 6.10), and two regions appeared in all lists – one of which was the same region on chromosome 1 found in the preceding analysis. The other prevalent region was on chromosome 1, positions 145,000,001-145,100,000, which contained five genes. In order for consistent inter-chromosomal contacts to be made, one may conclude that genomic regions must spend most of their time on the periphery of the folded globule. With our centrality measures in mind, we hypothesise that regions found to be central for inter-chromosomal networks must have this property. We also identify these regions as candidates for our subsequent targeted analysis.

From our intra-chromosomal Hi-C network, thresholded at $F > 0$, we found very few overlaps of central nodes. Despite this, we did find an interesting property from our degree and eigenvector centrality top five lists: all of the most central nodes were located on chromosome 2, between positions 85,000,000 and 110,100,000. A similar property was also found from our PageRank list, but for a slightly longer region (chromosome 9, positions 38,000,000-81,100,000). In contrast to central nodes from inter-chromosomal networks, this suggests that intra-chromosomally central nodes are also literally central within the folded globule, i.e. they are found at the interior, where relatively long stretches of DNA are knotted in the centre.

The local clustering coefficient was used to determine the connectedness of immediately neighbouring nodes, with scores ranging from zero (completely sparse) to one (completely connected), in an effort to identify small groups of communicating genomic regions. Results from our intra-chromosomal network were fairly trivial, since the network was small-world (93.75% of nodes had a local clustering coefficient of at least 0.92). Hence, there were no areas of the network that were relatively more connected than anywhere else. For our inter-chromosomal networks, we found that 0.079% and 0.125% of nodes had a local clustering coefficient of at least 0.9 for the median and third quartile thresholds, respectively. Regions with a perfect local clustering coefficient ($C = 1$), corresponding to neighbouring nodes being completely connected, are listed in table 6.5.

Table 6.5 Regions which have the maximum possible local clustering coefficient, from unweighted *in situ* Hi-C networks at both median and third quartile thresholds.

Regions (chromosome:100kb bin)					
1:1210	3:758	7:632	10:480	16:187	23:586
1:1483	5:465	7:1023	10:816	16:331	23:887
2:51	6:619	8:469	11:512	17:348	23:890
2:874	7:572	9:664	14:193	17:446	23:900
2:894	7:617	9:692	14:196	18:155	23:909
2:1071	7:628	9:700	14:201	19:206	23:923
2:1309	7:631	9:1412	16:164	22:207	

Since regions in table 6.5 are each part of a cluster of inter-chromosomal contacts, we can infer a couple of things: (1) in order to form inter-chromosomal 3D interaction clusters, these regions are very likely to lie on the surface of their respective folded chromosomes because this is where they would be most accessible; (2) at the periphery of two (or more) colliding chromosomes, regions participating in clusters are likely to be tangled at these locations for at least a small amount of time. Even brief communications between one-dimensionally remote regions can alter the expression of genes, therefore the clusters we identified are perhaps the focal point of expression variation.

Component Analysis

Using a raw threshold of 80, an INTERKR threshold of 42 and an INTERVC threshold of 41, we investigated the components of the resulting binary networks (figures 6.11, 6.12 and 6.13, respectively) to find a similarity of network architecture (table 6.6).

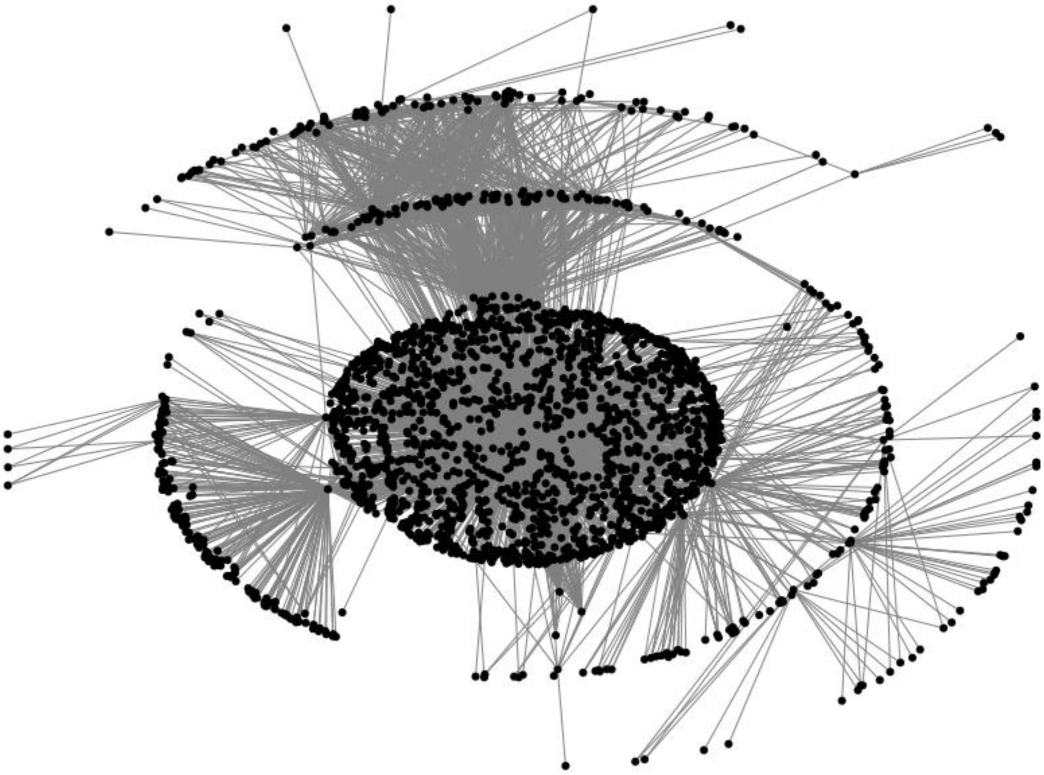


Figure 6.11 Harel-Koren (HK) fast multi-scale layout (Harel & Koren, 2000) of 3D interaction network using raw Hi-C data with interaction frequency threshold F≥80.

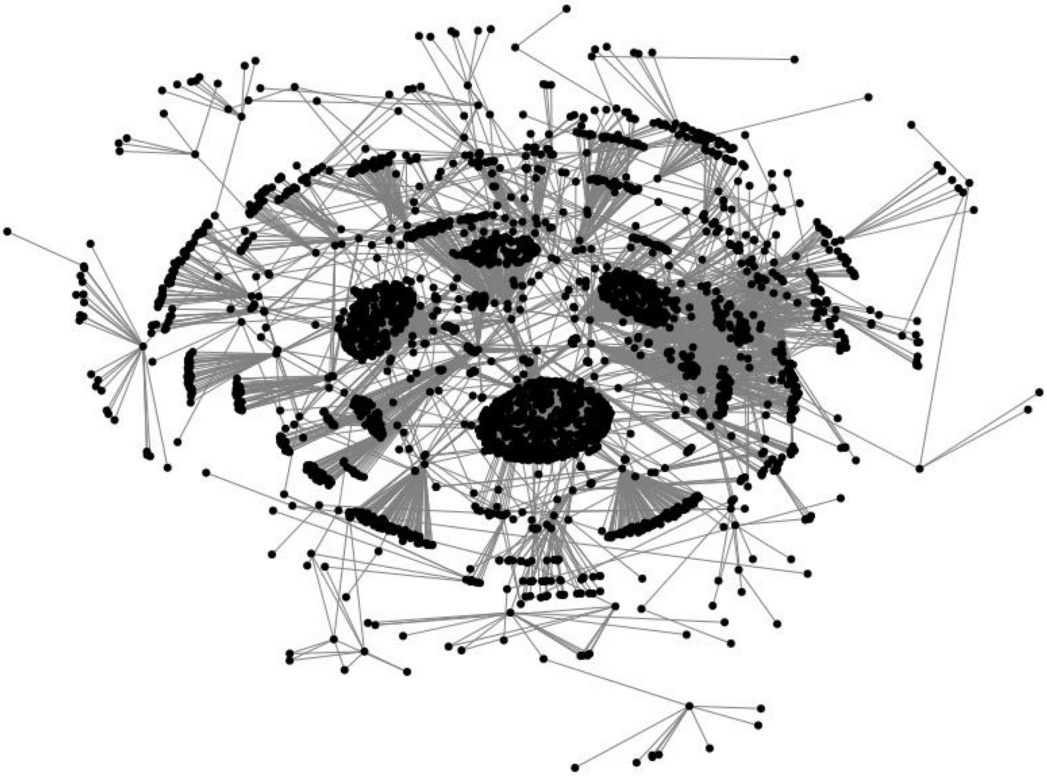


Figure 6.12 HK fast multi-scale layout of 3D interaction network using normalised (INTERKR) Hi-C data with interaction frequency threshold F≥42.

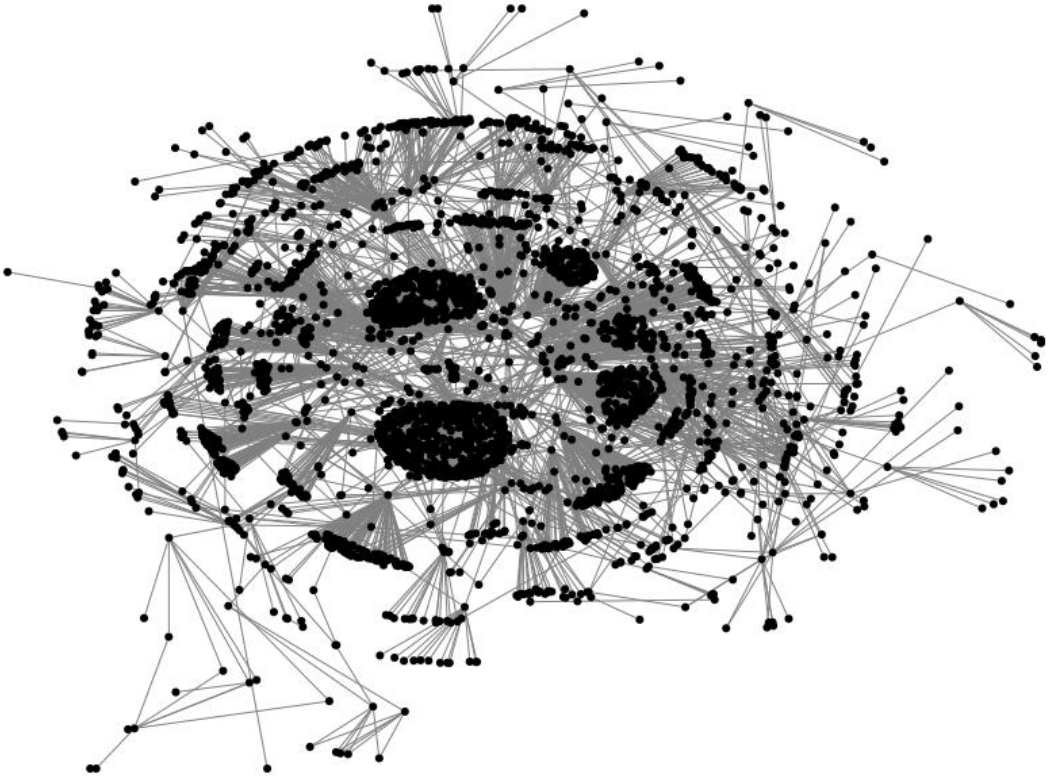


Figure 6.13 HK fast multi-scale layout of 3D interaction network using normalised (INTERVC) Hi-C data with interaction frequency threshold F≥41.

Table 6.6 Overview of binary *in situ* Hi-C networks (all with 30,367 total nodes).

	RAW	INTERKR	INTERVC
Number of edges	4459	4316	4305
Size of largest component	2189	2799	3219
Largest component % of global network	7.21	9.22	10.6
Next three largest component sizes	22, 7, 7	18, 14, 13	24, 15, 12
Number of isolated nodes	28074	27198	26799

All three networks contained a single, predominant component, followed by a string of much smaller components. Since our pool of schizophrenia-associated genes were mostly either isolated nodes or contained within our largest components, we decided to further investigate these components, to see if there was a common subset of genes which interacted in most/all largest components.

From a total of 347 schizophrenia-associated genes, three were found in the largest components of all Hi-C networks (figure 6.14). These were: (1) *HCN1* – a gene known to contribute to spontaneous rhythmic activity in both the heart and brain (Pan et al., 2015), (2) the family of *PCDHA* genes – which are predominantly expressed in the developing nervous system (Sano et al., 1993), and (3) *ZSWIM6* – a mutant allele of which is linked to brain malformation (Smith et al., 2014). Interestingly, all three of these genes are on chromosome 5 (*HCN1*: positions 45,255,052-45,696,220; *ZSWIM6*: positions 60,628,100-60,841,999; *PCDHA*: positions 140,144,410-140,391,929), which could be a signpost that particular locations on this chromosome could have the most profound influence.

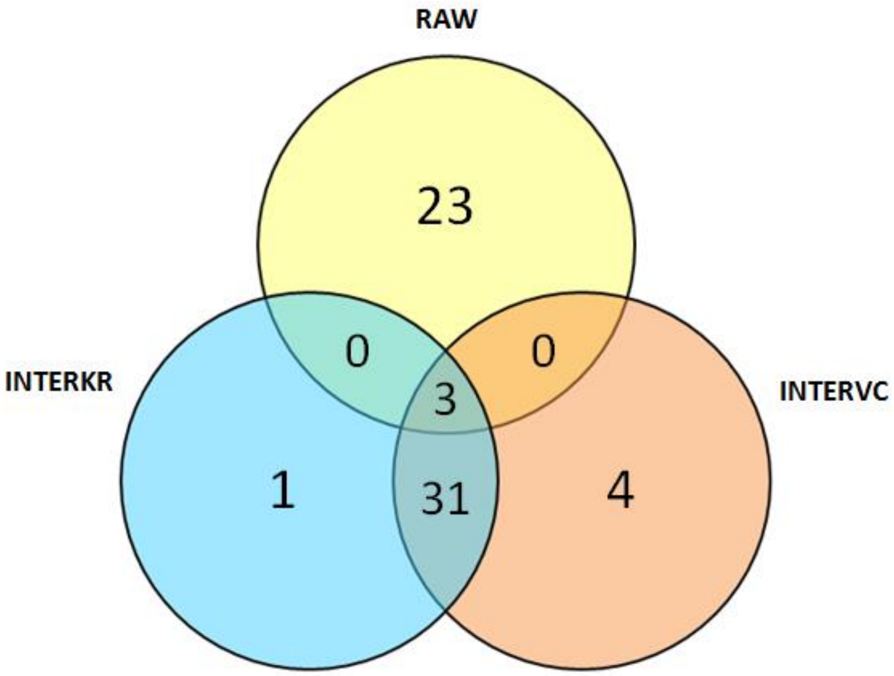


Figure 6.14 The number of schizophrenia-associated genes found in the largest component of our inter-chromosomal binary Hi-C networks.

There was also a large overlap between both normalised networks, with 34 out of 39 genes being located in the largest component (a list of these genes can be found in supplementary table A6.3). Perhaps this is due to similarities between the KR and VC procedures (described in chapter 2). Since there is also very little overlap between processed and unprocessed Hi-C networks, one should be wary of the effect of normalisation procedures and the possibility that genuine interactions are diluted and therefore missed. However, it is also possible that unprocessed Hi-C data contains particular biases that overwrite genuine interactions. We hence preserve genes found from all largest components to avoid the risk of discarding important regions.

After closer inspection of our raw *in situ* Hi-C network, we noticed one node in particular that overwhelmingly acted as the main hub in the largest component, connecting to 1,670 (or 76.33% of) other nodes within the component. This node corresponds to a fragment on chromosome 1, positions 121,400,001-121,500,000, which is a gene-poor region. Further investigation revealed that this fragment was added to the hg19 assembly and was not present in previous reference genomes. We therefore concluded that either (1) this region is poorly understood and is in fact an important candidate region that requires more investigation, or (2) this region is an anomaly, or a false positive, in Hi-C data. We included this region in the pool of candidate regions for our targeted approach, in case (1) was true, but we also decided to remove this region from our raw network and run the component analysis once more, which accounted for the possibility that (2) was true.

With our interaction frequency threshold still set at 80, we found that the largest connected component had significantly reduced in size when the potential anomaly was removed (from 2,189 nodes to 699 nodes; figure 6.15). The number of schizophrenia-associated genes found in the largest component reduced from 26 to 12, none of which overlapped with either largest component from normalised networks, although all 12 were found from our initial raw Hi-C network analysis.

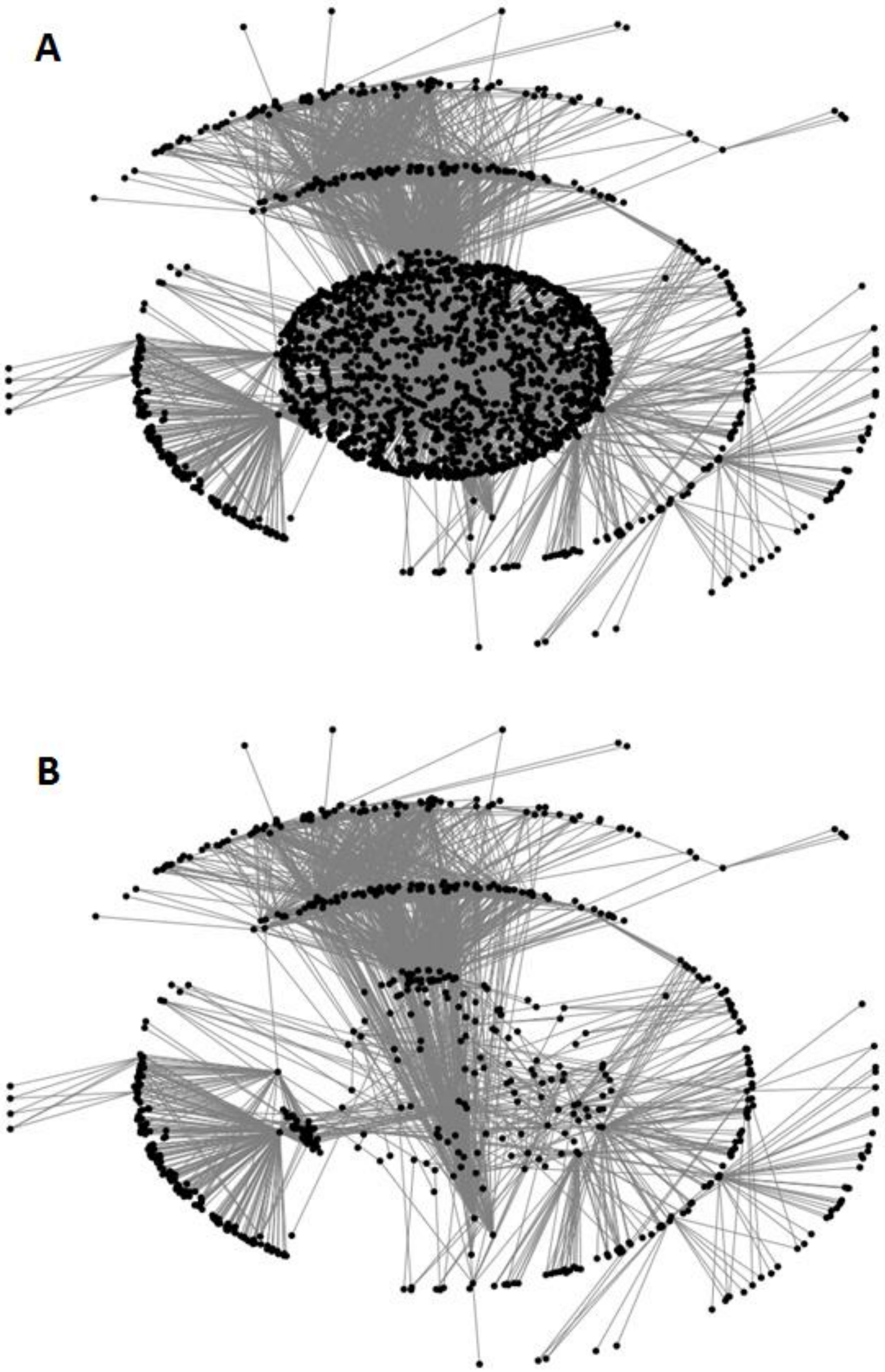


Figure 6.15 Comparison of 3D interaction networks using (A) raw Hi-C data with interaction frequency threshold F≥80 and (B) identical conditions, except a potentially anomalous node corresponding to chromosome 1, bin 1,215 has been removed.

6.4.3. SNP Distributions

From the master list constructed by our algorithm, the candidate nodes from set W were shown to have a highly variable number of SNPs (table 6.7). Of the 35 novel regions found, 22 harboured schizophrenia-associated SNPs within their genomic boundaries. Only one of these 22 regions contained SNPs with genome-wide significance (p-values below the threshold of 5×10^{-8}): the fragment on chromosome 6 between positions 26,000,000-26,500,000. This is the same fragment which contains a number of schizophrenia-associated genes: *HIST1H2BC, HIST1H2BD, HIST1H2BG, HIST1H2BH* (Sanders et al., 2013)*, HIST1H1E* (Föcking et al., 2014) and *BTN3A3* (Chen et al., 2014).

Table 6.7 Genomic regions from network analysis which are ranked in ascending order of minimum p-value (corresponding to the SNP with the lowest p-value found in this region; a small p-value indicates a strong disease association). Regions which contain known schizophrenia-associated genes are highlighted in **bold**.

Node (hg19 position)	Observed number of SNPs	SNP density (SNPs per 100 kb)	Minimum p-value
chr6:26,000,000-26,500,000	391	78.2	1.46×10^{-9}
chr19:300,000-1,400,000	514	46.73	2.98×10^{-5}
chr16:89,600,000-90,100,000	251	50.2	1.11×10^{-4}
chr16:100,000-900,000	386	48.25	3.31×10^{-4}
chr9:137,400,000-138,900,000	922	61.47	1.09×10^{-3}
chr21:15,000,000-15,200,000	26	13	0.013502
chr11:100,000-800,000	324	46.29	0.0137088

Entries in table 6.7 which have a considerably higher SNP density compared to the chromosomal average (table 6.8) are perhaps the most interesting cases. Despite only one node containing SNPs within the genome-wide significance threshold, the sheer number of SNPs could be an indicator for schizophrenia association. One could argue that a high concentration of SNPs which lie just outside of the required significance threshold could have a larger combined effect on disease formation than an individual, significant SNP does, which also ties into the possibility that such SNPs could be in linkage disequilibrium. This conjecture is perhaps supported by the fragment on chromosome 19 between positions 300,000-1,400,000. It has a higher than expected SNP density (46.73, compared to the expected density of 39.37 from table 6.8), there are no individual SNPs which satisfy genome-wide significance (all SNPs satisfy $p > 2.98 \times 10^{-5}$) and yet it contains two genes not found in our original dataset that are known to have an association with schizophrenia: *FGF22* (Terauchi et al., 2010) and *GRIN3B* (Lin et al., 2014).

Table 6.8 SNP density across 22 chromosomes.

Chromosome	Chromosome length (in 100 kb)	Observed number of SNPs	Expected SNP density
1	2493	103559	41.56
2	2432	104572	43.00
3	1980	87436	44.16
4	1911	77631	40.62
5	1810	79438	43.89
6	1711	83730	48.94
7	1592	68689	43.15
8	1464	68130	46.54
9	1412	57968	41.05
10	1356	67121	49.50
11	1350	63688	47.18
12	1339	61944	46.26
13	1152	47170	40.95
14	1073	41313	38.50
15	1026	37926	36.96
16	903	39549	43.80
17	812	34161	42.07
18	781	36805	47.13
19	592	23306	39.37
20	630	32764	52.01
21	482	17735	36.79
22	513	18226	35.53

6.4.4. Gene Expression in Brain Tissues

After probing our master list obtained from network analysis of 327 genes, a total of four promoters corresponding to three genes were identified via high tags per million (TPM) counts in brain tissues (table 6.9). Three of these promoters –

p4@ZNF276, p3@ZNF276 and p3@TCF25 – correspond to two genes (*ZNF276* and *TCF25*) that are neighbours in a single genomic region identified by our network analysis (chromosome 16, positions 89,600,000-90,100,000). *TCF25* is part of a family of transcription factors that are important for embryonic development (Steen & Lindholm, 2008) and is highly expressed in brain tissues (third highest of 27 human tissues tested; Fagerberg et al., 2014). *ZNF276* belongs to the family of zinc finger genes that are responsible for a number of protein functions. Furthermore, a member of the zinc finger family, *ZNF323*, has previously been identified as a risk gene for schizophrenia via brain eQTL and GWAS analysis (Luo et al., 2015). The region containing *TCF25* and *ZNF276* was found to have an incident edge of high weighting in our network analysis (RAW data) and contains a total of 26 genes. There are no genes in this region which have any previous implications in schizophrenia.

Identifying a cluster of promoters corresponding to two genes that are in close linear proximity (within 200 kb) could lead us to hypothesise that there is evidence of digenic activity between *ZNF276* and *TCF25* that contributes to schizophrenia development. This is especially likely, considering that this region contains 251 SNPs, which is higher than the expected SNP density for chromosome 16 (an observed density of 50.2, versus an expected density of 43.8; table 6.7).

Table 6.9 Relative expression (normalised TPM) values across all brain tissue samples (FANTOM5 data) for promoters identified from expression analysis of master list regions.

TSS region	TPM (fetal)	TPM (adult, 18)	TPM (adult, 77/79/81)
p4@ZNF276	0.58	0.87	-
p3@ZNF276	-	0.5	-
p3@TCF25	-	0.43	-
p1@DEAF1	0.94	0.6	0.57

The final TSS identified was p1@DEAF1, a promoter corresponding to the gene *DEAF1* with the highest tag support. *DEAF1* regulates transcription through a protein that contains a zinc finger structure, which perhaps implies an association with the *ZNF* gene family. The encoded protein is also important in the regulation of embryonic development and is very highly expressed in brain tissues (significantly higher expression than 26 other human tissues; Fagerberg et al., 2014). Interestingly, p1@DEAF1 is the only TSS found within our master list regions which has high TPM counts across all three brain tissue samples. The *DEAF1* gene, located in a region on chromosome 11, was found in four *in situ* Hi-C lists from network analysis (RAW, SQRTVC, KR, GWKR; positions 100,000-800,000) which contains a total of 47 genes and 324 SNPs. Perhaps even more interestingly, this region contains six genes with previous schizophrenia implication (*IFITM1*, *IFITM2*, *IFITM3*, *HRAS*, *IRF7*, *DRD4*) – *HRAS* in particular has a high expression in brain tissue (second highest of 27 human tissues; Fagerberg et al., 2014). These findings either (1) solidify the assumption that *DEAF1* may contribute to the development of schizophrenia through a small neighbourhood of interactions, or (2) are a misinterpretation of the involvement of *DEAF1* in schizophrenia; other neighbouring/overlapping genes may be responsible. Indeed, both interpretations are plausible, but further functional ascertainment is required in order to confidently resolve the issue.

The collection of promoters found within this analysis is perhaps too small to make any feasible conclusions about the differences in age between the three brain tissue samples used. Although from the promoters we did find, all four had a high TPM count in the male, 18 year-old sample (and two for the fetal sample), suggesting that the foundations for schizophrenia development are laid at a young adult age, or even as soon as the post-embryonic stage of pregnancy. Recent publications of Hi-C data for the developing brain confirm this (Won et al., 2016).

6.4.5. Enrichment of *cis*-eQTL Pairs in Distinct Brain Regions

Using our global approach (see section 6.4.2), we identified a total of 35 genomic regions which were deemed as potential candidates for schizophrenia association. As part of our subsequent targeted approach (see section 6.4.3 for other results), we found 22 candidate regions that contained a minimum of one SNP from our dataset of schizophrenia-associated SNPs described in section 6.2.2 (Ripke et al., 2011). We continued our targeted approach by using the Braineac web tool – we interrogated all SNPs within our candidate regions with the aim of discovering the presence of *cis*-eQTL pairs. That is, after entering a SNP (or group of SNPs) into Braineac, we hoped to receive an output of affected genes with corresponding p-values which would describe the magnitude of the SNP's influence on gene expression in particular brain regions. The lower the p-value, the greater the SNP's effect on the expression of a target gene. After interrogation of SNPs within these 22 genomic regions, we found that *cis*-eQTL pairs existed within eight distinct regions.

The relative strength of these pairs in terms of gene expression change is visualised in figure 6.16. This shows a heat map of the average expression scores for each genomic region. Vertical blocks of red indicate expression change in particular brain regions, whereas horizontal blocks indicate a particular genomic region that has a high influence on gene expression. Whilst most of the heat map seems fairly homogenous, there are areas which display a relatively high expression score. For example, particular genomic regions seem to have more of an impact on gene expression in the whole brain (such as 6:261-265, 11:2-8 and 16:2-9) than others (such as 6:3-4 and 9:1375-1389). These hot zones also tend to exhibit variable expression levels between distinct brain regions (CRBL having consistently high expression in comparison to other brain regions, for example).

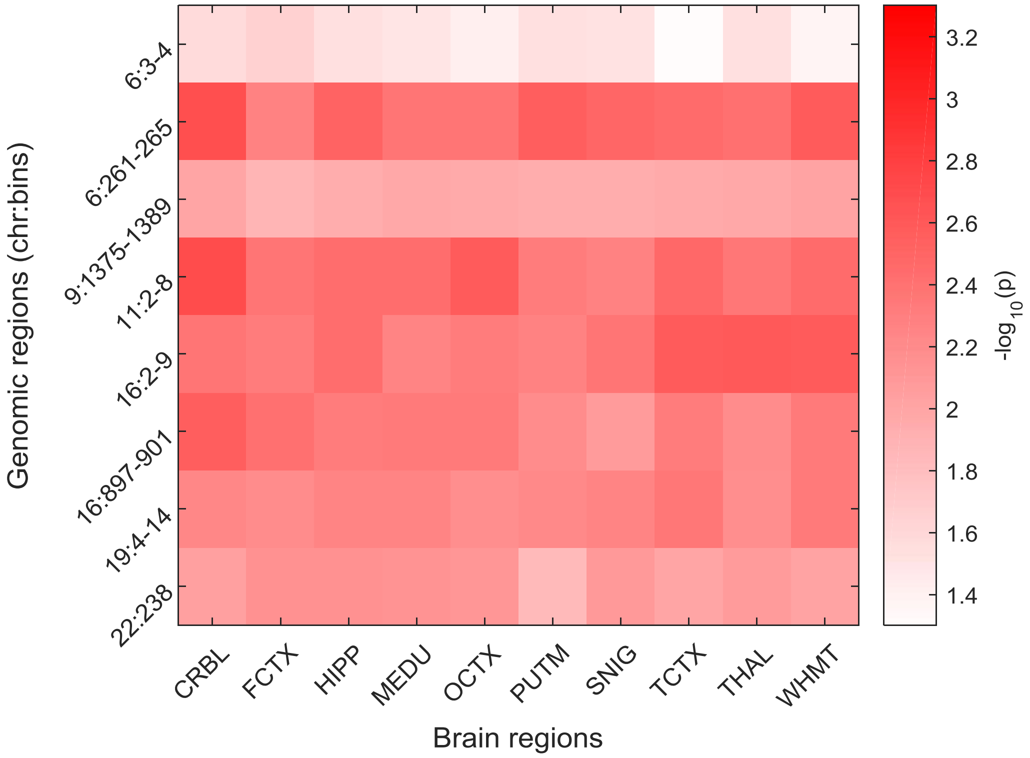


Figure 6.16 Heat map of genomic regions (described in chr:bin form on the y-axis at 100 kb resolution) potentially harbouring *cis*-acting elements influencing expression in ten brain regions (x-axis).

The macro scale of expression scores gives us an indication of where to look in more detail. Therefore, we have identified three areas which exhibit high expression to investigate in greater detail – 6:261-265, 16:2-9 and 11:2-8.

Figure 6.17 is a heat map of expression resulting from individual SNPs on chromosome 6, positions 26,000,001-26,500,000. The aim with such micro scale heat maps is to identify clusters of *cis*-acting SNPs leading to high expression. This is in order to find smaller fragments of the genomic region which affect expression change in corresponding genes, or to identify one or more distinct brain regions where expression change is commonplace. We have identified both of these traits within figure 6.17 by identifying clusters which have expression scores of higher than 5 at each endpoint (and consistently high expression in between). Firstly, we see a block of high expression scores in a number of brain regions, but most pronounced in WHMT and HIPP. The genomic region this spans is positions 26,184,041-26,251,601, therefore we suggest that SNPs found within this boundary are warrant a closer inspection – particularly with regard to their association with schizophrenia development and importantly, whether this region is in linkage disequilibrium (LD). Outside of this cluster, our most expressed region is CRBL at positions 26,351,597-26,405,990. Hence, for this genomic region alone, we suggest that schizophrenia development is most associated with the CRBL region of the brain. But, looking at the bigger picture, a sensible conclusion would be that this region on chromosome 6 is most expressed in WHMT, HIPP and CRBL.

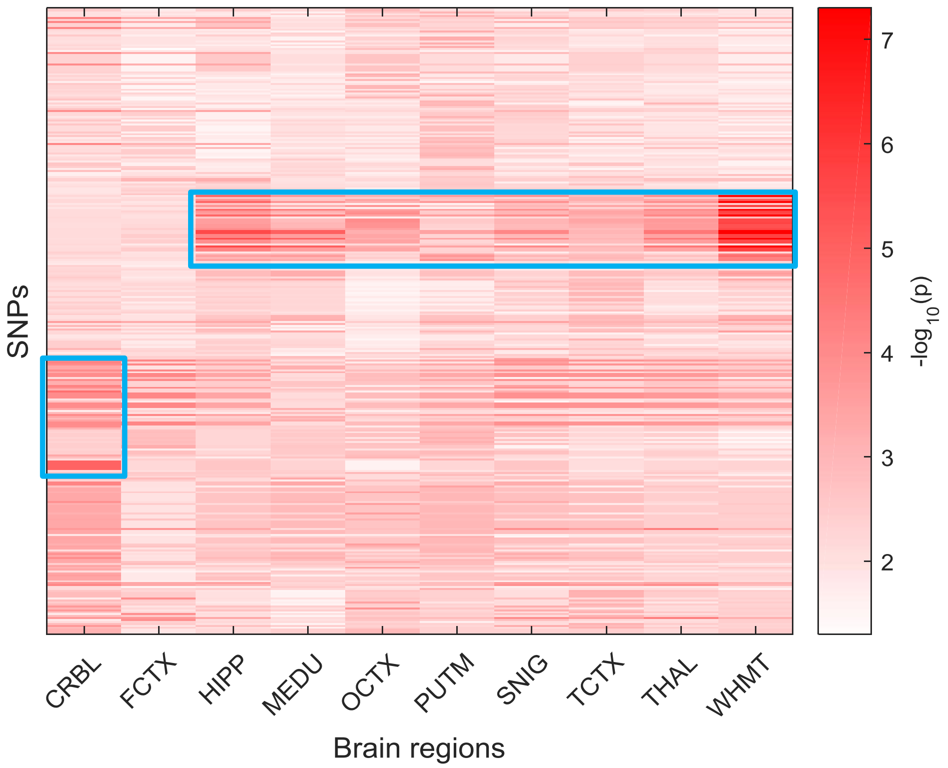


Figure 6.17 Expression heat map from SNPs on chromosome 6, positions 26,000,001-26,500,000. Areas of relatively high expression (satisfying $-\log_{10}(p) > 5$ for boundary endpoints) are enclosed by blue rectangles.

Our second area of interest lies on chromosome 16, positions 100,001-900,000 (figure 6.18). The clusters here are somewhat less pronounced, therefore our threshold of endpoint expression scores was reduced from five to four (corresponding to a p-value of 1×10^{-4}). We found the highest expression scores in small clusters of brain regions TCTX, THAL, WHMT and HIPP and genomic regions covering positions 105,444-176,743 (HIPP and TCTX), 449,162-536,686 (THAL), 709,001-719,933 (WHMT) and 811,433-852,137 (TCTX). We identify SNPs contained within these fragments as being potentially influential in schizophrenia development, but we must first take a closer look and determine whether SNPs are individually associated, and secondly we must find whether this region is in LD.

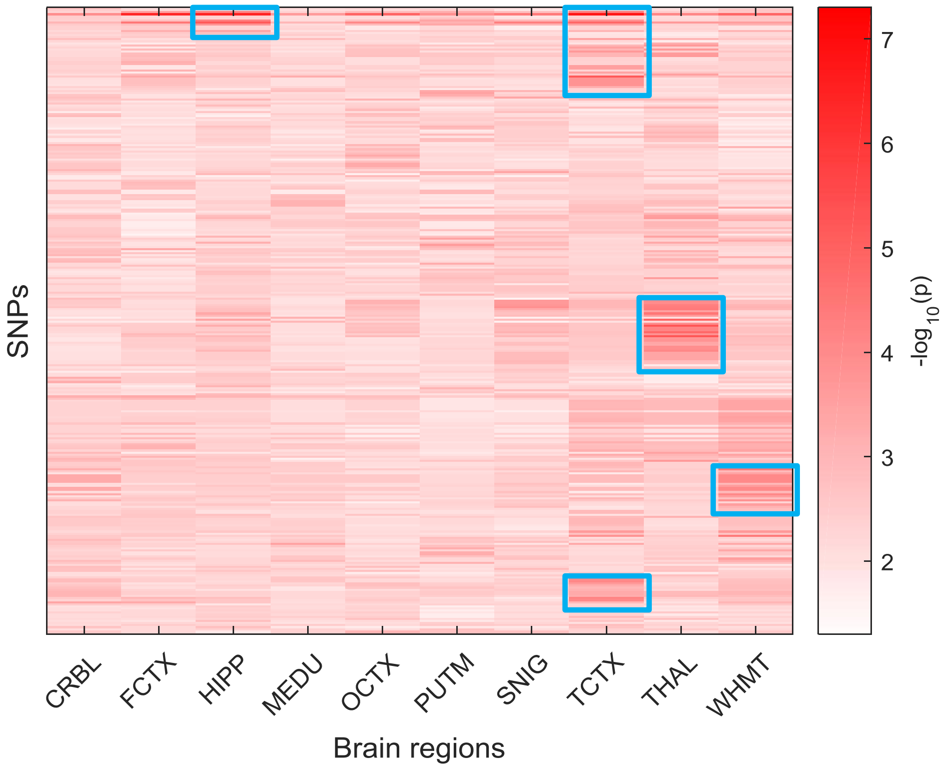


Figure 6.18 Expression heat map from SNPs on chromosome 16, positions 100,001-900,000. Areas of relatively high expression (satisfying $-\log_{10}(p) > 4$ for boundary endpoints) are enclosed by blue rectangles.

Our final area of interest spans positions 100,001-800,000 on chromosome 11. The behaviour of clusters in our heat map for this area (figure 6.19) is slightly different to the preceding heat maps, since they have discontinuous sections of high expression within a small genomic boundary. Although not in solid blocks of red, we identify brain regions CRBL, OCTX and TCTX and genomic positions 641,563-694,257 (CRBL) and 721,570-791,462 (OCTX and TCTX) to be candidates of *cis*-eQTL-containing regions responsible for high expression in schizophrenia.

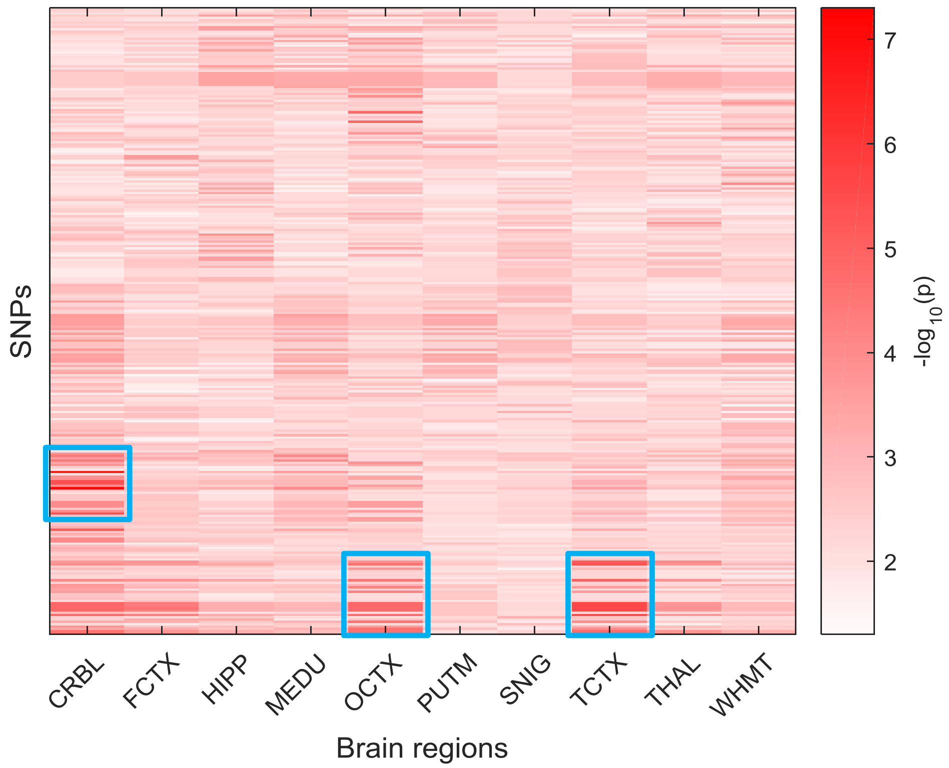


Figure 6.19 Expression heat map from SNPs on chromosome 11, positions 100,001-800,000. Areas of relatively high expression (satisfying $-\log_{10}(p) > 4$ for boundary endpoints) are enclosed by blue rectangles.

After finding SNP-containing fragments within our schizophrenia candidate regions and investigating their effect on expression levels, we wished to determine whether these SNPs had an individual association with schizophrenia, and whether the fragments we found exhibited LD. We therefore used the three regions analysed above (on chromosomes 6, 16 and 11 respectively) as an input into LDlink – a web tool that visualises LD by producing heat maps of R^2 values (Machiela & Chanock, 2015). From our outputs, we were able to hone in on particular regions which had strong LD blocks and subsequently investigate the SNPs lying within these blocks. We hypothesised that SNP-containing fragments which highly influenced gene expression in brain tissues were in LD. We also suggested that within these LD blocks were SNPs with a relatively high individual association with schizophrenia.

Blocks enclosed by blue in figures A6.1, A6.2 and A6.3 contain one or more fragments which are in LD and also SNPs which influence gene expression in the brain. Table 6.10 gives a full description of these regions, including the number of

SNPs found and whether any have a relatively small individual schizophrenia association.

Table 6.10 Genomic fragments in linkage disequilibrium containing SNPs that highly affect gene expression in brain tissues. P-values indicate the strength of association with schizophrenia for an individual SNP.

Fragment (hg19)	Observed number of SNPs	Number of SNPs satisfying p<0.05	Minimum p-value
chr6:26,180,000-26,235,000	39	1	0.0029
chr6:26,352,000-26,410,000	70	13	4.756×10^{-7}
chr16:80,000-200,000	69	2	0.0141
chr16:475,000-575,000	44	1	0.0139
chr16:610,000-785,000	87	0	0.1554
chr16:790,000-880,000	36	9	3.309×10^{-4}
chr11:625,000-725,000	53	0	0.0559
chr11:735,000-810,000	32	0	0.0921

The fragments on chromosome 6, positions 26,352,000-26,410,000 and chromosome 16, positions 790,000-880,000 seem the most interesting, since they have a substantial group of SNPs that have a relatively strong individual association with schizophrenia ($p < 0.05$; 13 and 9 SNPs respectively). Figure 6.20 shows the distribution of p-values (hence schizophrenia association) for all SNPs contained within our regions containing LD blocks.

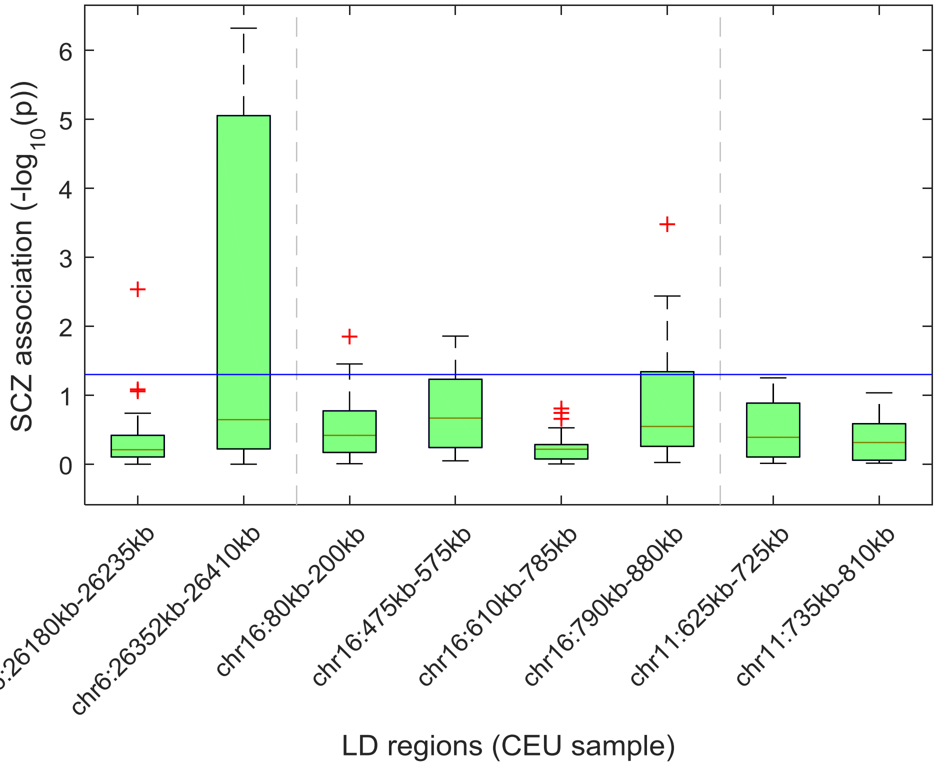


Figure 6.20 Box plots of schizophrenia association levels of SNPs lying within regions of linkage disequilibrium (LD). The web tool LDlink was used to find LD blocks (Machiela & Chanock, 2015) with the CEU population (Utah residents from north and west Europe ancestry). The blue line indicates a p-value threshold of $p = 0.05$. Entries above this line are SNPs which have an association with schizophrenia to a minimum of 95% confidence.

Although there are no SNPs which meet a genome-wide significance threshold of $p < 5 \times 10^{-8}$, we conclude that a network of interactions between SNPs found in these LD regions might play a role in schizophrenia development – particularly those which already have a mild association with the disease (such SNPs are identified in supplementary table A6.4).

Chapter VII

Conclusions

In this thesis we employed a mix of mathematical, computational and statistical approaches in order to test the general hypothesis that linearly remote genomic fragments come into close proximity when folded in the cell nucleus, which consequently affects function (e.g. through regulation of gene expression) and therefore phenotype. Data describing the 3D architecture of the human genome was used to analyse such long-range interactions, in order to reveal whether they regulate various diseases and genetic phenomena.

The many components of the genome, including genes, promoters, enhancers and silencers are well studied, whereas their communication is the area much less well-understood. For this reason, we often represented the 3D genome as a network, where our units were genomic fragments of equal length and interactions were added if pairs were neighbours in 3D space. Thus, fragments potentially containing biological units of expression or regulation (sometimes mutated) acted as nodes and the extent of proximity was modelled by weighted edges. That is, larger weights corresponded to smaller distances between nodes. In some cases, we preprocessed our 3D interaction data in an effort to remove bias from experimental procedures (described in chapter 2).

In this chapter we summarise our results, discuss their implications and finish by suggesting directions for future work.

7.1. Summary

7.1.1. Reduced Penetrance

The challenge of accurately measuring penetrance by detecting variants that genuinely reduce occurrences of expected phenotype is still rife, despite attempts being made in family- and population-based studies. Although the literature acknowledges polygenic activity in complex diseases and cites this as a contributor to the difficulties of identifying penetrance-associated variants (reviewed by Cooper et al., 2013), there is a lack of methods that account for this when analysing data. Typically, variants are investigated independently and are given individual appraisals with regards to their disease associations, whereas in this dissertation we placed the spotlight firmly on polygenic activity by using 3D interaction (Hi-C) data to show that linearly remote regions do cooperate and ultimately alter gene expression.

Our first hypothesis was that gene partners that are known to influence penetrance via digenic activity are non-randomly chosen, based on them sharing a neighbourhood in 3D. We tested for a difference in proportions of Hi-C interaction enrichment between fragments containing genes associated with reduced penetrance and non-associated control regions. No differences were found, suggesting that either the 3D structure of the human genome plays a diluted role in affecting penetrance, or additional regulatory regions are responsible for connecting putative penetrance genes.

As a result of the above findings, we considered that there could be an added layer of complexity regarding the 3D interactions between partner genes. This led us to introduce third-party fragments, suggesting that these unknown fragments could have acted as a connecting bridge between our remote gene pairs. We first hypothesised that our collection of gene pairs associated with reduced penetrance were managed by common third-party regulators. After testing for the presence of fragments consistently having enriched interactions with known penetrance regions, we found a handful of regions satisfying this criteria. However, many of these regions were also found to have enriched interactions for controls. Therefore, one must be wary of the possibility that these fragments ubiquitously interact, rather than interacting specifically with regions associated with reduced penetrance.

Consequently, we then hypothesised that instead of all digenic pairs sharing the same third-party regulators, perhaps each pair was regulated by its own unique third-party fragment. From a dataset of gene partners responsible for influencing disease penetrance, we sought the fragments of most interaction with each gene and looked for any commonality (or intersect) between the lists of highly interacting fragments. There was no significant difference of intersect frequency found between cases and controls, suggesting a lack of evidence for the presence of third-party penetrance regulators.

Whilst our analysis seems inconclusive, it is important to note a couple of methodological shortcomings that may have inhibited our ability to obtain results. Firstly, our intersect analysis focused on the number of intersects, rather than which regions were intersecting. A qualitative method, focused on identifying the locations of intersecting regions is perhaps suitable here as a follow-up approach – this would bring the added benefit of being able to recognise and deal with false positives. Secondly, the database of genes known to affect penetrance is perhaps too small. Repeating our quantitative analysis with a larger and more reliable set of genes (when it becomes available) could give way to a better understanding of the role of the 3D genome structure in penetrance.

7.1.2. Gene Fusion Events

Next generation sequencing (NGS) methods provide an effective means of detecting fusion events in hindsight. That is, signatures of translocations, deletions or inversions are identified from sequence data of diseased patients and variants common in these individuals are suggested to be the catalyst for fusion events. Whilst this method does partially address what causes gene fusion events (although the underlying mechanisms are not yet fully understood), it does not address why. We were therefore motivated in chapter 5 to investigate why fusions occur by observing the 3D behaviour of known fusion gene partners in healthy cell lines (i.e. a *foresight* method).

Our general hypothesis was that the occurrence of gene fusion events is governed by the 3D proximity of fusion partners. Moreover, we suggested that a required condition for genes to fuse after a translocation/deletion/inversion is the 3D neighbouring of breakpoint sites. Hence, our first specific hypothesis was that genomic fragments that harboured a known fusion gene were enriched in interactions from Hi-C data. We tested this using dilution and *in situ* Hi-C data at 1 Mb and 100 kb resolutions and found that in most cases, there was a significant enrichment of Hi-C interactions in cases versus controls. Furthermore, the difference in cases and controls was huge for inter-chromosomal data. This was perhaps because intra-chromosomal data suffers from a one-dimensional proximity bias that leads to a higher mean and lower variance for intra-chromosomal interaction frequencies, thus resulting in smaller differences.

We then hypothesised that the choice of fusion partner was non-random; a head gene would fuse to a tail gene as a result of the two partners sharing the same spatial domain. We tested each fusion pair by determining whether they were relatively *local* or *remote*, which we defined by setting an interaction frequency rank threshold. At various threshold choices, we found that the number of fusion pairs deemed local was significantly more than the number of local controls. These findings support our hypothesis of finding evidence for 3D neighbourhoods of fusion genes in healthy cell lines. Hence, the 3D structure of the human genome is either a catalyst for fusion events, or provides the ideal conditions for fusions to occur.

7.1.3. Schizophrenia

Initial attempts to individually identify single nucleotide polymorphisms (SNPs) associated with schizophrenia via genome-wide association studies (GWAS) have yielded relative success (Ripke et al., 2011, 2013), and subsequent studies have acknowledged the complex nature of schizophrenia by calculating the polygenic risk score (PRS) of SNPs (Euesden et al., 2014). Pathway- and set-based analyses are methods that also account for polygenic activity – they are categorise SNPs and transcription start sites (TSSs) into distinct biological processes or functions, therefore grouping units of perceived similarity (de Leeuw et al., 2015). These methods for detecting influential SNPs assume, however, that SNPs primarily affect the expression of the nearest downstream gene. Moreover, SNPs are combined if they are within a certain distance from the gene (such as <100 kb or <1 Mb), despite some regulatory elements (such as enhancers) occurring at a distance beyond 1 Mb. These initial assumptions are not always correct, and have resulted in the incorrect SNP-to-gene pairings of approximately 86% of all cases (Mumbach et al., 2017). In chapter 6, we therefore proposed an alternative method that resolves this issue by incorporating Hi-C data to identify the true, longer-range target genes and therefore more accurately predict schizophrenia associated genomic fragments.

At the time of study, we had access to a limited number of Hi-C cell lines – this did not include any form of brain tissue. Since schizophrenia is a disorder of the brain, we first wished to select Hi-C data that best emulated cell behaviour in the brain. We achieved this by finding overlaps between eQTL pairs (SNPs which affect the expression of corresponding genes in brain tissues; Ramasamy et al., 2014) and 3D interactions in our available Hi-C data (Rao et al., 2014). From three available Hi-C cell lines: GM12878 (lymphoblastoid), HMEC (mammary epithelial) and IMR90 (lung fibroblast), we found that the blood cell line (GM12878) was the best emulator of brain tissues, with 73% overlap between eQTL pairs and Hi-C interactions, compared to 23.8% and 6.6% for IMR90 and HMEC, respectively.

After choosing the most suitable Hi-C data to emulate brain cells, we started work on our aim to identify novel genomic regions that influence schizophrenia via long-range interactions. Generally speaking, our analysis consisted of two parts. The first part was termed the *global approach* – we constructed networks of Hi-C interactions and sought out nodes of highest importance through various network measures (described in chapter 3). Nodes from our Hi-C networks that exhibited significantly higher interaction frequencies with our extended gene regions (EGRs), as well as nodes that had high centrality scores were classified as schizophrenia *candidate regions*. We were able to construct a database of candidate regions corresponding to important nodes from these networks – the majority of which were gene-rich fragments (positions listed in table 6.4). Furthermore, our approach found four candidate regions containing a total of 15 genes independent from the original dataset. Interestingly, these genes were found to be associated with schizophrenia through a subsequent literature search (positions listed in table 6.3). Our methodology was therefore validated, since we were not only finding unique candidate regions independently, but we had also found gene-containing regions already associated with schizophrenia.

We also investigated the connected components of raw and normalised Hi-C networks and found common regions contained within each of the largest components that shared connections with schizophrenia genes from our original dataset. These were added to our dataset of candidate regions. Thus, our global approach indiscriminately identified novel schizophrenia candidate regions (some of which were found via independent literature to be associated with schizophrenia), ready for further analysis.

Once we had collected our database of candidate regions, we implemented the second part of our analysis: the targeted approach. In essence, we wished to solidify claims that our candidate regions were indeed associated with schizophrenia by running each region through a series of checks. Those that satisfied most, or all, of the checks were consequently deemed the strongest candidates for schizophrenia association. We found that of all candidate regions analysed, most satisfied at least one check and many satisfied numerous checks. Our first check showed that 22 candidate regions contained at least one schizophrenia associated SNP within its boundaries, with most containing more SNPs than expected, assuming SNPs were uniformly distributed throughout each chromosome. Secondly, we found groups of genes from novel candidate regions that had functional similarities. Perhaps the most interesting group was a haemoglobin gene family that is responsible for transporting oxygen to the brain. It is clear that any change in the expression of these genes could easily affect the health of the brain and therefore possibly induce schizophrenia. Thirdly, four promoters affecting the expression of three genes in brain tissue samples of different ages were found using FANTOM5 data. All four promoters resided on one of two particular candidate regions. We also found that the four promoters all affected expression in the young adult brain tissue sample, confirming that schizophrenia is a disorder of the developing brain. Our final check was to utilise eQTL data once more to investigate the expression of target genes in specific brain regions. eQTL pairs in eight of our candidate regions were found, with many exhibiting a large variation in expression between brain regions. On a micro-scale, we discovered that some linear sequences of SNPs were in clusters displaying high expression change. Further analysis using LDlink revealed that some of these SNP clusters were in linkage disequilibrium, whilst some SNPs were shown via GWAS to have a high individual association with schizophrenia.

Each of the above targeted approaches revealed a number of candidate regions that exhibited the expected characteristics for schizophrenia association. We have explicitly stated these regions throughout chapter 6, and we consequently suggest that the most mentioned novel candidate regions (corresponding to the most checks satisfied) are the most likely to influence schizophrenia, although further functional ascertainment is required.

7.2. Outlook

Historically, approaches have tended to isolate one factor attributed to a disease or genetic phenomenon and independently test the strength of its association. More recently, however, studies have acknowledged that aberrant traits can occur as a result of the cooperation of many genetic components, often at seemingly remote distances from one another. The term *polygenic* best describes the nature of complex disorders, although the underlying causes of many polygenic diseases are yet to be elucidated.

This thesis has continued the recent trend of inclusive analysis by incorporating prior knowledge and datasets to the relatively new concept of functional change induced by the 3D genome architecture. We have effectively added a layer to existing knowledge by transforming what were once studies of the one-dimensional genome into contextually relevant, three-dimensional methods. In particular, we have investigated the extent to which long-range interactions of genomic elements influence genetic phenomena, and have even predicted novel genomic regions that are likely to affect the complex brain disorder, schizophrenia.

One key limitation of our studies was the lack of cell diversity from available Hi-C libraries. Therefore, a direction for future work would certainly include the use of cell-specific 3D interaction data, such as brain Hi-C data for schizophrenia studies which became available recently (Won et al., 2016). Since this thesis centres on the application of an overarching method to different case studies, it follows that a plethora of routes for future work in cell-specific disorders will emerge with Hi-C availability of corresponding cell lines. For example, future prediction and identification techniques for schizophrenia would be conceivably much more accurate and reliable with brain Hi-C as opposed to blood Hi-C data, despite the high overlap shown in our analysis. As for the use of new data, the methodological foundations have been laid in this thesis: the designed algorithms require little editing to be effective with new Hi-C libraries and are only limited by cell line availability of such data. Experimental Hi-C methodologies and the subsequent removal of biases via normalisation techniques is also a fast-moving area, therefore future Hi-C libraries are likely to be more comprehensive, have better fragment resolutions and be more reliable. It is expected that the methods seen in this thesis will branch out into a broader range of case studies, given the right data. Revisions or improvements of bioinformatics tools used in our subsequent targeted analysis also promise to benefit the prediction/identification methodology.

We also envisage future work continuing to aggregate methods and results published from various literatures, in an attempt to include as much knowledge as possible to validate assertions about genetic variants and disease. In particular, the inclusion of network theory in various analyses of cell biology has changed the way we think about genetic components – especially their communications. We propose that Hi-C networks could form part of a greater multi-layered network (including protein-protein interaction networks, regulatory networks and disease gene networks), where connections can describe more than one type of communication. Although seemingly complex, a holistic multi-layered network such as this could be key to revealing previously overlooked properties.

Bibliography

Abugessaisa, I., Shimoji, H., Sahin, S., Kondo, A., Harshbarger, J., Lizio, M., Hayashizaki, Y., Carninci, P., Forrest, A. and Kasukawa, T., 2016. FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database,* 2016.

Abyzov, A., Li, S., Kim, D.R., Mohiyuddin, M., Stutz, A.M., Parrish, N.F., Mu, X.J., Clark, W., Chen, K., Hurles, M., Korbel, J.O., Lam, H.Y., Lee, C. and Gerstein, M.B., 2015. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications,* 6, 7256.

Adeva, M., El-Youssef, M., Rossetti, S., Kamath, P.S., Kubly, V., Consugar, M.B., Milliner, D.M., King, B.F., Torres, V.E. and Harris, P.C., 2006. Clinical and molecular characterization defines a broadened spectrum of autosomal recessive polycystic kidney disease (ARPKD). *Medicine,* 85 (1), 1-21.

Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A., 2008. McKusick's online Mendelian inheritance in man (OMIM®). *Nucleic Acids Research,* 37 (suppl_1), D793-D796.

Antoniou, A.C., Sinilnikova, O.M., Simard, J., Léoné, M., Dumont, M., Neuhausen, S.L., Struewing, J.P., Stoppa-Lyonnet, D., Barjhoux, L. and Hughes, D.J., 2007. RAD51 135G→ C modifies breast cancer risk among BRCA2 mutation carriers: results from a combined analysis of 19 studies. *The American Journal of Human Genetics,* 81 (6), 1186-1200.

Ashley-Koch, A., Yang, Q. and Olney, R.S., 2000. Sickle hemoglobin (Hb S) allele and sickle cell disease: a HuGE review. *American Journal of Epidemiology,* 151 (9), 839-845.

Badano, J.L., and Katsanis, N., 2002. Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews Genetics,* 3 (10), 779-789.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R. and Wooster, R., 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer,* 91 (2), 355-358.

Barutcu, A.R., Fritz, A.J., Zaidi, S.K., van Wijnen, A.J., Lian, J.B., Stein, J.L., Nickerson, J.A., Imbalzano, A.N. and Stein, G.S., 2016. C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. *Journal of Cellular Physiology,* 231 (1), 31-35.

Bass, A.J., Lawrence, M.S., Brace, L.E., Ramos, A.H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G. and Sivachenko, A., 2011. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature Genetics,* 43 (10), 964-968.

Berge, K., Haugaa, K., Früh, A., Anfinsen, O., Gjesdal, K., Siem, G., Øyen, N., Greve, G., Carlsson, A. and Rognum, T., 2008. Molecular genetic analysis of long QT syndrome in Norway indicating a high prevalence of heterozygous mutation carriers. *Scandinavian Journal of Clinical and Laboratory Investigation,* 68 (5), 362-368.

Beutler, E., 2003. The HFE Cys282Tyr mutation as a necessary but not sufficient cause of clinical hereditary hemochromatosis. *Blood,* 101 (9), 3347-3350.

Bolland, J.M., 1988. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. *Social Networks,* 10 (3), 233-253.

Boyle, E.A., Li, Y.I. and Pritchard, J.K., 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell,* 169 (7), 1177-1186.

Bruno, A.E., Miecznikowski, J.C., Qin, M., Wang, J. and Liu, S., 2013. FUSIM: a software tool for simulating fusion transcripts. *BMC Bioinformatics,* 14 (1), 13.

Buchanan, M., 2010. *Networks in cell biology.* Cambridge University Press.

Bullard, J.E., and Nogee, L.M., 2007. Heterozygosity for ABCA3 mutations modifies the severity of lung disease associated with a surfactant protein C gene (SFTPC) mutation. *Pediatric Research,* 62 (2), 176-179.

Bullmore, E., and Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience,* 10 (3), 186-198.

Bush, W.S., and Moore, J.H., 2012. Genome-wide association studies. *PLoS Computational Biology,* 8 (12), e1002822.

Cancer Genome Atlas Research Network, 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature,* 455 (7216), 1061-1068.

Cardno, A.G., and Gottesman, I.I., 2000. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *American Journal of Medical Genetics Part A,* 97 (1), 12-17.

Chen, C., Zhang, C., Cheng, L., Reilly, J.L., Bishop, J.R., Sweeney, J.A., Chen, H., Gershon, E.S. and Liu, C., 2014. Correlation between DNA methylation and gene expression in the brains of patients with bipolar disorder and schizophrenia. *Bipolar Disorders,* 16 (8), 790-799.

Chmielecki, J., Peifer, M., Jia, P., Socci, N.D., Hutchinson, K., Viale, A., Zhao, Z., Thomas, R.K. and Pao, W., 2010. Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer. *Nucleic Acids Research,* 38 (20), 6985-6996.

Clarke, G.M., Anderson, C.A., Pettersson, F.H., Cardon, L.R., Morris, A.P. and Zondervan, K.T., 2011. Basic statistical analysis in genetic case-control studies. *Nature Protocols,* 6 (2), 121-133.

Coleman, W.B., and Tsongalis, G.J., 2010. *Essential concepts in molecular pathology.* Academic Press.

Cooper, D.N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. and Kehrer-Sawatzki, H., 2013. Where genotype is not predictive of phenotype: Towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics,* 132 (10), 1077-1130.

Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L. and Pericak-Vance, M.A., 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science (New York, N.Y.),* 261 (5123), 921-923.

Costenbader, E., and Valente, T.W., 2003. The stability of centrality measures when networks are sampled. *Social Networks,* 25 (4), 283-307.

Dächsel, J.C., Mata, I.F., Ross, O.A., Taylor, J.P., Lincoln, S.J., Hinkle, K.M., Huerta, C., Ribacoba, R., Blazquez, M. and Alvarez, V., 2006. Digenic parkinsonism: investigation of the synergistic effects of PRKN and LRRK2. *Neuroscience Letters,* 410 (2), 80-84.

Dahm, R., 2008. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics,* 122 (6), 565-581.

Dalman, C., Thomas, H.V., David, A.S., Gentz, J., Lewis, G. and Allebeck, P., 2001. Signs of asphyxia at birth and risk of schizophrenia. Population-based case-control study. *The British Journal of Psychiatry : The Journal of Mental Science,* 179, 403-408.

de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D., 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology,* 11 (4), e1004219.

de Wit, E., and de Laat, W., 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development,* 26 (1), 11-24.

Dedoussis, G., Luo, Y., Starremans, P., Rossetti, S., Ramos, A., Cantiello, H., Katsareli, E., Ziroyannis, P., Lamnissou, K. and Harris, P.C., 2008. Co-inheritance of a PKD1 mutation and homozygous PKD2 variant: a potential modifier in autosomal dominant polycystic kidney disease. *European Journal of Clinical Investigation,* 38 (3), 180-190.

Dekker, J., 2006. The three'C's of chromosome conformation capture: controls, controls, controls. *Nature Methods,* 3 (1), 17.

Dekker, J., 2008. Gene regulation in the third dimension. *Science (New York, N.Y.),* 319 (5871), 1793-1794.

Dobson, A.J., and Barnett, A., 2008. *An introduction to generalized linear models.* CRC press.

Dong, X., and Weng, Z., 2013. The correlation between histone modifications and gene expression.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., Green, R.D. and Dekker, J., 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research,* 16 (10), 1299-1309.

Drumm, M.L., Konstan, M.W., Schluchter, M.D., Handler, A., Pace, R., Zou, F., Zariwala, M., Fargo, D., Xu, A. and Dunn, J.M., 2005. Genetic modifiers of lung disease in cystic fibrosis. *New England Journal of Medicine,* 353 (14), 1443-1453.

Eaton, W.W., Martins, S.S., Nestadt, G., Bienvenu, O.J., Clarke, D. and Alexandre, P., 2008. The burden of mental disorders. *Epidemiologic Reviews,* 30 (1), 1-14.

Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I.H., Nyberg, S., Wolf, M. and Borresen-Dale, A., 2011. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology,* 12 (1), R6.

Edwards, P.A., 2010. Fusion genes and chromosome translocations in the common epithelial cancers. *The Journal of Pathology,* 220 (2), 244-254.

Eeles, R.A., Kote-Jarai, Z., Giles, G.G., Al Olama, A.A., Guy, M., Jugurnauth, S.K., Mulholland, S., Leongamornlert, D.A., Edwards, S.M. and Morrison, J., 2008. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics,* 40 (3), 316-321.

Elgar, G., and Vavouri, T., 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics,* 24 (7), 344-352.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489 (7414), 57-74.

Estrada, E., 2012. *The structure of complex networks: theory and applications.* Oxford University Press.

Euesden, J., Lewis, C.M. and O'Reilly, P.F., 2014. PRSice: polygenic risk score software. *Bioinformatics,* 31 (9), 1466-1468.

Fagerberg, L., Hallstrom, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjostedt, E., Lundberg, E., Szigyarto, C.A., Skogs, M., Takanen, J.O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J.M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, A., von Feilitzen, K., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F. and Uhlen, M., 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics : MCP,* 13 (2), 397-406.

Fairley, C., Zimran, A., Phillips, M., Cizmarik, M., Yee, J., Weinreb, N. and Packman, S., 2008. Phenotypic heterogeneity of N370S homozygotes with type I Gaucher disease: an analysis of 798 patients from the ICGG Gaucher Registry. *Journal of Inherited Metabolic Disease,* 31 (6), 738-744.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), et al.,2014. A promoter-level mammalian expression atlas. *Nature,* 507 (7493), 462-470.

Ferraiuolo, M.A., Sanyal, A., Naumova, N., Dekker, J. and Dostie, J., 2012. From cells to chromatin: capturing snapshots of genome organization with 5C technology. *Methods,* 58 (3), 255-267.

Föcking, M., Lopez, L., English, J., Dicker, P., Wolff, A., Brindley, E., Wynne, K., Cagney, G. and Cotter, D., 2015. Proteomic and genomic evidence implicates the postsynaptic density in schizophrenia. *Molecular Psychiatry,* 20 (4), 424-432.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E. and Ponting, L., 2016. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research,* 45 (D1), D777-D783.

Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C. and Ward, S., 2014. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research,* 43 (D1), D805-D811.

Ford, D., Easton, D., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D., Weber, B., Lenoir, G. and Chang-Claude, J., 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *The American Journal of Human Genetics,* 62 (3), 676-689.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. *Sociometry,* , 35-41.

Freeman, L.C., Roeder, D. and Mulholland, R.R., 1979. Centrality in social networks: II. Experimental results. *Social Networks,* 2 (2), 119-141.

Frenkel-Morgenstern, M., Gorohovski, A., Lacroix, V., Rogers, M., Ibanez, K., Boullosa, C., Andres Leon, E., Ben-Hur, A. and Valencia, A., 2012. ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Research,* 41 (D1), D142-D151.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R., 2004. A census of human cancer genes. *Nature Reviews Cancer,* 4 (3), 177-183.

Galetzka, D., Hansmann, T., El Hajj, N., Weis, E., Irmscher, B., Ludwig, M., Schneider-Rätzke, B., Kohlschmidt, N., Beyer, V. and Bartsch, O., 2012. Monozygotic twins discordant for constitutive BRCA1 promoter methylation, childhood cancer and secondary cancer. *Epigenetics,* 7 (1), 47-54.

Gandhi, M., Medvedovic, M., Stringer, J. and Nikiforov, Y., 2006. Interphase chromosome folding determines spatial proximity of genes participating in carcinogenic RET/PTC rearrangements. *Oncogene,* 25 (16), 2360.

Gan-Or, Z., Bar-Shira, A., Gurevich, T., Giladi, N. and Orr-Urtreger, A., 2011. Homozygosity for the MTX1 c. 184T> A (p. S63T) alteration modifies the age of onset in GBA-associated Parkinson's disease. *Neurogenetics,* 12 (4), 325-332.

Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A. and Cruciat, C., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* 415 (6868), 141-147.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A.L., 2007. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America,* 104 (21), 8685-8690.

Gong, G., Hannon, N. and Whittemore, A.S., 2010. Estimating gene penetrance from family data. *Genetic Epidemiology,* 34 (4), 373-381.

Goodwin, S., McPherson, J.D. and McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17 (6), 333-351.

Green, N.M., 1975. Avidin. *Advances in Protein Chemistry,* 29, 85-133.

Grossmann, V., Kohlmann, A., Klein, H., Schindela, S., Schnittger, S., Dicker, F., Dugas, M., Kern, W., Haferlach, T. and Haferlach, C., 2011. Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure. *Leukemia,* 25 (4), 671-680.

Ha, K.C., Lalonde, E., Li, L., Cavallone, L., Natrajan, R., Lambros, M.B., Mitsopoulos, C., Hakas, J., Kozarewa, I. and Fenwick, K., 2011. Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC Medical Genomics,* 4 (1), 75.

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T.M., Troakes, C., Turecki, G., O'Donovan, M.C., Schalkwyk, L.C., Bray, N.J. and Mill, J., 2016. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience,* 19 (1), 48-54.

Harel, D., and Koren, Y., 2000. A fast multi-scale method for drawing large graphs. *In: Proceedings of the working conference on Advanced visual interfaces,* ACM, pp. 282-285.

Havlin, S., Kenett, D.Y., Ben-Jacob, E., Bunde, A., Cohen, R., Hermann, H., Kantelhardt, J., Kertész, J., Kirkpatrick, S. and Kurths, J., 2012. Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics,* 214, 273-293.

He, P., Lei, X., Yuan, D., Zhu, Z. and Huang, S., 2017. Accumulation of minor alleles and risk prediction in schizophrenia. *Scientific Reports,* 7 (1), 11661-017-12104-0.

Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C., Craddock, N. and Wellcome Trust Case-Control Consortium, 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *The American Journal of Human Genetics,* 85 (1), 13-24.

Holmans, P., Moskvina, V., Jones, L., Sharma, M., International Parkinson's Disease Genomics Consortium (IPDGC), Vedernikov, A., Buchel, F., Sadd, M., Bras, J.M. and Bettella, F., 2012. A pathway-based analysis provides additional support for an immune-related genetic susceptibility to Parkinson's disease. *Human Molecular Genetics,* 22 (5), 1039-1049.

Huang, D.W., Sherman, B.T. and Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols,* 4 (1), 44.

Hunter, T., 2007. Treatment for chronic myelogenous leukemia: the long road to imatinib. *The Journal of Clinical Investigation,* 117 (8), 2036-2043.

International HapMap Consortium, 2005. A haplotype map of the human genome. *Nature,* 437 (7063), 1299-1320.

International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F. and Sklar, P., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature,* 460 (7256), 748-752.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabási, A., 2000. The large-scale organization of metabolic networks. *Nature,* 407 (6804), 651-654.

Jia, P., Wang, L., Meltzer, H.Y. and Zhao, Z., 2010. Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizophrenia Research,* 122 (1), 38-42.

Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C.O., Carninci, P., Forrest, A.R. and Hayashizaki, Y., 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Research,* 21 (7), 1150-1159.

Kass, S.U., Pruss, D. and Wolffe, A.P., 1997. How does DNA methylation repress transcription? *Trends in Genetics,* 13 (11), 444-449.

Keller, M.C., Simonson, M.A., Ripke, S., Neale, B.M., Gejman, P.V., Howrigan, D.P., Lee, S.H., Lencz, T., Levinson, D.F. and Sullivan, P.F., 2012. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genetics,* 8 (4), e1002656.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D., 2002. The human genome browser at UCSC. *Genome Research,* 12 (6), 996-1006.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. and Tsui, L., 1989. Identification of the cystic fibrosis gene: genetic analysis. *Trends in Genetics,* 5, 363-363.

Kim, N., Kim, P., Nam, S., Shin, S. and Lee, S., 2006. ChimerDB—a knowledgebase for fusion sequences. *Nucleic Acids Research,* 34 (suppl_1), D21-D24.

Kim, P., Yoon, S., Kim, N., Lee, S., Ko, M., Lee, H., Kang, H., Kim, J. and Lee, S., 2009. ChimerDB 2.0—a knowledgebase for fusion genes updated. *Nucleic Acids Research,* 38 (suppl_1), D81-D85.

Kiss, C., and Bichler, M., 2008. Identification of influencers—measuring influence in customer networks. *Decision Support Systems,* 46 (1), 233-253.

Knight, P.A., and Ruiz, D., 2013. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis,* 33 (3), 1029-1047.

Kumar, A., Agarwal, S., Phadke, S.R. and Jaiswal, Y., 2014. Genetic insight of schizophrenia: past and future perspectives. *Gene,* 535 (2), 97-100.

Latora, V., and Marchiori, M., 2001. Efficient behavior of small-world networks. *Physical Review Letters,* 87 (19), 198701.

Lee, M., Lee, K., Yu, N., Jang, I., Choi, I., Kim, P., Jang, Y.E., Kim, B., Kim, S. and Lee, B., 2017. ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. *Nucleic Acids Research,* 45 (D1), D784-D789.

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., Wray, N.R. and Schizophrenia Psychiatric Genome-Wide Association Study Consortium, 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics,* 44 (3), 247-250.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-Luria, A.H., Ware, J.S., Hill, A.J. and Cummings, B.B., 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536 (7616), 285-291.

Lewontin, R., and Kojima, K., 1960. The evolutionary dynamics of complex polymorphisms. *Evolution,* 14 (4), 458-472.

Li, W., Li, Y., Zhu, W. and Chen, X., 2014. Changes in brain functional network connectivity after stroke. *Neural Regeneration Research,* 9 (1), 51-60.

Li-Chang, H.H., Driman, D.K., Levin, H., Siu, V.M., Scanlan, N.L., Buckley, K., Cairney, A.E. and Ainsworth, P.J., 2013. Colorectal cancer in a 9-year-old due to combined EPCAM and MSH2 germline mutations: case report of a unique genotype and immunophenotype. *Journal of Clinical Pathology,* 66 (7), 631-633.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J. and Dorschner, M.O., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science,* 326 (5950), 289-293.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B., Ohgi, K.A., Zhang, J., Rose, D.W., Fu, X. and Glass, C.K., 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell,* 139 (6), 1069-1083.

Lin, Y., Hsieh, M.H., Liu, C., Hwang, T., Chien, Y., Hwu, H. and Liu, C., 2014. A recently-discovered NMDA receptor gene, GRIN3B, is associated with duration mismatch negativity. *Psychiatry Research,* 218 (3), 356-358.

Liu, J., Xiong, Q., Shi, W., Shi, X. and Wang, K., 2016. Evaluating the importance of nodes in complex networks. *Physica A: Statistical Mechanics and its Applications,* 452, 209-219.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J. and Axel, R., 2006. Interchromosomal interactions and olfactory receptor choice. *Cell,* 126 (2), 403-413.

Luo, X., Mattheisen, M., Li, M., Huang, L., Rietschel, M., Børglum, A.D., Als, T.D., Van Den Oord, Edwin J, Aberg, K.A. and Mors, O., 2015. Systematic integration of brain eQTL and GWAS identifies ZNF323 as a novel schizophrenia risk gene and suggests recent positive selection based on compensatory advantage on pulmonary function. *Schizophrenia Bulletin,* 41 (6), 1294-1308.

Machiela, M.J., and Chanock, S.J., 2015. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics,* 31 (21), 3555-3557.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A.M., 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature,* 458 (7234), 97-101.

Maher, C.A., Palanisamy, N., Brenner, J.C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T.R., Grasso, C., Yu, J., Lonigro, R.J., Schroth, G., Kumar-Sinha, C. and Chinnaiyan, A.M., 2009. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America,* 106 (30), 12353-12358.

Mani, R.S., Tomlins, S.A., Callahan, K., Ghosh, A., Nyati, M.K., Varambally, S., Palanisamy, N. and Chinnaiyan, A.M., 2009. Induced chromosomal proximity and gene fusions in prostate cancer. *Science (New York, N.Y.),* 326 (5957), 1230.

Mathias, A., Moss, A.J., Lopes, C.M., Barsheshet, A., McNitt, S., Zareba, W., Robinson, J.L., Locati, E.H., Ackerman, M.J. and Benhorin, J., 2013. Prognostic implications of mutation-specific QTc standard deviation in congenital long QT syndrome. *Heart Rhythm,* 10 (5), 720-725.

McCarthy, S.E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., Mistry, M., Pavlidis, P., Solomon, R., Ghiban, E., Antoniou, E., Kelleher, E., O'Brien, C., Donohoe, G., Gill, M., Morris, D.W., McCombie, W.R. and Corvin, A., 2014. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular Psychiatry,* 19 (6), 652-658.

McPherson, A., Wu, C., Hajirasouliha, I., Hormozdiari, F., Hach, F., Lapuk, A., Volik, S., Shah, S., Collins, C. and Sahinalp, S.C., 2011. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics,* 27 (11), 1481-1488.

Mendel, G., 1996. Experiments in plant hybridization (1865). *Verhandlungen Des Naturforschenden Vereins Brünn.) Available Online: Www.Mendelweb.org/Mendel.Html (Accessed on 1 January 2013),* .

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W. and Ewels, P.A., 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics,* 47 (6), 598-606.

Minikel, E.V., Zerr, I., Collins, S.J., Ponto, C., Boyd, A., Klug, G., Karch, A., Kenny, J., Collinge, J. and Takada, L.T., 2014. Ascertainment bias causes false signal of anticipation in genetic prion disease. *The American Journal of Human Genetics,* 95 (4), 371-382.

Mitelman, F., Johansson, B. and Mertens, F., 2007. The impact of translocations and gene fusions on cancer causation. *Nature Reviews.Cancer,* 7 (4), 233.

Morissette, J., Clépet, C., Moisan, S., Dubois, S., Winstall, E., Vermeeren, D., Nguyen, T., Polansky, J., Côté, G. and Anctil, J., 1998. Homozygotes carrying an autosomal dominant TIGR mutation do not manifest glaucoma. *Nature Genetics,* 19 (4), 319-321.

Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M. and Kazane, K.R., 2017. Enhancer connectome in primary human cells reveals target genes of disease-associated DNA elements. *Biorxiv,* , 178269.

Nacu, S., Yuan, W., Kan, Z., Bhatt, D., Rivers, C.S., Stinson, J., Peters, B.A., Modrusan, Z., Jung, K. and Seshagiri, S., 2011. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Medical Genomics,* 4 (1), 11.

Neudorfer, O., Pastores, G.M., Zeng, B.J., Gianutsos, J., Zaroff, C.M. and Kolodny, E.H., 2005. Late-onset Tay-Sachs disease: phenotypic characterization and genotypic correlations in 21 affected patients. *Genetics in Medicine,* 7 (2), 119-123.

Newman, M., 2010. *Networks: an introduction.* Oxford university press.

Novo, F.J., de Mendíbil, I.O. and Vizmanos, J.L., 2007. TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics,* 8 (1), 33.

Nowell, P.C., 2007. Discovery of the Philadelphia chromosome: a personal perspective. *The Journal of Clinical Investigation,* 117 (8), 2033-2035.

Page, L., Brin, S., Motwani, R. and Winograd, T., 1999. *The PageRank Citation Ranking: Bringing Order to the Web.,* .

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X. and Suleman, K., 2010. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nature Medicine,* 16 (7), 793-798.

Pan, Y., Laird, J.G., Yamaguchi, D.M. and Baker, S.A., 2015. A di-arginine ER retention signal regulates trafficking of HCN1 channels from the early secretory pathway to the plasma membrane. *Cellular and Molecular Life Sciences,* 72 (4), 833-843.

Pei, Y., Paterson, A.D., Wang, K.R., He, N., Hefferton, D., Watnick, T., Germino, G.G., Parfrey, P., Somlo, S. and George-Hyslop, P.S., 2001. Bilineal disease and trans-heterozygotes in autosomal dominant polycystic kidney disease. *The American Journal of Human Genetics,* 68 (2), 355-363.

Pern, F., Bogdanova, N., Schürmann, P., Lin, M., Ay, A., Länger, F., Hillemanns, P., Christiansen, H., Park-Simon, T. and Dörk, T., 2012. Mutation analysis of BRCA1, BRCA2, PALB2 and BRD7 in a hospital-based series of German patients with triple-negative breast cancer. *PloS One,* 7 (10), e47993.

Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P.C., Svensson, M.A., Kitabayashi, N., Moss, B.J., MacDonald, T.Y., Cao, X., Barrette, T., Tewari, A.K., Chee, M.S., Chinnaiyan, A.M., Rickman, D.S., Demichelis, F., Gerstein, M.B. and Rubin, M.A., 2011. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Research,* 21 (1), 56-67.

Phizicky, E.M., and Fields, S., 1995. Protein-protein interactions: methods for detection and analysis. *Microbiological Reviews,* 59 (1), 94-123.

Pitteloud, N., Quinton, R., Pearce, S., Raivio, T., Acierno, J., Dwyer, A., Plummer, L., Hughes, V., Seminara, S., Cheng, Y.Z., Li, W.P., Maccoll, G., Eliseenkova, A.V., Olsen, S.K., Ibrahimi, O.A., Hayes, F.J., Boepple, P., Hall, J.E., Bouloux, P., Mohammadi, M. and Crowley, W., 2007. Digenic mutations account for variable phenotypes in idiopathic hypogonadotropic hypogonadism. *The Journal of Clinical Investigation,* 117 (2), 457-463.

Pleasance, E.D., Stephens, P.J., O'meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M., Beare, D., Lau, K.W. and Greenman, C., 2010. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature,* 463 (7278), 184-190.

Ramalingam, A., Zhou, X., Fiedler, S.D., Brawner, S.J., Joyce, J.M., Liu, H. and Yu, S., 2011. 16p13. 11 duplication is a risk factor for a wide spectrum of neuropsychiatric disorders. *Journal of Human Genetics,* 56 (7).

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Coin, L., De Silva, R. and Cookson, M.R., 2014. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience,* 17 (10), 1418-1428.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D. and Lander, E.S., 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell,* 159 (7), 1665-1680.

Ratjen, F., and Döring, G., 2003. *Cystic fibrosis.*

Rees, E., O'Donovan, M.C. and Owen, M.J., 2015. Genetics of schizophrenia. *Current Opinion in Behavioral Sciences,* 2, 8-14.

Reich, D.E., and Lander, E.S., 2001. On the allelic spectrum of human disease. *TRENDS in Genetics,* 17 (9), 502-510.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J. and Fromer, M., 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics,* 45 (10), 1150-1159.

Robinson, C.V., Sali, A. and Baumeister, W., 2007. The molecular sociology of the cell. *Nature,* 450 (7172), 973-982.

Robinson, D.R., Kalyana-Sundaram, S., Wu, Y., Shankar, S., Cao, X., Ateeq, B., Asangani, I.A., Iyer, M., Maher, C.A. and Grasso, C.S., 2011. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nature Medicine,* 17 (12), 1646-1651.

Roix, J.J., McQueen, P.G., Munson, P.J., Parada, L.A. and Misteli, T., 2003. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nature Genetics,* 34 (3), 287.

Rubinov, M., and Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage,* 52 (3), 1059-1069.

Saini, S., Robinson, P.N., Singh, J.R. and Vanita, V., 2012. A novel 7 bp deletion in PRPF31 associated with autosomal dominant retinitis pigmentosa with incomplete penetrance in an Indian family. *Experimental Eye Research,* 104, 82-88.

Sanders, A.R., Göring, H.H., Duan, J., Drigalenko, E.I., Moy, W., Freda, J., He, D., Shi, J., Mgs and Gejman, P.V., 2013. Transcriptome study of differential expression in schizophrenia. *Human Molecular Genetics,* 22 (24), 5001-5014.

Sano, K., Tanihara, H., Heimark, R.L., Obata, S., Davidson, M., St John, T., Taketani, S. and Suzuki, S., 1993. Protocadherins: a large family of cadherin-related molecules in central nervous system. *The EMBO Journal,* 12 (6), 2249-2256.

Sati, S., and Cavalli, G., 2017. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma,* 126 (1), 33-44.

Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D.Z., Rozowsky, J.S., Tewari, A.K., Kitabayashi, N., Moss, B.J. and Chee, M.S., 2010. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biology,* 11 (10), R104.

Schaffer, A.A., 2013. Digenic inheritance in medical genetics. *Journal of Medical Genetics,* 50 (10), 641-652.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium, 2011. Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics,* 43 (10), 969-976.

Shabalin, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics,* 28 (10), 1353-1358.

Shawky, R.M., 2014. Reduced penetrance in human inherited disease. *Egyptian Journal of Medical Human Genetics,* 15 (2), 103-111.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B. and De Laat, W., 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics,* 38 (11), 1348.

Sinkhorn, R., and Knopp, P., 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics,* 21 (2), 343-348.

Smith, J.D., Hing, A.V., Clarke, C.M., Johnson, N.M., Perez, F.A., Park, S.S., Horst, J.A., Mecham, B., Maves, L. and Nickerson, D.A., 2014. Exome sequencing identifies a recurrent de novo ZSWIM6 mutation associated with acromelic frontonasal dysostosis. *The American Journal of Human Genetics,* 95 (2), 235-240.

Spurdle, A.B., Whiley, P.J., Thompson, B., Feng, B., Healey, S., Brown, M.A., Pettigrew, C., kConFab, Van Asperen, C.J., Ausems, M.G., Kattentidt-Mouravieva, A.A., van den Ouweland, A.M., Dutch Belgium UV Consortium, Lindblom, A., Pigg, M.H., Schmutzler, R.K., Engel, C., Meindl, A., German Consortium of Hereditary Breast and Ovarian Cancer, Caputo, S., Sinilnikova, O.M., Lidereau, R., French COVAR group collaborators, Couch, F.J., Guidugli, L., Hansen, T., Thomassen, M., Eccles, D.M., Tucker, K., Benitez, J., Domchek, S.M., Toland, A.E., Van Rensburg, E.J., Wappenschmidt, B., Borg, A., Vreeswijk, M.P., Goldgar, D.E. and ENIGMA Consortium, 2012. BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk. *Journal of Medical Genetics,* 49 (8), 525-532.

Staudt, C.L., Sazonovs, A. and Meyerhenke, H., 2016. NetworKit: A tool suite for large-scale complex network analysis. *Network Science,* 4 (4), 508-530.

Steen, H., and Lindholm, D., 2008. Nuclear localized protein-1 (Nulp1) increases cell death of human osteosarcoma cells and binds the X-linked inhibitor of apoptosis protein. *Biochemical and Biophysical Research Communications,* 366 (2), 432-437.

Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D. and Cooper, D.N., 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics,* , 1-13.

Sugawara, H., Iwamoto, K., Bundo, M., Ueda, J., Ishigooka, J. and Kato, T., 2011. Comprehensive DNA methylation analysis of human peripheral blood leukocytes and lymphoblastoid cell lines. *Epigenetics,* 6 (4), 508-515.

Symmons, O., and Spitz, F., 2013. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philosophical Transactions of the Royal Society of London.Series B, Biological Sciences,* 368 (1620), 20120358.

Terauchi, A., Johnson-Venkatesh, E.M., Toth, A.B., Javed, D., Sutton, M.A. and Umemori, H., 2010. Distinct FGFs promote differentiation of excitatory and inhibitory synapses. *Nature,* 465 (7299), 783-787.

Thauvin-Robinet, C., Munck, A., Huet, F., Genin, E., Bellis, G., Gautier, E., Audrezet, M.P., Ferec, C., Lalau, G., Georges, M.D., Claustres, M., Bienvenu, T., Gerard, B., Boisseau, P., Cabet-Bey, F., Feldmann, D., Clavel, C., Bieth, E., Iron, A., Simon-Bouy, B., Costa, C., Medina, R., Leclerc, J., Hubert, D., Nove-Josserand, R., Sermet-Gaudelus, I., Rault, G., Flori, J., Leroy, S., Wizla, N., Bellon, G., Haloun, A., Perez-Martin, S., d'Acremont, G., Corvol, H., Clement, A., Houssin, E., Binquet, C., Bonithon-Kopp, C., Alberti-Boulme, C., Morris, M.A., Faivre, L., Goossens, M., Roussey, M., Collaborating Working Group on R117H and Girodon, E., 2009. The very low penetrance of cystic fibrosis for the R117H mutation: a reappraisal for genetic counselling and newborn screening. *Journal of Medical Genetics,* 46 (11), 752-758.

Tropeano, M., Ahn, J.W., Dobson, R.J., Breen, G., Rucker, J., Dixit, A., Pal, D.K., McGuffin, P., Farmer, A. and White, P.S., 2013. Male-biased autosomal effect of 16p13. 11 copy number variation in neurodevelopmental disorders. *PloS One,* 8 (4), e61365.

Uhrhammer, N., and Bignon, Y., 2008. Report of a family segregating mutations in both the APC and MSH2 genes: juvenile onset of colorectal cancer in a double heterozygote. *International Journal of Colorectal Disease,* 23 (11), 1131-1135.

Van De Werken, Harmen JG, Landan, G., Holwerda, S.J., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Öz, Y. and Bouwman, B.A., 2012. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods,* 9 (10), 969-972.

van der Kolk, Dorina M, de Bock, G.H., Leegte, B.K., Schaapveld, M., Mourits, M.J., de Vries, J., van der Hout, Annemieke H and Oosterwijk, J.C., 2010. Penetrance of breast cancer, ovarian cancer and contralateral breast cancer in BRCA1 and BRCA2 families: high cancer incidence at older age. *Breast Cancer Research and Treatment,* 124 (3), 643-651.

Wacholder, S., Hartge, P., Struewing, J.P., Pee, D., McAdams, M., Brody, L. and Tucker, M., 1998. The kin-cohort study for estimating penetrance. *American Journal of Epidemiology,* 148 (7), 623-630.

Wang, Q., Xia, J., Jia, P., Pao, W. and Zhao, Z., 2012. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Briefings in Bioinformatics,* 14 (4), 506-519.

Wang, Y., Ottman, R. and Rabinowitz, D., 2006. A method for estimating penetrance from families sampled for linkage analysis. *Biometrics,* 62 (4), 1081-1088.

Wang, Y., Wu, N., Liu, J., Wu, Z. and Dong, D., 2015. FusionCancer: a database of cancer fusion genes derived from RNA-seq data. *Diagnostic Pathology,* 10 (1), 131.

Watson, I.R., Takahashi, K., Futreal, P.A. and Chin, L., 2013. Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics,* 14 (10), 703-718.

Watson, J.D., and Crick, F.H., 1953. Molecular structure of nucleic acids. *Nature,* 171 (4356), 737-738.

Watts, D.J., and Strogatz, S.H., 1998. Collective dynamics of'small-world'networks. *Nature,* 393 (6684), 440.

Welch, J.S., Westervelt, P., Ding, L., Larson, D.E., Klco, J.M., Kulkarni, S., Wallis, J., Chen, K., Payton, J.E. and Fulton, R.S., 2011. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *Jama,* 305 (15), 1577-1584.

Wetterstrand, K.A., 2017. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP),* .

Wijchers, P.J., and de Laat, W., 2011. Genome organization influences partner selection for chromosomal rearrangements. *TRENDS in Genetics,* 27 (2), 63-71.

Wilson, R.J., 1970. *An introduction to graph theory.* Pearson Education India.

Won, H., de La Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F. and Lu, D., 2016. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature,* 538 (7626), 523-527.

Wong, A.H., Gottesman, I.I. and Petronis, A., 2005. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics,* 14 (suppl_1), R11-R18.

Wu, C., Kannan, K., Lin, S., Yen, L. and Milosavljevic, A., 2013. Identification of cancer fusion drivers using network fusion centrality. *Bioinformatics,* 29 (9), 1174-1181.

Würtele, H., and Chartrand, P., 2006. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Research,* 14 (5), 477-495.

Yaffe, E., and Tanay, A., 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics,* 43 (11), 1059-1065.

Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H. and Verhaak, R.G., 2015. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene,* 34 (37), 4845-4854.

Zhang, Y., McCord, R.P., Ho, Y., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W. and Dekker, J., 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell,* 148 (5), 908-921.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S. and Singh, U., 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature Genetics,* 38 (11), 1341.

Zhao, Q., Caballero, O.L., Levy, S., Stevenson, B.J., Iseli, C., de Souza, S.J., Galante, P.A., Busam, D., Leversha, M.A., Chadalavada, K., Rogers, Y.H., Venter, J.C., Simpson, A.J. and Strausberg, R.L., 2009. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proceedings of the National Academy of Sciences of the United States of America,* 106 (6), 1886-1891.

Zlotogora, J., 2003. Penetrance and expressivity in the molecular age. *Genetics in Medicine,* 5 (5), 347-352.

Appendix A

Table A4.1 Putative primary and modifier gene pairs which exhibit epistatic properties, hence modulating penetrance levels (Cooper et al., 2013).

Disease	Primary gene	Modifier gene	Reference
X-linked retinitis pigmentosa	*RPGR*	*IQCB1*	Fahim et al., 2011
X-linked retinitis pigmentosa	*RPGR*	*RPGRIP1L*	Fahim et al., 2011
Retinoblastoma	*RB1*	*MDM2*	Castéra et al., 2010
Familial hypercholesterolaemia	*LDLR*	*PCSK9*	Abifadel et al., 2009
Familial hypercholesterolaemia	*LDLR*	*APOB*	Benlian et al., 1996/Taylor et al., 2010
Familial hypercholesterolaemia	*LDLR*	*CFH*	Koeijvoets et al., 2009
Familial hypercholesterolaemia	*LDLR*	*APOH*	Takada et al., 2003a
Familial hypercholesterolaemia	*LDLR*	*GHR*	Takada et al., 2003b
Familial hypercholesterolaemia	*LDLR*	*EPHX2*	Sato et al., 2004
Breast cancer	*BRCA2*	*RAD51*	Antoniou et al., 2007
Ovarian cancer	*BRCA1*	*IRS1*	Ding et al., 2012
Ovarian cancer	*BRCA2*	*IRS1*	Ding et al., 2012
Lynch syndrome	*MSH2*	*RNASEL*	Krüger et al., 2007
Lynch syndrome	*MLH1*	*RNASEL*	Krüger et al., 2007
Lynch syndrome	*MSH2*	*TP53*	Krüger et al., 2007
Lynch syndrome	*MLH1*	*TP53*	Krüger et al., 2007
Cystic fibrosis	*CFTR*	*TGFB1*	Drumm et al., 2005
Familial pulmonary arterial hypertension	*BMPR2*	*TGFB1*	Phillips et al., 2008
Paget's disease	*SQSTM1*	*TNFRSF11A*	Gianfrancesco et al., 2012
X-linked variable immunodeficiency	*XIAP*	*CD40LG*	Rigaud et al., 2011
Haemochromatosis	*HFE*	*CYBRD1*	Constantine et al., 2009
Parkinson's disease	*GBA*	*MTX1*	Gan-Or et al., 2011
Recessive dystrophic epidermolysis bullosa	*COL7A1*	*MMP1*	Titeux et al., 2008
Amyotrophic lateral sclerosis	*SOD1*	*CHGB*	Gros-Louis et al., 2009
Huntingdon disease	*HTT*	*HAP1*	Metzger et al., 2008

Fatal kernicterus	*G6PD*	*UGT1A1*	Zangen et al., 2009
Atypical haemolytic uraemic syndrome	*CFH*	*C4BPA*	Blom et al., 2008
Spinal muscular atrophy	*SMN1*	*SMN2*	Prior et al., 2009
Long QT syndrome	*KCNQ1*	*KCNH2*	Cordeiro et al., 2010
Long QT syndrome	*KCNQ1*	*ADRB1*	Paavonen et al., 2007
Long QT syndrome	*KCNQ1*	*NOS1AP*	Crotti et al., 2009
Familial venous thrombosis	*PROC*	*F5*	Koeleman et al., 1994/Gandrille et al., 1995/Cafolla et al., 2012
Familial venous thrombosis	*PROS1*	*F5*	Koeleman et al., 1995
Familial venous thrombosis	*SERPINC1*	*F5*	Van Boven et al., 1996
Hypertrophic cardiomyopathy	*MYBPC3*	*CALM3*	Friedrich et al., 2009
Hypertrophic cardiomyopathy	*MYH7*	*CALM3*	Friedrich et al., 2009
Familial Mediterranean fever	*MEFV*	*SAA1*	Migita et al., 2013

Table A4.2 Putative examples of digenic mutations causing inherited disease (Cooper et al., 2013).

Disease	Gene 1	Gene 2	Reference
Waardenburg syndrome type 2	*MITF*	*PAX3/OCA3/TYR/GJB2*	Morell et al., 1997; Chiang et al., 2009; Yan et al., 2011/Yang et al., 2013
Retinitis pigmentosa	*PRPH2*	*ROM1/RHO/PDE6B*	Kajiwara et al., 1994; Loewen et al., 2001; Sullivan et al., 2006/Jin et al., 2008
Retinitis pigmentosa	*RHO*	*PRPF31*	Lim et al., 2009
Retinitis pigmentosa	*PDE6B*	*GPR98*	Hmani-Aifa et al., 2009
Progressive cone dystrophy	*CNGA3*	*CNGB3*	Thiadens et al., 2010
Frontotemporal dementia	*PSEN1*	*PRNP*	Bernardi et al., 2011
Leber congenital amaurosis	*RPE65*	*GUCY2D*	Silva et al., 2004
Idiopathic hypogonadotropic hypogonadism	*FGFR1*	*GNRHR/NELF*	Pitteloud et al., 2007

Bilateral cystic renal dysplasia	*DACH1*	*BMP4*	Schild et al., 2013
Glaucoma, early onset	*MYOC*	*CYP1B1/LTBP2*	Vincent et al., 2002/Geyer et al., 2011/Azmanov et al., 2011
Severe insulin resistance	*PPARG*	*PPP1R3A*	Savage et al., 2002
Usher syndrome type 2	*PDZD7*	*GPR98*	Ebermann et al., 2010
Usher syndrome type 1-associated deafness	*CDH23*	*PCDH15*	Zheng et al., 2005
Hidrotic ectodermal dysplasia	*GJB2*	*GJA1*	Kellermayer et al., 2005
Non-syndromic deafness	*GJB2*	*GJB3*	Liu et al., 2009
Hearing loss	*GJB2*	*SLC26A4*	Sagong et al., 2013
Non-syndromic hearing loss associated with an enlarged vestibular aqueduct/Pendred syndrome	*KCNJ10*	*SLC26A4*	Yang et al., 2009
Porphyria	*CPOX*	*PPOX*	van Tuyll van Serooskerken et al., 2011
Atypical haemolytic uremic syndrome	*CFI*	*CD46/C3/CFB/CFHR1*	Esparza-Gordillo et al., 2006/Westra et al., 2010/Bresin et al., 2013
Atypical haemolytic uraemic syndrome	*CFH*	*CD46/CFI/C3/THBD*	Sullivan et al., 2011/Bresin et al., 2013/Fan et al., 2013
Epidermolysis bullosa simplex	*KRT14*	*KRT5*	Padalon-Brauch et al., 2012
Junctional epidermolysis bullosa	*COL17A1*	*LAMB3*	Floeth/Bruckner-Tuderman, 1999
Long QT syndrome	*KCNQ1*	*KCNH2/KCNE1/SCN5A*	Schwartz et al., 2003/Westenskow et al., 2004/Tester et al., 2005/Itoh et al., 2010
Long QT syndrome	*KCNH2*	*SCN5A/KCNE*	Schwartz et al.,

		1	2003/Westenskow et al., 2004/Tester et al., 2005
Long QT syndrome	*SCN5A*	*SNTA1/KCNE1*	Westenskow et al., 2004/Hu et al., 2013
Haemochromatosis	*HFE*	*HAMP/TFR2*	Merryweather-Clarke et al., 2003/Jacolot et al., 2004/Island et al., 2009/Altès et al., 2009/Del-Castillo-Rueda et al., 2012
Kallmann syndrome	*PROK2*	*PROKR2*	Cole et al., 2008/Sarfati et al., 2010/Shaw et al., 2011
Kallmann syndrome	*NELF*	*KAL1/TACR3*	Xu et al., 2011/Quaynor et al., 2011
Kallmann syndrome	*PROKR2*	*KAL1*	Dodé et al., 2006/Canto et al., 2009/Shaw et al., 2011
Kallmann syndrome	*KAL1*	*TACR3/WDR11/CHD7*	Quaynor et al., 2011/Shaw et al., 2011
Normosmic idiopathic hypogonadotrophic hypogonadism	*GNRH*	*KAL1*	Quaynor et al., 2011
Normosmic idiopathic hypogonadotrophic hypogonadism	*WDR11*	*GNRHR*	Quaynor et al., 2011
Normosmic idiopathic hypogonadotrophic hypogonadism	*FGFR1*	*GNRHR/PROKR2/FGF8/KAL1/GPR54*	Raivio et al., 2009/Sykiotis et al., 2010/Shaw et al., 2011
Systemic amyloid A amyloidosis	*TNFRSF1A*	*MEFV*	Cigni et al., 2006/Mereuta et al., 2013
Familial hypercholesterolaemia	*LDLR*	*PCSK9*	Pisciotta et al., 2006/Noguchi et al.,

			2010/Bertolini et al., 2013
Familial hypercholesterolaemia	*LDLR*	*APOB*	Bertolini et al., 2013
Familial hypercholesterolaemia	*LDLR*	*LDLRAP1*	Tada et al., 2011
Severe congenital neutropenia	*ELANE*	*G6PC3/HAX1*	Germeshausen et al., 2010
McArdle's disease	*PYGM*	*CPT2*	Vockley et al., 2000
Parkinson's disease, early onset	*PINK1*	*PARK2/PARK7*	Tang et al., 2006/Funayama et al., 2008
Parkinson's disease	*LRRK2*	*PRKN*	Dächsel et al., 2006
Emery–Dreifuss muscular dystrophy	*LMNA*	*DES*	Muntoni et al., 2006
Joubert syndrome and nephronophthisis	*NPHP1*	*NPHP6*	Tory et al., 2007
Axenfeld–Rieger syndrome	*FOXC1*	*PITX2*	Kelberman et al., 2011
Cortisone reductase deficiency	*HSD11B1*	*H6PD*	Draper et al., 2003/San Millán et al., 2005
Hypertrophic cardiomyopathy	*MYBPC3*	*TNNT2/TNNI3/MYH7/TPM1*	Richard et al., 2003/Van Driest et al., 2004 Ingles et al., 2005/Millat et al., 2010/Kubo et al., 2011/Zou et al., 2013
Hypertrophic cardiomyopathy	*MYH7*	*TNNT2/MYL2/TNNI3/ACTC1*	Millat et al., 2010/Zou et al., 2013
Restrictive cardiomyopathy	*MYL2*	*MYL3*	Caleshu et al., 2011
Rasopathy phenotype with severe hypertrophic cardiomyopathy	*PTPN11*	*SOS1*	Fahrner et al., 2012
Arrhythmogenic right ventricular cardiomyopathy	*DES*	*PKP2*	Lorenzon et al., 2013
Arrhythmogenic right ventricular cardiomyopathy	*DES*	*DSG2*	Rasmussen et al., 2013

Arrhythmogenic right ventricular cardiomyopathy	*PKP2*	*DSP/DSG2/PKP4/DSC2*	Xu et al., 2010
Arrhythmogenic right ventricular dysplasia/cardiomyopathy	*DSG2*	*DSC/PKP2*	Bhuiyan et al., 2009/Nakajima et al., 2012
Familial dilated cardiomyopathy	*LMNA*	*TTN*	Roncarati et al., 2013
Dent’s disease	*CLCN5*	*OCRL*	Addis et al., 2013
Amyotrophic lateral sclerosis	*SOD1*	*CNTF*	Giess et al., 2002
Amyotrophic lateral sclerosis	*VAPB*	*C9orf72*	van Blitterswijk et al., 2012b
Dravet syndrome	*PCDH19*	*TSPYL4*	Kwong et al., 2012
Dravet syndrome	*SCN9A*	*SCN1A*	Singh et al., 2009
Dravet syndrome	*CACNA1A*	*SCN1A*	Ohmori et al., 2013
Severe myoclonic epilepsy	*CACNB4*	*SCN1A*	Ohmori et al., 2008
Severe myoclonic epilepsy	*POLG*	*SCN1A*	Bolszak et al., 2009
Progressive external ophthalmoplegia	*POLG*	*SLC25A4*	Galassi et al., 2008
Bartter syndrome	*CLCNKA*	*CLCNKB*	Nozu et al., 2008
Chronic pancreatitis	*SPINK1*	*CASR/CFTR/CTRC/PRSS1*	Felderbauer et al., 2003/Masson et al., 2007/Tzetis et al., 2007/Schneider et al., 2011/LaRusch et al., 2012/Rosendahl et al., 2013
Oculocutaneous albinsim	*OCA2*	*TYRP1/SLC45A2/TYR*	Chiang et al., 2008/Wei et al., 2013
Oculocutaneous albinsim	*TYR*	*SLC45A2*	Wei et al., 2013
Cystinuria	*SLC3A1*	*SLC7A9*	Font-Llitjós et al., 2005
Transposition of the great arteries	*ZIC3*	*FOXH1/NKX2-5*	De Luca et al., 2010
Congenital heart disease	*MYH6*	*NKX2-5/GATA4*	Granados-Riveron et al., 2012
Charcot–Marie–Tooth disease	*PMP22*	*ABCD1/LITAF*	Meggouh et al.,

			2005/Hodapp et al., 2006
Charcot–Marie–Tooth disease	*GJB1*	*EGR2*	Chung et al., 2005
Charcot–Marie–Tooth disease	*GDAP1*	*MFN2*	Vital et al., 2012
Refractory auto-inflammatory syndrome	*TNFRSF1A*	*CIAS1*	Touitou et al., 2006
Short-rib polydactyly syndrome type 2	*NEK1*	*DYNC2H1*	Thiel et al., 2011
Maturity-onset diabetes of the young	*HNF1A*	*HNF1B*	Karges et al., 2007
Maturity-onset diabetes of the young	*HNF1A*	*HNF4A*	Forlani et al., 2010/Shankar et al., 2013
Polycystic kidney disease	*PKD1*	*PKD2*	Pei et al., 2001/Dedoussis et al., 2008
Hyperimmunoglobulinaemia D and periodic fever syndrome	*MVK*	*TNFRSF1A*	Hoffmann et al., 2005
Obesity, hyperinsulinaemia and insulin resistance	*TCF1*	*NROB2*	Tonooka et al., 2002
Progressive external ophthalmoplegia	*POLG*	*C10orf2*	Van Goethem et al., 2003
Neuronal ceroid lipofuscinosis	*POLG*	*CLN5*	Staropoli et al., 2012
Chronic lung disease	*SFTPC*	*ABCA3*	Bullard/Nogee, 2007
Lafora disease	*EPM2B*	*PPP1R3C*	Guerrero et al., 2011
Congenital erythropoietic porphyria	*UROS*	*ALAS2*	To-Figueras et al., 2011
Familial venous thrombosis	*PROC*	*PROS1*	Formstone et al., 1996/Brenner et al., 1996/Boinot et al., 2003/Knoll et al., 2001/Hayashida et al., 2003
Familial venous thrombosis	*PROC*	*SERPIND1*	Bernardi et al., 1996

Breast cancer	*BRCA1*	*BRCA2*	Leegte et al., 2005/Lavie et al., 2011/Heidemann et al., 2012
Breast cancer	*BRCA1*	*PALB2*	Pern et al., 2012
Multiple tumours of different types	*BRCA1*	*MLH1*	Pedroni et al., 2013
Familial pulmonary arterial hypertension	*BMPR2*	*THBS1*	Maloney et al., 2012
Hereditary nonpolyposis colorectal cancer	*MUTYH*	*MSH6*	Van Puijenbroek et al., 2007/Giráldez et al., 2009
Colorectal cancer	*EPCAM*	*MSH2*	Li-Chang et al., 2013
Colorectal cancer, juvenile onset	*APC*	*MSH2*	Uhrhammer/Bignon, 2008
Autoimmune lymphoproliferative syndrome	*FAS*	*CASP10*	Cerutti et al., 2007
Autoimmune lymphoproliferative syndrome	*FAS*	*PRF1*	Clementi et al., 2004
Steroid-resistant focal segmental glomerulosclerosis	*NPHS2*	*NPHS1/CD2AP*	Löwik et al., 2008
Severe infantile liver disease	*AKR1D1*	*SKIV2L*	Morgan et al., 2013
Ataxia, dementia and hypogonadotropism	*RNF216*	*OTUD4*	Margolin et al., 2013
Paediatric inflammatory bowel disease	*NOD2*	*GSDMB/ZNF365/ERAP2/SEC16A/GMPBB*	Christodoulou et al., 2013
Paediatric inflammatory bowel disease	*BACH2*	*IL10*	Christodoulou et al., 2013
Hutchinson-Gilford progeria syndrome	*ZMPSTE24*	*LMNA*	Denecke et al., 2006

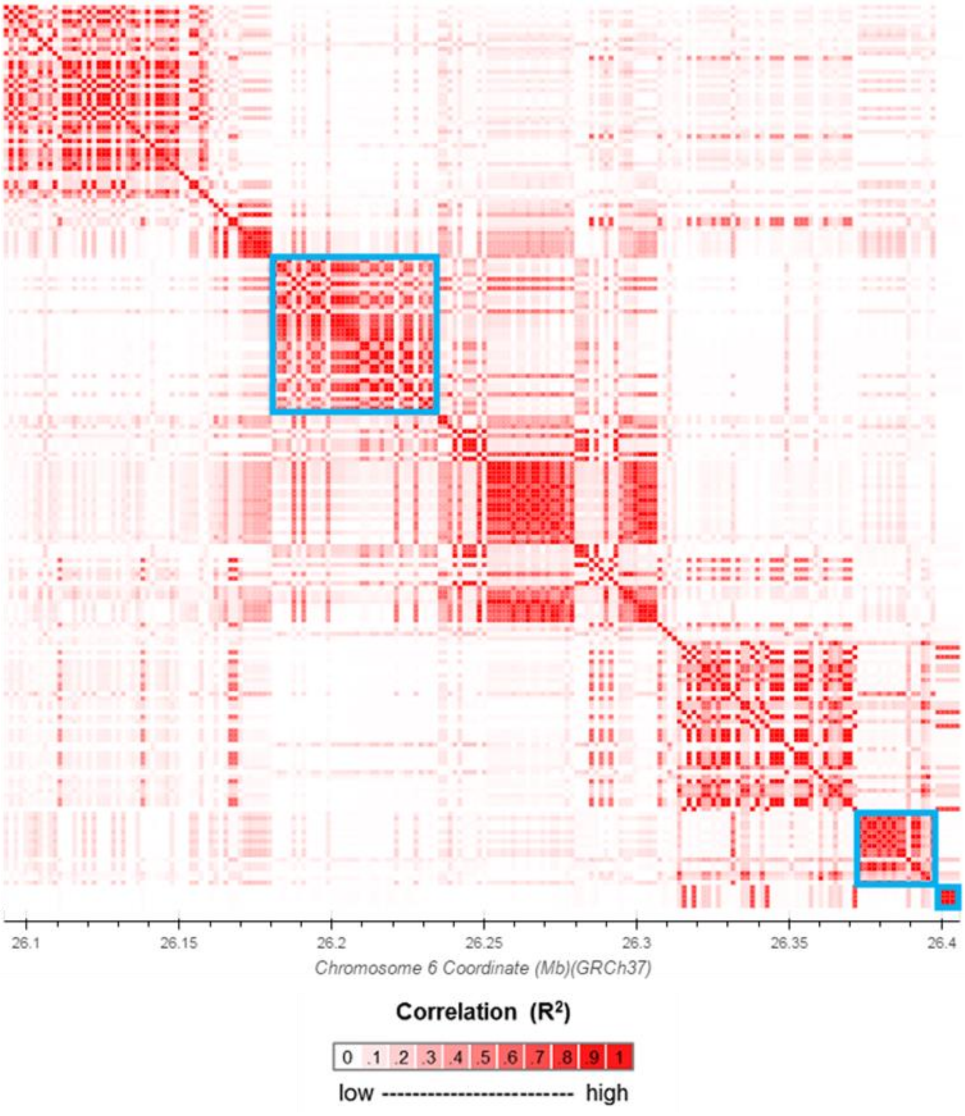


Figure A6.1 Linkage disequilibrium (LD) heat map of chromosome 6, positions 26,100,000-26,410,000 (via LDlink; Machiela & Chanock, 2015). Blocks of red indicate high R^2 values, corresponding to genomic regions in LD. Blocks enclosed by blue squares are regions which have been found to harbour SNPs that affect gene expression change in the brain.

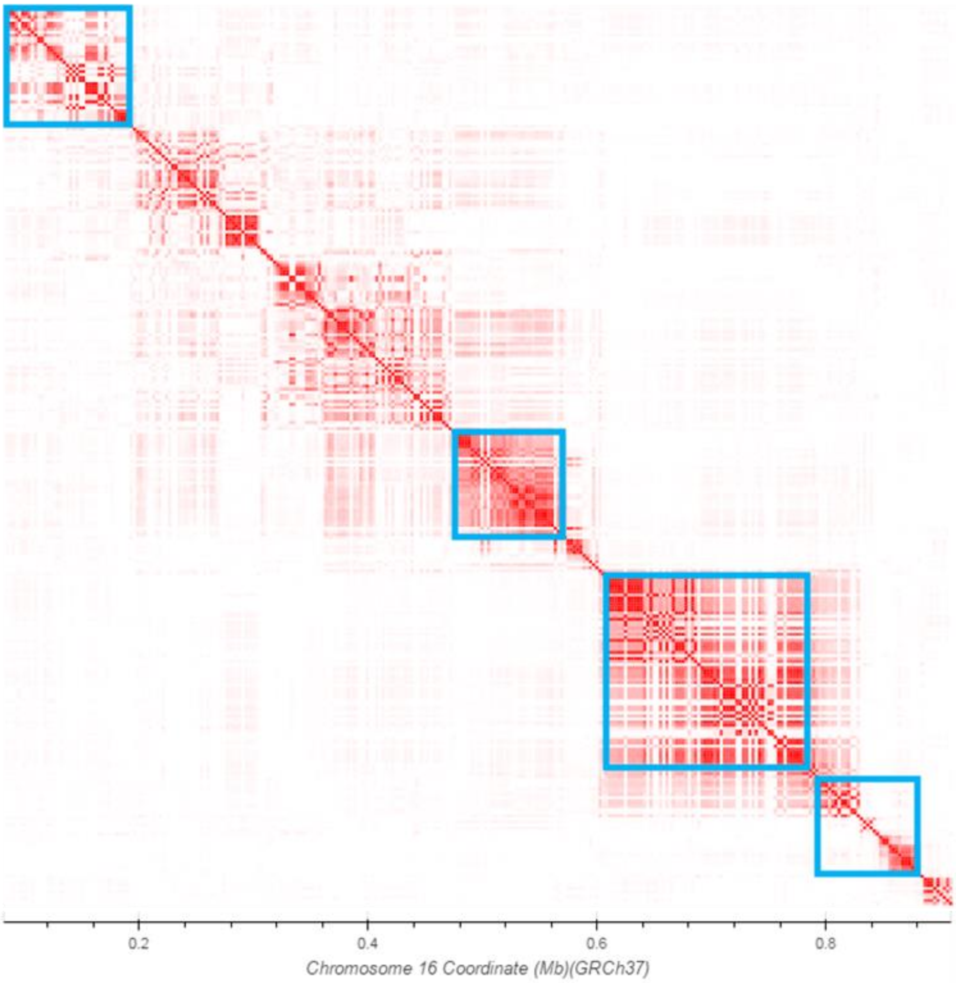


Figure A6.2 Linkage disequilibrium (LD) heat map of chromosome 16, positions 100,000-890,000 (via LDlink; Machiela & Chanock, 2015). Blocks of red indicate high R^2 values, corresponding to genomic regions in LD. Blocks enclosed by blue squares are regions which have been found to harbour SNPs that affect gene expression change in the brain.

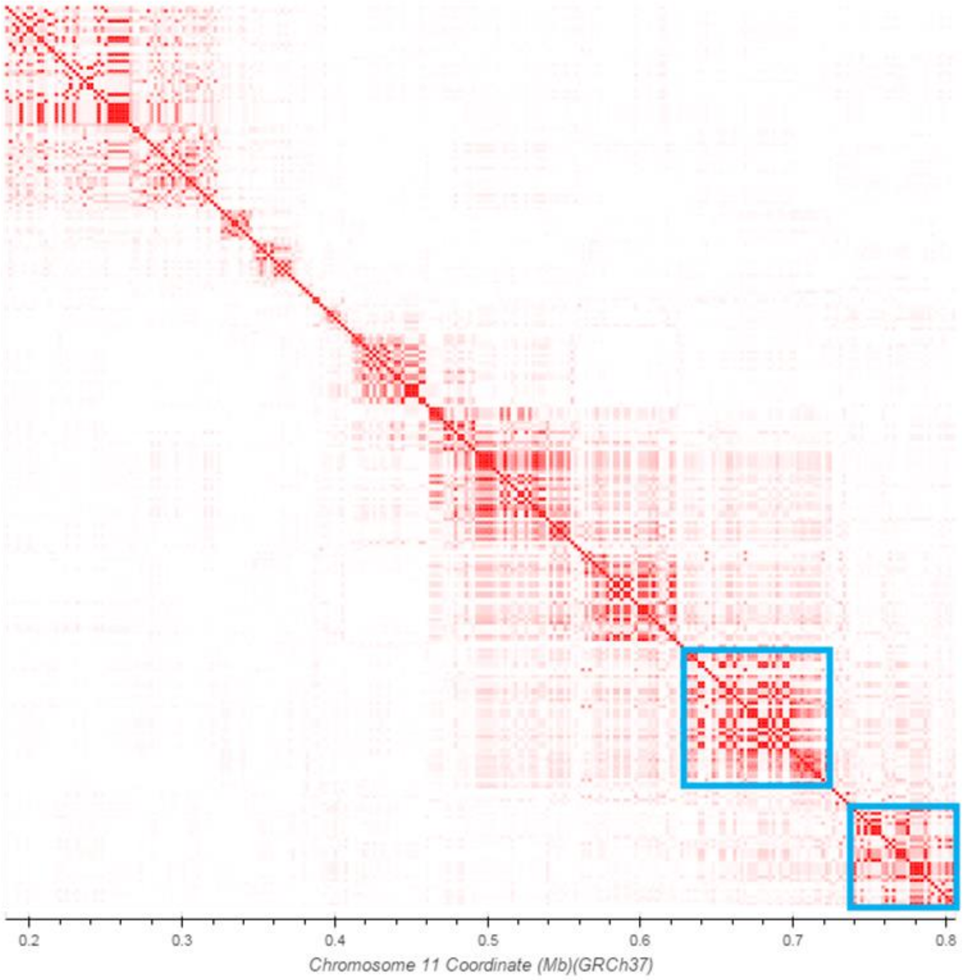


Figure A6.3 Linkage disequilibrium (LD) heat map of chromosome 11, positions 100,000-800,000 (via LDlink; Machiela & Chanock, 2015). Blocks of red indicate high R^2 values, corresponding to genomic regions in LD. Blocks enclosed by blue squares are regions which have been found to harbour SNPs that affect gene expression change in the brain.

Table A6.1 List of 347 schizophrenia genes, which were identified by GWAS (Ripke et al., 2013).

		Gene IDs		
ABCB9	*CNKSR2*	*HIRIP3*	*PAK6*	*SLC39A8*
AC005477.1	*CNNM2*	*HSPA9*	*PARD6A*	*SLC45A1*
AC005609.1	*CNOT1*	*HSPD1*	*PBRM1*	*SLC4A10*
AC027228.1	*CNTN4*	*HSPE1*	*PBX4*	*SLC7A6*
AC073043.2	*COQ10B*	*IGSF9B*	*PCCB*	*SLC7A6OS*
ACD	*CR1L*	*IK*	*PCDHA1*	*SMDT1*
ACTR5	*CREB3L1*	*IMMP2L*	*PCDHA10*	*SMG6*
ADAMTSL3	*CSMD1*	*INA*	*PCDHA2*	*SMIM4*
AGPHD1	*CTNNA1*	*INO80E*	*PCDHA3*	*SNAP91*

AKT3
AL049840.1
ALDOA
AMBRA1
ANKRD44
ANKRD63
ANP32E
APH1A
APOPT1
ARHGAP1
ARL3
ARL6IP4
AS3MT
ASPHD1
ATG13
ATP2A2
ATPAF2
ATXN7
BAG5
BCL11B
BOLL
BTBD18
C10orf32
C11orf31
C11orf87
C12orf42
C12orf65
C12orf79
C16orf86
C16orf92
C1orf132
C1orf51
C1orf54
C2orf47
C2orf69
C2orf82
C3orf49
C4orf27
CA14
CA8
CACNA1C
CACNA1I
CACNB2
CCDC39
CD14
CD46
CTNND1
CTRL
CUL3
CYP17A1
CYP26B1
CYP2D6
DDX28
DFNA5
DGKI
DGKZ
DNAJC19
DND1
DOC2A
DPEP2
DPEP3
DPP4
DPYD
DRD2
DRG2
DUS2L
EDC4
EFHD1
EGR1
ENKD1
EP300
EPC2
EPHX2
ERCC4
ESAM
ESRP2
ETF1
F2
FAM109B
FAM53C
FAM57B
FAM5B
FANCL
FES
FURIN
FUT9
FXR1
GALNT10
GATAD2A
GDPD3
GFOD2
GFRA3
IREB2
ITIH1
ITIH3
ITIH4
KCNB1
KCNJ13
KCNV1
KCTD13
KDM3B
KDM4A
KLC1
L3MBTL2
LCAT
LRP1
LRRC48
LRRIQ3
LUZP2
MAD1L1
MAN2A1
MAN2A2
MAPK3
MARS2
MAU2
MDK
MED19
MEF2C
MIR137
MIR548AJ2
MLL5
MMP16
MPHOSPH9
MPP6
MSANTD2
MSL2
MUSTN1
MYO15A
MYO1A
NAB2
NAGA
NCAN
NCK1
NDUFA13
NDUFA2
NDUFA4L2
NDUFA6
NEK1
PCDHA4
PCDHA5
PCDHA6
PCDHA7
PCDHA8
PCDHA9
PCGEM1
PCGF6
PDCD11
PITPNM2
PJA1
PLA2G15
PLCB2
PLCH2
PLCL1
PLEKHO1
PODXL
PPP1R13B
PPP1R16B
PPP2R3A
PPP4C
PRKD1
PRR12
PRRG2
PSKH1
PSMA4
PSMB10
PSMD6
PTGIS
PTN
PTPRF
PUS7
R3HDM2
RAI1
RANBP10
RANGAP1
RCN3
REEP2
RERE
RFTN2
RGS6
RILPL2
RIMS1
RRAS
SATB2
SBNO1
SNX19
SPCS1
SREBF1
SREBF2
SRPK2
SRR
STAB1
STAC3
STAG1
STAT6
SUGP1
TAC3
TAF5
TAOK2
TBC1D5
TBX6
TCF20
TCF4
THAP11
THOC7
TLE1
TLE3
TM6SF2
TMCO6
TMEM110-MUSTN1
TMEM194A
TMEM219
TMTC1
TMX2
TNFRSF13C
TOM1L2
TRANK1
TRIM8
TRMT61A
TSNARE1
TSNAXIP1
TSR1
TSSK6
TYW5
USMG5
VPS14C
VPS45
VRK2
VSIG2
WBP1L
WBP2NL

CDC25C	*GID4*	*NEK4*	*SCAF1*	*WDR55*
CDK2AP1	*GIGYF2*	*NFATC3*	*SDCCAG8*	*XRCC3*
CENPM	*GLT8D1*	*NGEF*	*Sep-03*	*YPEL3*
CENPT	*GNL3*	*NISCH*	*SERPING1*	*YPEL4*
CHADL	*GOLGA6L4*	*NLGN4X*	*SETD8*	*ZDHHC5*
CHRM4	*GPM6A*	*NOSIP*	*SEZ6L2*	*ZFYVE21*
CHRNA3	*GRAMD1B*	*NRGN*	*SF3B1*	*ZMAT2*
CHRNA5	*GRIA1*	*NRN1L*	*SFXN2*	*ZNF408*
CHRNB4	*GRIN2A*	*NT5C2*	*SGSM2*	*ZNF536*
CILP2	*GRM3*	*NT5DC2*	*SHISA8*	*ZNF804A*
CKAP5	*HAPLN4*	*NUTF2*	*SHMT2*	*ZSCAN2*
CKB	*HARBI1*	*NXPH4*	*SLC12A4*	*ZSWIM6*
CLCN3	*HARS*	*OGFOD2*	*SLC32A1*	
CLP1	*HARS2*	*OSBPL3*	*SLC35G2*	
CLU	*HCN1*	*OTUD7B*	*SLC38A7*	

Table A6.2 List of 97 extended gene regions (EGRs) and corresponding genomic positions.

EGR ID	**hg19 position (Chr, Start, End)**		
PLCH2	1	2407754	2436964
SLC45A1{RERE}	1	7136076	9502487
PTPRF{KDM4A}	1	43166725	44779619
LRRIQ3	1	74309811	74794591
DPYD-MIR137	1	96758959	99100978
VPS45{APH1A OTUD7B}-PLEKHO1{ANP32E APH1A OTUD7B}-C1orf51{ANP32E APH1A C1orf54 CA14 OTUD7B}	1	149815778	151245170
FAM5B	1	176238088	178364194
C1orf132{CD46 CR1L}	1	207005216	209798531
SDCCAG8-AKT3	1	243419307	244282764
VRK2-FANCL	2	57655500	60917597
CYP26B1	2	71908225	72807021
EPC2	2	148355764	150085782
SLC4A10	2	162480845	162841786
DPP4	2	162848755	162931052
ZNF804A	2	184137898	186082304
PCGEM1	2	193614571	193641625
SF3B1{AC073043.2 ANKRD44 BOLL COQ10B HSPD1 HSPE1 MARS2 PLCL1 RFTN2 SATB2 TYW5}-C2orf47{AC073043.2 C2orf69 TYW5}	2	196972133	201568864
CUL3	2	224775116	225814198
GIGYF2{C2orf82 EFHD1 KCNJ13}-NGEF{C2orf82 EFHD1 KCNJ13}	2	232630330	234661428
CNTN4	3	1103727	3458808
TBC1D5	3	17198654	19068576

TRANK1	3	36704648	37283621
GNL3{GLT8D1 ITIH1 ITIH4 NEK4 NISCH NT5DC2 PBRM1 SMIM4 SPCS1 STAB1}-MUSTN1{GLT8D1 ITIH1 ITIH3 ITIH4 NEK4 NISCH NT5DC2 PBRM1 SMIM4 SPCS1 TMEM110-MUSTN1}	3	51523089	53938766
C3orf49-THOC7{ATXN7 PSMD6}	3	63327080	64504050
PCCB{MSL2 NCK1 PPP2R3A SLC35G2}-STAG1{MSL2 NCK1 PPP2R3A SLC35G2}	3	134764021	138361288
CCDC39{FXR1}-DNAJC19{FXR1}	3	179424471	181029187
MIR548AJ2	4	23464696	23464726
SLC39A8	4	103182821	103266655
NEK1{C4orf27 CLCN3}	4	169960467	170720959
GPM6A	4	175672336	178256786
HCN1	5	45255052	46147700
ZSWIM6	5	60347967	60841999
MEF2C	5	88014058	88199922
MAN2A1	5	107783599	110444200
GFRA3{CDC25C EGR1 ETF1 FAM53C KDM3B REEP2}-HSPA9{CDC25C EGR1 ETF1 FAM53C KDM3B REEP2}	5	136974195	139092464
NDUFA2{AC005609.1 CD14 DND1 HARS HARS2 IK PCDHA1 PCDHA10 PCDHA2 PCDHA3 PCDHA4 PCDHA5 PCDHA6 PCDHA7 PCDHA8 PCDHA9 TMCO6 WDR55 ZMAT2}	5	139346411	141030658
GRIA1-GALNT10	5	151894201	154222766
RIMS1	6	71769468	73112845
SNAP91	6	83409589	85562549
FUT9	6	96463845	96663488
MAD1L1	7	1142906	2690329
DFNA5{MPP6 OSBPL3}	7	23393355	25261123
GRM3	7	85207255	87309798
PUS7{SRPK2}	7	104536115	105796193
PODXL	7	130644170	131943940
PTN-DGKI	7	136384831	138163595
CSMD1	8	2792875	4916754
CLU{EPHX2}	8	26433842	28250633
CA8	8	59643721	62485987
MMP16	8	87936967	90632845
KCNV1	8	110809503	111262738
TSNARE1	8	142508513	144110524
TLE1	9	83123419	85560764
CACNB2	10	18223287	18830688
ARL3{AS3MT C10orf32 NT5C2 PDCD11 SFXN2 TRIM8 USMG5 WBP1L}-CYP17A1{AS3MT C10orf32 NT5C2 PDCD11 SFXN2 TAF5 TRIM8 USMG5 WBP1L}-CNNM2{AS3MT NT5C2 PDCD11 SFXN2 TAF5 USMG5 WBP1L}-INA{AS3MT NT5C2 PDCD11 TAF5 USMG5 WBP1L}-PCGF6{AS3MT	10	103581302	105408987

NT5C2 PDCD11 TAF5 USMG5}			
LUZP2	11	23723074	25104186
HARBI1{AMBRA1 ATG13 CHRM4 CKAP5 CREB3L1 DGKZ MDK ZNF408}-F2{AMBRA1 ARHGAP1 ATG13 CKAP5 CREB3L1 ZNF408}	11	45850311	47765368
SERPING1{BTBD18 C11orf31 MED19 TMX2 YPEL4}-CTNND1{BTBD18 C11orf31 CLP1 MED19 TMX2 YPEL4 ZDHHC5}	11	56887198	59411566
C11orf87	11	109203086	109339543
DRD2	11	112171763	114011104
GRAMD1B-NRGN{ESAM MSANTD2 VSIG2}	11	122727700	126065367
SNX19	11	129734531	131543436
IGSF9B	11	132848137	134271219
CACNA1C	12	1243304	3279035
TMTC1	12	29213919	30967424
TMEM194A{NAB2 R3HDM2 SHMT2 TAC3}-LRP1{MYO1A NAB2 NDUFA4L2 NXPH4 R3HDM2 SHMT2 STAC3 STAT6 TAC3}	12	56677347	58245642
C12orf79	12	92378752	92536447
C12orf42	12	103128804	104278072
ATP2A2	12	110719032	110788897
OGFOD2{ABCB9 ARL6IP4 CDK2AP1 MPHOSPH9 PITPNM2 RILPL2 SBNO1 SETD8}-C12orf65{ABCB9 ARL6IP4 CDK2AP1 MPHOSPH9 PITPNM2 RILPL2 SBNO1 SETD8}	12	121722819	124308277
PRKD1	14	29224481	30931651
RGS6{AC005477.1}	14	72399156	73033238
BCL11B	14	99635625	99737822
TRMT61A{APOPT1 BAG5 CKB KLC1 XRCC3 ZFYVE21}-AL049840.1{APOPT1 BAG5 CKB KLC1 PPP1R13B XRCC3 ZFYVE21}	14	103489114	105289120
ANKRD63{PAK6}	15	39906215	40937199
TLE3	15	69724037	71047194
IREB2{AC027228.1 AGPHD1 CHRNA3 CHRNA5 CHRNB4 PSMA4}	15	76785048	80148768
ADAMTSL3	15	83975968	84722793
GOLGA6L4	15	84904525	84914120
ZSCAN2	15	85096833	85681159
MAN2A2{FES FURIN}	15	90682138	92618436
GRIN2A	16	8982345	10981350
ERCC4	16	12704748	14769348
DOC2A{ALDOA ASPHD1 C16orf92 FAM57B GDPD3 HIRIP3 INO80E KCTD13 MAPK3 SEZ6L2 TAOK2 TBX6 TMEM219 YPEL3}-PPP4C{ALDOA ASPHD1 FAM57B GDPD3 HIRIP3 KCTD13 MAPK3 SEZ6L2 TBX6 YPEL3}	16	28839158	30771265
CNOT1{SLC38A7}	16	58134256	58833839
ACD{C16orf86 CENPT CTRL DDX28 DPEP2 DPEP3 DUS2L EDC4 ENKD1 ESRP2 GFOD2 LCAT NRN1L	16	66861562	69771395

NUTF2 PARD6A PLA2G15 PSKH1 PSMB10 RANBP10 SLC12A4 THAP11 TSNAXIP1}-SLC7A6{C16orf86 CTRL DDX28 DPEP2 DPEP3 ENKD1 ESRP2 GFOD2 NRN1L NUTF2 PARD6A PLA2G15 PSKH1 PSMB10 SLC12A4 THAP11 TSNAXIP1}-NFATC3{C16orf86 CTRL DDX28 DPEP2 DPEP3 ENKD1 ESRP2 GFOD2 NRN1L NUTF2 PARD6A PLA2G15 PSKH1 PSMB10 SLC12A4 THAP11 TSNAXIP1}-SLC7A6OS{C16orf86 CTRL DDX28 DPEP2 ENKD1 ESRP2 NUTF2 PARD6A PLA2G15 PSMB10 SLC12A4 TSNAXIP1}			
TSR1{SGSM2 SMG6 SRR}	17	1024099	2933986
RAI1{ATPAF2 DRG2 GID4 LRRC48 MYO15A SREBF1 TOM1L2}	17	16112939	19671395
HAPLN4{CILP2 GATAD2A MAU2 NCAN NDUFA13 PBX4 SUGP1 TM6SF2 TSSK6}	19	18316601	20008019
ZNF536	19	29557055	32850323
RCN3{NOSIP PRR12 PRRG2 RRAS}-SCAF1{NOSIP PRR12 PRRG2 RRAS}	19	49404637	50661517
PPP1R16B{ACTR5 SLC32A1}	20	37291161	37678722
KCNB1-PTGIS	20	47638985	48768940
CACNA1I-L3MBTL2{CHADL RANGAP1}-EP300{CHADL RANGAP1}-CENPM{CYP2D6 FAM109B NAGA SEPT3 SHISA8 SMDT1 SREBF2 TNFRSF13C WBP2NL}-NDUFA6{CYP2D6 FAM109B NAGA SEPT3 SHISA8 SMDT1 SREBF2 TCF20 TNFRSF13C WBP2NL}	22	39291157	43480828
NLGN4X	23	5808083	6666551
CNKSR2	23	20595156	21693982
PJA1	23	67636502	69240034

Table A6.3 Genes found to overlap in the largest components of INTERKR and INTERVC binary Hi-C networks.

Genes			
AC073043.2	*ESAM*	*MSANTD2*	*TCF4*
ATXN7	*ETF1*	*NLGN4X*	*TLE1*
C2orf69	*FAM5B*	*NRGN*	*TMTC1*
CNTN4	*GALNT10*	*PCDHA*	*TYW5-C2orf47*
CSMD1	*GPM6A*	*PDCD11*	*VSIG2*
CTNNA1	*HCN1*	*PLCL1*	*ZNF804A*
CYP26B1	*HSPA9*	*PRKD1*	*ZSWIM6*
DYPD	*IMMP2L*	*PSMD6*	
EGR1	*MMP16*	*RIMS1*	

Table A6.4 SNPs within linkage disequilibrium blocks which satisfy schizophrenia association $p < 0.05$ (figure 6.20).

rsID	Chromosome	Position (hg19)	P-value
rs1978	6	26377573	4.76×10^{-7}
rs9393714	6	26373740	6.93×10^{-7}
rs1977	6	26377546	7.54×10^{-7}
rs12176317	6	26372786	7.80×10^{-7}
rs9358932	6	26362705	9.17×10^{-7}
rs2073529	6	26375159	9.76×10^{-7}
rs9379855	6	26364930	1.06×10^{-6}
rs9393713	6	26373678	1.15×10^{-6}
rs9393708	6	26362643	1.19×10^{-6}
rs9393710	6	26367833	1.21×10^{-6}
rs9379851	6	26354780	1.31×10^{-6}
rs9379858	6	26367689	1.35×10^{-6}
rs9393705	6	26361011	1.52×10^{-6}
rs9366653	6	26354247	1.60×10^{-6}
rs9379859	6	26369549	1.60×10^{-6}
rs9379856	6	26366836	1.66×10^{-6}
rs13218591	6	26376832	6.64×10^{-6}
rs3734536	6	26365346	8.81×10^{-6}
rs12199613	6	26367218	9.59×10^{-6}
rs1985732	6	26376161	9.60×10^{-6}
rs9393709	6	26365147	1.07×10^{-5}
rs6933583	6	26355283	1.38×10^{-5}
rs4712980	6	26355758	1.46×10^{-5}
rs3799378	6	26404374	1.87×10^{-5}
rs4712981	6	26361430	2.15×10^{-5}
rs9379870	6	26374410	2.67×10^{-5}
rs4712984	6	26379537	0.00011
rs2076029	6	26390830	0.000154
rs2073526	6	26374658	0.000161
rs1796518	6	26388672	0.000183
rs12214031	6	26376628	0.00026
rs10456045	6	26404958	0.000437
rs1614887	6	26393021	0.000589
rs10946808	6	26233387	0.002906
rs2858944	16	193586	0.014088
rs6600233	16	143503	0.035136
rs12596306	16	572882	0.013867
rs742285	16	872531	0.000331
rs12445974	16	866053	0.003636
rs3765263	16	840378	0.019352

rs4984931	16	852137	0.021463
rs742319	16	841966	0.030402
rs3765264	16	840409	0.031203
rs3765266	16	840769	0.033247
rs2294451	16	847743	0.035097
rs1078903	16	854689	0.040022

Appendix B

```
PROGRAM Penetrance regulators:
    S = (Size of reduced penetrance gene dataset);
    FOR 1 to S
        Find top ten interacting fragments;
        Store fragments in list;
        Find top ten interacting fragments for controls;
        Store fragments in list for controls;
    END FOR;
    Concatenate case list into one large list;
    Concatenate control list into one large list;
    FOR (Each unique region in case list)
        Construct 2-by-2 table;
        Count frequency of region in case list;
        Store frequency in index (1, 1) of table;
        Subtract frequency from S;
        Store result in index (1, 2) of table;
        Count frequency of region in control list;
        Store frequency in index (2, 1) of table;
        Subtract frequency from S;
        Store result in index (2, 2) of table;
        Perform Fisher's exact test on table;
        Store region, table and result in new table;
    END FOR;
    Sort new table in descending order of table(1, 1) magnitude;
    Print sorted new table;
END.
```

Figure B4.1 Pseudocode for analysis described in section 4.3.2.

```
PROGRAM Third-party regulators:
    Define interaction frequency rank threshold, h;
    S = (Size of reduced penetrance pairs dataset);
    Construct 2-by-2 table;
    FOR 1 to S
        Find top h interacting fragments for gene 1 of pair;
        Store fragments in list 1;
        Find top h interacting fragments for gene 2 of pair;
        Store fragments in list 2;
        Find top h interacting fragments for gene 1 of control pair;
        Store fragments in control list 1;
        Find top h interacting fragments for gene 2 of control pair;
        Store fragments in control list 2;
        Count frequency of identical regions appearing in both list 1 and list 2;
        IF (Frequency > 0)
            THEN Add one to index (1, 1) of table;
            ELSE Add one to index (1, 2) of table;
        END IF;
        Count frequency of identical regions appearing in both control list 1 and control list 2;
        IF (Frequency > 0)
            THEN Add one to index (2, 1) of table;
            ELSE Add one to index (2, 2) of table;
        END IF;
    END FOR;
    Perform Fisher’s exact test on table;
    Print result;
END.
```

Figure B4.2 Pseudocode for analysis described in section 4.3.3.

```
PROGRAM Fusion interactions:
    S = (Size of gene fusions dataset);
    FOR 1 to S
        Find mean (excluding zeroes) of all interactions with fusion gene region;
        Store mean in case list;
        Find mean (excluding zeroes) of all interactions with control region;
        Store mean in control list;
    END FOR;
    Perform two-sample t-test on case list versus control list;
    Print result;
END.
```

Figure B5.1 Pseudocode for analysis described in section 5.3.2.

```
PROGRAM Fusion pair proximity:
    Define interaction frequency rank threshold, h;
    S = (Size of gene fusion pairs dataset);
    Construct 2-by-2 table;
    FOR 1 to S
        IF (Interaction frequency rank of gene fusion pair <= h)
            THEN Add one to index (1, 1) of table;
            ELSE Add one to index (1, 2) of table;
        END IF;
        IF (Interaction frequency rank of control pair <= h)
            THEN Add one to index (2, 1) of table;
            ELSE Add one to index (2, 2) of table;
        END IF;
    END FOR;
    Perform Fisher's exact test on table;
    Print result;
END.
```

Figure B5.2 Pseudocode for analysis described in section 5.3.3.

```
PROGRAM Cell line choice:
    Set interaction frequency threshold, h;
    S = (Size of eQTL pairs dataset);
    Create 3-by-2 table;
    FOR 1 to S
        IF (eQTL pair interaction frequency >= h for GM12878 Hi-C)
            THEN Add one to index (1, 1) of table;
            ELSE Add one to index (1, 2) of table;
        ENDIF;
        IF (eQTL pair interaction frequency >= h for HMEC Hi-C)
            THEN Add one to index (2, 1) of table;
            ELSE Add one to index (2, 2) of table;
        ENDIF;
        IF (eQTL pair interaction frequency >= h for IMR90 Hi-C)
            THEN Add one to index (3, 1) of table;
            ELSE Add one to index (3, 2) of table;
        ENDIF;
    ENDFOR;
    Perform Fisher's exact test between cell line groups of table;
    Print result;
END.
```

Figure B6.1 Pseudocode for analysis described in section 6.3.5.

```
PROGRAM eQTL and Hi-C overlap:
    Set interaction frequency threshold, h;
    Set expression p-value threshold, x;
    S = (Size of eQTL pairs dataset);
    Create 2-by-2 table;
    FOR 1 to S
        IF (Interaction frequency of pair >= h AND expression p-value <= x)
            THEN Add one to index (1, 1) of table;
        ELSEIF (Interaction frequency of pair >= h AND expression p-value > x)
            THEN Add one to index (1, 2) of table;
        ELSEIF (Interaction frequency of pair < h AND expression p-value <= x)
            THEN Add one to index (2, 1) of table;
        ELSEIF (Interaction frequency of pair < h AND expression p-value > x)
            THEN Add one to index (2, 2) of table;
        ENDIF;
    ENDFOR;
    Perform Fisher’s exact test on table;
    Print result;
END.
```

Figure B6.2 Pseudocode for analysis described in section 6.3.6.