Identifying disease-relevant interactions in Schizophrenia

Bathilde Ambroise

Thesis submitted for the degree of Doctor of Philosophy at

Cardiff University

2018

Supervisors:

School of Medicine

Dr Andrew Pocklington

Dr Valentina Escott-Price

School of Computer Science

Dr Irena Spasic

Dr Jianhua Shao

Acknowledgments

Firstly, I would like to thank Andrew Pocklington for his support, advice and patience during all these years, especially towards the end. Without his kindness and his precious help even at the last minute, I would not have made it this far.
I would also like to thank Valentina Escott-Price for her precious advice and the mathematical discussions that we were able to have together.
I am grateful to both of them for enabling me to make it to the finish line.
I must also express gratitude to my two other supervisors Irena Spasic and Jianhua Shao for their valuable guidance.

I would also like to thank Antonio, Katherine and Johanna, Maria, Elliot and June as well as the rest of the BBU and the psychosis team for all their help and many advices.

A special thank to Hayley, Will, Cath, Rose and Neil for including me in the gang and for your kindness.
Evelyne and Robin, my friends through this long journey. For all those times we supported each other, for all the time at the graduate center working on our thesis and for encouraging me at the end, many thanks!
Bibi, pour m'avoir supportée, pour avoir toujours été à portée de messages quand j'en avais besoin. Merci!
Sandrine, Mika, Nico, Max, Mariam et Julie: pour tout votre soutien, votre amitie si precieuse malgre la distance. J'ai hate de celebrer avec vous!

I would like to thank my parents and my sister Adélaïde for helping me throughout the past few years, allowing me to rant and complain but never doubting me. A special mention also goes to my Dad for all his statistical help and little sis for the proofreading. You've been there when I needed it the most and even when everything fell apart you believed in me. Merci énormément à vous trois, je n'y serais pas arrivé sans vous, je vous aime.

Summary

Analyses of genome-wide association study data have demonstrated that there are potentially thousands of loci associated with schizophrenia (Sullivan et al. 2003). Although risk is partially explained by the additive effects of top-ranking polymorphisms, genetic interactions may help to explain additional heritability (Hemani et al. 2014; Zuk et al. 2012). However, attempts to identify disease-associated pair-wise interactions through exhaustive testing have so far been unsuccessful due to the large burden of multiple testing and the absence of easily discoverable interactions of large effect (Moskvina et al. 2011). Here we investigate whether evidence for a contribution to disease risk from SNP-SNP interactions can be found by searching for sets of genes enriched for nominally associated interactions.

When performing interaction analyses covariates were introduced to account for population structure. Where the effect of covariates needs to be accounted for, the most widely used method modifies the basic logistic regression interaction analysis by simply adding covariate terms into the model. The performance of this method was compared to two alternative approaches: adding covariate-SNP interactions terms in addition to the individual covariate terms, as suggested by (Yzerbyt et al. 2004); and testing for interactions in each population separately, then using meta-analysis to combine interaction effects. Results and running time were similar whether SNP-covariate terms were included or not, while the meta-analytic approach was found to be the most efficient in terms of running time.

To try and identify sets of genes enriched for nominally associated interactions, two approaches were investigated: one based on genetic information alone, and one based on functional information using protein-protein interactions (PPI). The first approach analyzed the distribution of interaction p-values after ranking them by the gene-wide main effects of the contributing genes, allowing a comparison to be made between genes with high/low gene-wide association. The second approach asked whether genes encoding directly interacting proteins were enriched for nominally associated interactions, drawing upon two PPI datasets: one from a large experimental (yeast two-hybrid) screen, the other consisting of PPI data curated from the literature. In both of the genetic datasets studied there was evidence for enrichment of nominally associated interactions amongst genes with highest gene-wide association for schizophrenia. There was no evidence for an excess of nominally associated interactions when investigating either PPI dataset.

Table of Contents

Table of Figures

Tables:

Chapter 1 - Introduction

1.1 Schizophrenia

1.1.1 History

The concept of schizophrenia has recent origins despite the fact that some of the specific aspects of the disorder appeared to have been known from the Middle Ages (Jeste et al. 1985). Dr Emil Kraepelin was the first to identify the disease in 1887 under the name of "dementia praecox". It was the first time that schizophrenia was separated from other forms of psychosis. Kraepelin thought that his dementia praecox was a brain disease differentiated from dementia by the fact that it occurred early in life.

Eugen Bleuler introduced the term schizophrenia in 1911. The word schizophrenia comes from the Greeks *schizo* and *phrene*. *Schizo* can be translated as split and *phrene* as mind. Bleuler noted that schizophrenia was very complex in the sense that it seemed to appear as a group of diseases, due to the wide variety of symptoms observed between individuals. He was also the first person to separate the symptoms in two categories: positive symptoms and negative symptoms.

1.1.2 Disease

Schizophrenia is a severe psychiatric disorder, estimated to affect approximately 1% of the population (Owen et al. 2005). The onset of the disease lies in adolescence or early adulthood and often results in a lifetime of illness and treatment. Consequently, schizophrenia has a strong impact on the patient's life, their family and on the public health service.

1.1.2.1 Symptoms

To diagnose schizophrenia, patients have to fulfil certain criteria. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) provides a list of symptoms and behaviours

that enable psychiatrists to identify the disease. As schizophrenia‘s phenotype is very complex and heterogeneous, patients show diverse symptoms. Moreover, schizophrenia is a long-term mental health disorder and symptoms can also evolve. The main symptoms of the disorder can be classified in four categories: positive, negative and cognitive symptoms and disorganization of thinking and behaviour.

Positive symptoms appear to imitate at the excess or distort normal behaviours. Such symptoms can appear and disappear. They can also depend on the treatments that the patient is taking. Patients generally live in a distorted reality. The degree to which a patient is affected can also vary: one person could show a severe form of the symptoms whereas in another patient it would be hardly noticeable. Hallucination is when a person sees, smells, hears or feels things that do not exist. The most well-known symptom amongst individuals suffering from schizophrenia is the hearing of voices. Other hallucinations include seeing people or sensations of being touched when no one is physically present. Delusion is a false belief based on a mistaken or unrealistic view and is not part of the person’s culture. Despite being shown that it is illogical or untrue, patients experiencing both hallucinations and delusions will still hold strongly to their beliefs, as they feel very real.

Negative symptoms include signs that disrupt emotions or behaviours including but not limited to lack of emotions, affective blunting, poverty of speech, or a loss of interest in everyday activities. It is sometimes hard to recognise whether such indications are part of the development of the disease or characteristics caused by something else. These symptoms are the least likely to improve in the patients.

Cognitive symptoms include cognitive deficits such as poor ability to make decisions, problems with memory, the inability to focus or a lack of understanding information. Although cognitive deficiencies are frequently observable, it has to be noted that such

observations vary dramatically across individuals. Moreover, cognitive impairment has only been recently recognised as a fundamental symptom of schizophrenia.

Disorganized thinking is an unusual way of thinking. This happens when the patient cannot connect their thoughts in a logical way or by poverty of thought content (Tandon et al. 2009). This can also be expressed though a patient's difficulty in holding a conversation.

1.1.2.2 Treatment

Schizophrenia is a complex disorder and the cause of the disease is not yet known to its full extent (Tandon et al. 2008). So far, available treatments have been focused on helping patients to deal with their symptoms or to eliminate them. For example Clozaril, a form of Clozapine is an antipsychotic medication.

1.1.3 Epidemiology

To characterise the epidemiology of a disease, two indicators are usually used: incidence and prevalence. The incidence is the number of observed new cases developed over a period of time in a population at risk with the disease. The prevalence is the number of new and previously existing cases who have the disease at a particular time within a defined population.

Numerous epidemiological studies have been done on schizophrenia. An analysis of the prevalence from 188 studies that covered 46 countries calculated the median prevalence estimate at 4.6 per 1,000 (Saha et al. 2005). Regarding the incidence rate, research showed that results range between 0.16 and 0.42 per 1,000 (Jablensky 2000). While differences have been observed between economically rich and more economically impoverished countries, these differences essentially concern the progression and consequence of the disease (Jablensky 2000): there is no evidence that the economic status of a country influences the incidence rates of schizophrenia (Saha et al. 2006). A meta-analysis of all published studies between 1965 and 2001 showed that variation in the incidence of

schizophrenia was highly associated with sex, urbanity and migration (McGrath et al. 2004). Another review also noted that the clinical features of the disorder vary significantly from one patient to another (Rössler et al. 2005).

1.1.4 Genetics of schizophrenia

1.1.4.1 Heritability

Heritability is the proportion of variance in a particular trait, in a particular population, at a particular time that is due to genetic factors.

A given phenotype (P) for a particular trait at a particular time in a particular population can be modelled as the sum of the unobserved genotype (G) and the unobserved environment (E) as follow:

Consequently the variance of the observable phenotype can then be defined:

Where is the phenotypic variance, is the genetic variance and the environmental variance.

The broad sense heritability H^2 is defined as the proportion of the observed trait variances that is due to all genetic factors:

The genetic variance can then be defined as the following:

Where is the variance of additive effects, is the variance of dominant effects and the variance of interaction effects.

The narrow-sense of heritability h^2 is defined as the proportion of the observed trait variances that is due to additive genetic factors only:

Where is the variance of additive effects, is the phenotypic variance.

Schizophrenia is a highly heritable disorder in which genetic factors account for 80% of the variability in liability (Sullivan et al. 2003). Estimates of heritability from twin studies (Cardno and Gottesman 2000) varies from 80-85%.

1.1.4.2 Genome Wide Association Studies

Association studies such as Genome Wide Association Studies (GWAS) take the following approach: the location of the causal variants is not assumed to allow an unbiased search using association principles (Hirschhorn and Daly 2005). Moreover, such a method is more suited to identify variants of smaller effect than those detected by early linkage studies (Risch and Merikangas 1996).

GWAS assume that common variations contribute to the heritability of common diseases (Reich and Lander 2001). This method focuses on comparing allele frequencies between individuals affected by schizophrenia (cases) and healthy individuals (controls). If an allele has a higher frequency in cases than in controls, then this allele will be considered to be associated with the disease. Using such a method has only been possible because of the characterisation of linkage disequilibrium (LD) between SNPs via the HapMap project (International HapMap Consortium 2003). LD exists when genotypes at different loci are not independent of another: there is a non-random association. By sequencing millions of variants, the HapMap project (International HapMap Consortium 2003) examined LD patterns between SNPs and concluded that because neighbouring SNPs are very likely to be correlated, only a subset of variants need to be selected in order to look for association across the genome (Hirschhorn and Daly 2005). Consequently, if a SNP has been found to be associated with the disease, either this marker is causal (directly associated) or it is in LD with the causal variant (Bergen and Petryshen 2012), which is more likely. Statistical

analysis is carried out to investigate the likelihood of each variant to be associated with the disease (or trait). The power of GWAS depends on the effect size, the frequency of the SNPs and the sample size (Klein 2007).

Using a case-control population, hundreds of thousands of variants are genotyped across the genome using commercial SNPs arrays. Imputation methods are used in order to increase the number of SNPs: the genotype of one variant can be used to predict the genotype of the neighbouring markers as neighbouring SNPs are almost always at least partially correlated with each other (Hirschhorn and Daly 2005).

In order to evaluate interesting results and to limit false positive results, a threshold for genome-wide significance is used. Conservative approaches such as Bonferroni determine the threshold by using the number of tests performed. However it has been showed that the number of tests should not be the only determining factor but rather best practice should estimate the probability to obtain a true association at any loci. Indeed the strength of evidence in a GWAS depends on the likely number of true associations and on the power to detect true interactions. But the latter depends on effect sizes and sample size. As a result a threshold of $5e^{-8}$ has become the significance standard in GWAS (International HapMap Consortium 2005).

One of the first GWAS of schizophrenia was published in 2006 (Mah et al. 2006) showing only suggestive association. Other GWAS results for schizophrenia then came from the results of three studies (Stefansson et al. 2009; Shi et al. 2009; International Schizophrenia Consortium 2008) using the following samples: the International Schizophrenia Consortium sample (3,322 cases and 3,587 controls), the Molecular Genetics of Schizophrenia (MGS) sample (3,967 cases, 3,626 controls), the SGENE sample (2,663 cases and 13,498 controls). Those three studies were combined and produced an important genome-wide significant result for schizophrenia: an association signal was found within the major histocompatibility complex (MHC) region. In addition, the SNP for the gene

ZNF804A (Williams et al. 2011) reached genome-wide significance. These studies showed that schizophrenia was a highly polygenic disorder, even more than expected (Kavanagh et al. 2015).

Further efforts have been made and sample size has been increased thanks to international collaborations. The Psychiatric Genomics Consortium has successfully coordinated the largest collaboration for schizophrenia. The latest GWAS results showed 108 loci associated with schizophrenia, 83 of which have not been observed before (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). This could lead to the identification of new mechanisms involved in schizophrenia (Doherty et al. 2012).

However, as with every method, GWAS have limitations and including data from different samples comes with challenges. The lack of well-defined cases and controls, heterogeneity across samples or population stratification are commonly cited as limitations of the method (McCarthy et al. 2008). In addition to a very high number of SNPs being processed, there is a pressing need for a strict control of multiple testing in order to limit the discovery of false positive results.

To overcome that difficulty, other methods have been used such as gene-based approaches. These methods consider a group of SNPs instead of a single marker. It is then possible to include all the variations from each SNP within a gene and to obtain more functional information. One commonly used method is based on the smallest SNP p-value within a gene and on the number of SNPs within the same gene. Moreover, as the number of genes is significantly smaller than the number of markers, it eases the multiple testing problem (Hirschhorn and Daly 2005). However the presence of LD between SNPs implies that tests are not independent rending the correction applied by some methods insufficient.

Other approaches have been based on the Brown method (Brown 1975) to combine non-independent p-values. This method is derived from Fisher's combined probability test

(Littell and Folks 1971) and uses a theoretical approximation of Fisher's statistics (Moskvina et al 2011).

The approximation Fisher's statistic combines probabilities and has a chi-square distribution (2N degrees of freedom, N being the number of variants) as described below:

with representing the gene wide p-values for the gene and is the number of values being combined (here k = 2).

It calculates the significance of a set of variants in a gene by combining p-values from SNPs and by taking into account the number of variants and the LD between them. The resulting gene-wide significance p-value reflects the degree of association of a gene with the disease.

Moreover, when looking for rare variants through linkage disequilibrium with common SNPs, detecting such variants with a GWAS can be very challenging. In conclusion, GWAS approaches have allowed for the discovery of new loci associated with susceptibility for schizophrenia. But this approach is also complementary with the previously known method such as linkage studies.

1.1.4.3 CNV studies

Copy Number Variations (CNVs) are structural variants. A CNV is a portion of DNA larger than 1000 bp that is either duplicated or deleted. As a result one individual could have a different number of copy of any genes. Algorithms have been developed over the years in order to detect such variations and to allow genome-wide calling using data from GWAS genotyping (McCarroll et al. 2006).

Several studies have showed that individuals affected by schizophrenia tend to have an increased number of CNVs (Walsh et al. 2008; International Schizophrenia Consortium 2008; Kirov et al. 2009). It seems that the effect size of CNVs is larger than those detected

by GWAS. Moreover it has also been reported that CNVs associated with schizophrenia are involved in other disorders including autism and ADHD (Sebat et al. 2007; Williams et al. 2010).

1.1.4.4 Rare variants

Rare variants correspond to alleles having a population frequency under 1%.

In order to detect rare risk variants that contribute to the heritability, variants need to have a relatively large effect size. However there might be rare variants of small effect that contribute to disease but that are extremely hard to detect even when using a large sample size.

Rare variants with large enough effects cannot be detected by common methods such as GWAS, Next Generation Sequencing (NGS) methods need to be used, for example exome sequencing or whole genome sequencing. Whole genome sequencing involves obtaining the sequences of the entire genome of an individual, whereas exome sequencing targets the protein coding genes in an individual's DNA. Exome sequencing is often preferred to whole genome sequencing as its cost is reduced hence preventing the limitation of small studies.

Recently two studies (Purcell et al. 2014; Fromer et al. 2014) took a closer look at rare variants that could potentially contribute to schizophrenia. The first study (Purcell et al. 2014) compared the exome sequences from 1,500 cases and 2,500 controls of a Swedish population.

The second study (Fromer et al. 2014) looked for de novo mutations within protein-coding genes by analysing 600 trios (affected probands with healthy parents). De novo mutations are mutations that are present in the offspring but that does not exist in both parents.

Both studies confirmed the polygenic nature of schizophrenia and that the disease tends to involve functionally related genes: genes related to neuronal function or synaptic signalling.

1.2 Interaction and epistasis

1.2.1 Gene-gene interactions

1.2.1.1 History

The idea of gene-gene interaction was first introduced by William Bateson in 1909 in his book Mendel's Principle of Heredity. By looking at flower colour in peas and combs in chicken, he observed that the transmission rates of some variants deviated from the Mendelian ratios. In his report he suggested that a pair of gene alleles could affect the alleles from another gene. Translating to genetics, one variant could prevent another one from expressing its effect. This can be viewed as an extension of the concept of dominant and recessive genes: one gene has an ascendant effect over another one. As a consequence, in the presence of a variant, a phenotypic change could be observed or could modify a mechanism of gene expression.

In 1918, Fisher gave another definition of epistasis. He argued that alleles from different genes could have an additive effect on the considered phenotype and that any deviation from this effect should be considered as epistasis. This definition made it possible to define a mathematical model that describes the relationship between a phenotype and a genotype under or not the influence of an interaction.

1.2.1.2 Definition

Epistasis could simply be defined as an interaction between genes that affect a phenotype. However distinction is often made between three type of epistasis (Moore and Williams 2005): compositional epistasis, functional epistasis and statistical epistasis.

Compositional epistasis is comparable to Bateson's original definition. It refers to the masking effect of one locus to another locus (Moore and Williams 2005).

Functional epistasis corresponds to physical interactions among proteins or other molecules with an impact on the phenotype. This type of epistasis cannot be inferred from genetic data, it needs biological experimental validation.

Statistical epistasis is closer to Fisher's definition of epistasis. It refers to a phenotypic deviation from an additive effect due to combinations of loci.

The latter definition allows the use of mathematical modelling to detect epistasis in disease traits and will be the one used in this thesis. The working hypothesis is that if two SNPs interact to increase the disease risk, then the combination of alleles associated with increased risk will occur more frequently within the case population than within the control population.

1.2.1.3 Heritability and gene-gene interactions

While it has been established that there are potentially hundreds of loci associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014) these findings do not fully explain the risk of disease: only a proportion of the heritability is explained (Lee et al. 2012). In the case of schizophrenia, the latest estimates suggests that one half to a third of the genetic contribution of risk is captured by the variants detected by GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). This 'missing' heritability could also be attributed to gene-gene or gene-environment interactions (Zuk et al. 2012). Indeed, other non-additive effects are likely to contribute to disease heritability although the size of their contribution is difficult to estimate (Zuk et al. 2012).

1.2.1.4 Evidence for gene-gene interactions

From a biological standpoint, there is no a priori reason to expect that traits should only be additive (Zuk et al. 2012). Evidence for genetic interactions has been reported in many model organisms; for example detection of two-loci interactions in a yeast cross (Bloom et

al. 2013), positive epistasis involving essential genes in Escherichia coli and Saccharomyces cerevisiae (He et al. 2010).

Many methods have been developed to identify genetic interactions in human GWAS data (Wei et al. 2014) and will be discussed in the next section.

Even though most of the studies only looked at pair-wise interactions between two SNPs, there is a possibility that high-order interactions also play a role. However these are harder to detect: the number of interactions to calculate is extremely high and as a result very large sample sizes are needed (Cordell 2009).

1.2.1.4.1 In complex disorder

Several large-scale interaction studies have now evaluated evidence for interactions across the genome for several complex disorders (Wei et al. 2012). Among those studies, different methodologies have been used to evaluate gene-gene interactions. For example a genome-wide interaction-based association using data from the Wellcome Trust Case–Control Consortium identified interaction for Crohn's disease and coronary artery disease (Liu et al. 2011). Still using the WTCCC data, two genome-wide searches for pairwise interactions in each of the seven traits studied reported significant interactions (Lippert et al. 2013; Wan et al. 2010) however many detected effects were in the MHC region and replication was not attempted.

Significant interactions between SNPs have also been reported for Bipolar Disorder (Prabhu and Pe'er 2012).

Multiple sclerosis is another complex trait in which epistasis has been demonstrated to have an impact (Lincoln et al. 2009; Gregersen et al. 2006). Analysis in human populations showed an association between a form of multiple sclerosis and interacting loci (Gregersen et al. 2006).

Unfortunately out of these potential interactions, only a few have a functional basis (Phillips 2008).

1.2.1.4.2 In schizophrenia

Genome-wide interaction studies have been quite unexplored in schizophrenia. A few studies have provided evidence supporting a role for gene-gene interactions in the disease. Epistatic interaction between DISC1, CIT and NDEL1 have been reported to impact risk for schizophrenia and was validated by functional neuroimaging (Nicodemus et al. 2010). The same group also reported significant interactions between the genes NRG1, ERBB4, and AKT1 (Nicodemus et al. 2010). Other studies have also shown evidence that genes interact in schizophrenia the two genes DISC1 and PDE4B have been found to interact (Millar et al. 2005).

1.2.2 Methods to detect gene-gene interactions

Different methods have been explored in order to detect epistasis (Cordell 2009). These methods can be classified into broad groups, described in more detail below: regression based methods, LD-haplotype based methods, Bayesian methods, Data filtering methods, artificial intelligence based methods.

1.2.2.1 Regression-based methods

Regression-based methods are the most frequently used approach to calculate statistical gene-gene interactions. Such models test whether the relationship between one or several predictor variables and an outcome (phenotype) variable is captured by a linear or logistic model (Cordell 2009). In order to perform such an analysis it is necessary to test for interactions by comparing two models: one containing the interaction term and one without.

Considering a case-control study this can be translated into a mathematical model.

Let and be two independent variables with three levels (0/1/2) corresponding to the genotypes of the considered markers (aa/Aa/AA).

Let be the dependent variable that represents the disease status of each individual. In a case-control analysis, this outcome variable would be the log odds of the phenotype: the disease status of an individual (i.e. case or control). This variable is a function of the previous variables , and , the latter represents the interaction.

Let the standardized beta coefficients: is the Y-intercept.

Let be the random error component.

Translating this into a mathematical formula would correspond to the following equation:

(1)

Testing for interaction between the two markers corresponds to a comparison of the above model (which contains both the main effect and the interaction terms) with another model where only the main effects are present, testing whether the regression coefficients associated with the interaction term in the above equation equal zero or not. This corresponds to a one-degree freedom test of .

While it is possible to use GWAS data to perform a genome-wide study of epistasis (Wei et al. 2014), this approach suffers from several limitations. The sample size required to detect interaction signals needs very large. Indeed it scales inversely with the square of the effect size (Zuk et al. 2012): for n loci, the sample size to detect the n^2 interactions scales with n^4. As a result the power to detect interactions in current studies is low (Zuk et al. 2012). Additionally, genome-wide interaction studies suffer from the high burden of multiple test correction. To account for all pair-wise interactions between markers, ——— interactions must be calculated. For example using 5,000 markers will result in 12,497,500 interactions calculated. If the Bonferroni multiple testing correction were to be used, the level of detection for a single test p-value would be $4e^{-9}$.

1.2.2.2 LD and haplotype based methods

LD based methods compare LD between SNPs within cases and controls in order to detect interactions. Studies have been performed and seem to detect interactions within genes using LD rather than within SNPs (He et al. 2011).

Haplotype based methods infer haplotypes from genotypes issued from GWAS and use estimates of linkage between SNPs (Zhang et al. 2012). However simulation studies have shown a greater risk of type I error when the two SNPs are highly correlated or have significant main effects (Ueki and Cordell 2012).

1.2.2.3 Bayesian methods

Bayesian methods are based upon Thomas Bayes theorem, which calculate conditional probabilities based on prior distributions of parameters in a model as well as in the observed data (Wei et al. 2014). Such methods offer a different approach for selecting key predictor variables (and interactions between them) in order to best predict the phenotype (Cordell 2009). Because of the specification of prior distributions of the parameters, this approach differs from frequentist-based statistics (Cordell 2009).

For example, the Bayesian epistasis association-mapping (BEAM) algorithm identified epistasis associations in case-control studies using a Markov chain Monte Carlo method, incorporating prior knowledge about each marker in order to identify interactions (Zhang and Liu 2007).

In addition several studies have tried to combine a Bayesian framework with generalized linear models in order to detect epistasis (Yi et al. 2011). This approach can be advantageous as it can take into account covariates or gene-environment interactions.

1.2.2.4 Data filtering methods

Data filtering methods make use of biological knowledge (Turner et al. 2011) such as disease pathways, protein-protein interactions or features such as the frequency of a

variant (Ackermann and Beyer 2012) in order to select a subset of SNPs prior to pair-wise interaction analysis. By identifying groups of markers between which interactions are more likely to appear, interactions may be easier to detect due to the reduced multiple testing burden. However prior selection of the SNPs could lead to miss detection of certain interactions as SNPs that interact together might not have been all selected (Wei et al. 2014).

Data-filtering methods considerably improve the interpretability of results especially when the analysis is driven by a biological hypothesis. However such methods can be a source of bias due to their initial hypothesis (Wei et al. 2014): analysis implies less tests to perform and less severe correction.

1.2.2.5 Machine Learning

In recent years, machine learning methods and data-mining algorithms have been used to search for epistasis (Wei et al. 2014). Despite limited success so far, these methods could be a potential asset in the search for higher order interactions. Regression models are limited in such analyses due to the curse of dimensionality: if the number of predictors increases, so does the number of interactions in an exponential way. Due to this combinatorial complexity, regression methods are limited in the number of predictors used in their analysis and machine learning based methods could be useful to deal with this issue (Hu et al. 2013). However, despite the fact that computational efficiency improves, even machine-learning methods struggle to analyse higher order interactions (Cordell 2009).

1.2.3 Challenges to detect epistasis

When performing genetic interaction analysis, many challenges arise and power to detect interactions can be influenced by many factors described below (Wei et al. 2014).

1.2.3.1 Linkage disequilibrium

While the most highly associated variants detected by GWAS may be true causal variants, they are more likely to be in LD with the true causal variants. In the latter case a problem arises as the variance explained by such SNPs would be less than the variance explained by the causal markers (Wei et al. 2014). Indeed, the additive effect of the observed marker is correlated with the LD between that marker and the causal variants, thus if there is low LD between the causal and the observed SNPs then the additive effect will be little and any epistatic effect will be very hard to detect. To counteract this, denser genotype or high quality imputation can be use and could help to detect epistasis.

1.2.3.2 Curse of dimensionality

When performing a gene-gene interactions analysis, all pair-wise interactions between all the SNPs are calculated, assuming the analysis is reduced to a search for binary interactions. In that case, the number of possible combinations increases exponentially. If N markers are considered, then the number of calculated interactions is ———. If higher-order interactions are calculated, this number rises even more. With so many results, there is a risk that any true signal will be obscured by the noise produced by such analysis: this is the curse of dimensionality (Wei et al. 2014).

Moreover the significance level required to survive multiple testing correction is much more stringent for interactions than main effect association. This reinforces the need to increase the sample size in order to detect any epistatic effect (Wei et al. 2014).

1.2.3.3 Replication

In order to confirm putative genetic interactions, results need to be replicated. However this has proven to be difficult for epistasis analysis (Combarros et al. 2009): replication rates for gene-gene interactions are expected to be lower than those for additive effect studies (Wei et al. 2014). Indeed, finding the same direction of effect for one interaction

within two independent populations will be extremely hard due to LD differences between the populations. When looking at variants in high LD within a sample, the same variants might not have the same LD within a population resulting in bias within the discovery sample (Cordell 2009). The use of very dense and similar data could help to overcome this problem.

1.2.4 Choice of methodology

In this thesis, two methods were combined to calculate SNP-SNP interactions: filtering on LD and logistic regression model. By filtering on LD, the number of variants used in the interaction analysis is significantly reduced. It allows avoiding collinearity problems and to limit the number of interactions to calculate as well as the multiple tests burden. As interaction analysis is time consuming and computationally intensive reducing the number of variants is vital for feasibility. However by applying such filter information can be lost (SNPs that interact together might bot be selected) but the variants analysed are semi-independent.

In addition the logistic regression model is the most natural and used model to calculate interactions (Cordell 2009). Many tools such as Plink (Purcell et al. 2007; Chang et al. 2015) are available to use this method.

1.3 Protein-protein interactions

1.3.1 Introduction

Protein-protein interactions (PPIs) have a key role in many processes in a cell. Activities of PPI complexes range from protein folding to transport, degradation, transcription, transduction or cell signalling for example. This makes PPI complexes one of the most important components in a cell.

The human genome contains approximately 19,000 protein-coding genes suggesting that the possible number of interactions is very large and that the number of discovered

interactions is probably very low (Rual et al. 2005). Several powerful methodologies and techniques have been developed to discover protein–protein interactions data such as Yeast two-hybrid (Ito et al. 2001), co-purification or fluorescence energy transfer.

Finding PPIs has been an important challenge in the last decade to identify all the complex mechanisms used at a cellular level. With a better knowledge of PPIs it is possible to understand a protein's function and behaviour or to characterise proteins complexes and pathways (Koh et al. 2012).

1.3.2 Databases

1.3.2.1 Primary and secondary databases

Primary databases include interactions that are experimentally determined. Those databases generally contain information on the data model and the data extraction method that has been used. New records are added after manual curation of interactions from the literature. In some cases, users are allowed to submit their own interactions pending verification. Most of those databases contain added information such as functional annotations, sequence information, gene references.

Some example of this resources includes but is not limited to the Biomolecular Interaction Network Database (BIND) (Bader et al. 2003), the Biological General Repository for Interaction Datasets (BioGRID) (Stark et al. 2011), the Database of Interaction Proteins (DIP) (Salwinski et al. 2004), the Human Protein Reference Database (HPRD) (Mishra et al. 2006), the protein InterAction database (IntAct)(Kerrien et al. 2007), the Molecular INTeraction database (MINT) (Ceol et al. 2010)and the Munich Information Center for Protein Sequence (MIPS)(Pagel et al. 2005). Table 1.1 displays the main characteristics of primary databases including the number of interactions available in each of them.

Often called meta-databases, secondary databases combine interactions from several primary databases in one repository. For example the IMEx consortium combines interaction data from 10 databases (Orchard et al. 2007).

1.3.2.2 Predictive databases

Predictive databases contain interactions that have been identified using either algorithms to predict potential interactions using existing data (verified interactions) from curated databases, or the protein structures to determine potential interactions.

1.3.2.3 Limitations

Despite the effective techniques available to experimentally detect protein-protein interactions (Klapa et al. 2013), the above-mentioned databases have limitations. Many false positives and false negatives rates are founds within the findings (Berggård et al. 2007). Moreover the content of the databases often overlaps and is redundant (Cusick et al. 2009).

Acronym	Full Name	Species	# proteins	# interactions	# proteins (human)	# interactions (human)
BIND	Biomolecular Interaction Network Database	All	31,972	58 266	NA	NA
BioGRID	Biological General Repository for Interaction Datasets	All(H)	18401	147500	18401	147500
DIP	Database of Interacting Proteins	All(H)	27098	78191	4283	7140
HPRD	Human Protein Reference Database	Human	30047	41327	30047	41327
IntAct	IntAct Molecular Interaction Database	All	83417	454515	NA	NA
MINT	Molecular Interaction Database	All(H)	35528	241458	8751	26830
MIPS-MPPI	MIPS Mammanlian Protein-protein Interaction Database	Mammals	NA	NA	NA	NA

Table 1.1: Main characteristics (acronym, name, species, number of proteins and interactions total, number of proteins and interactions in human data) of primary databases

Many newly discovered PPI published in small-scale studies are not captured by existing interaction databases, resulting in a huge amount of data only available through curating the literature using databases such as PubMed (Figure 1.1). Manually extracting such information is infeasible on a large scale, especially for data-driven approaches.

In addition there is a need to further analyse the details of PPIs experimentally identified on order to better verify the accuracy of experiments.

All those arguments explain the pressing need to use text-mining methods.

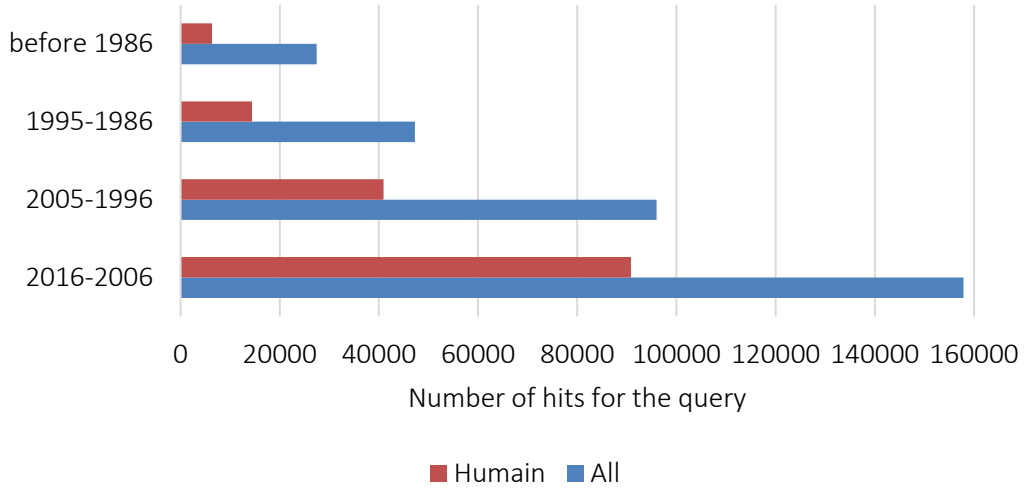


Figure 1.1: Growth of the PubMed database when searching for "protein-protein interactions"

1.3.3 Use of text-mining to extract PPI from the literature

The number of available literature is growing at an exponential rate as generating experimental data has become easier. Specifically for protein-protein interactions the growing trend is clearly visible (Figure 1.1).

Retrieving interactions from the published literature has become an important task. However such analysis can be extremely time-consuming due to the huge amount of data to process. Automated analysis of text can help researchers to evaluate the available literature.

Text mining was first defined by Marti Hearst as the following: “Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise "hidden" meanings.”

Text mining allows the extraction of precise information (PPI for example) based on more than just a simple search for keywords. It searches for entities, concepts, relationships, phrases, sentences or information in a specified context. In order to achieve that task, it is possible to use different techniques such as text processing or machine learning techniques. Three main steps are part of the text mining process: information retrieval, Name Entity Recognition and information extraction.

Information retrieval is the first phase of a text-mining process. In order to extract information from the literature, the relevant texts need to be retrieved. This step is usually performed using keywords to query a database such as PubMed. It is also possible to use other source such as patients’ records for example. PubMed is often chosen as it includes Medline as a subset and allow the use of Mesh (MEdical Subject Headings) terms.

The second step of the text-mining process is the Name Entity Recognition (NER). NER is the use of search algorithms in order to find occurrence of specific keyword such as protein or gene names for example. This complex process is the key to the text mining process. NER techniques can allow searching for several keywords associated with unique entity. For example, it is possible to search for a gene by its full name or its symbol and it will be identified as the same entity. Some text mining tools rely on machine learning techniques for NER using Hidden Markov Models (Zhang et al. 2004) or support vector machines (Habib and Kalita 2010). These techniques can be combined with rule-based methods (Fleuren and Alkema 2015).

The Information extraction process is purely the detection of the relationship between the elements identified by NER. Co-occurrence and natural language processing (NLP) are the

most commonly used methods. Co-occurrence methods rely on the idea that if two entities are related then they will often appear in the text together (Jensen et al. 2006). Such methods can correct for false positives by using a scoring system based on frequency or degree of significance (Alako et al. 2005). This method is easy to implement but tends to have a lower precision than NLP–based methods. NLP methods are based on the structure of a specific language and hence linked to detect information in that language. Additionally NLP methods can provide information about the type of the relationship between two identities (Ben-Hur and Noble 2005). Compared to co-occurrence method, the precision of NLP method is higher but this method can be limited as the detection of relationship between entities is only possible by pre-defining the relationships that are searched for.

Additionally, text mining competitions as the BioCreative challenge (Krallinger et al. 2011) have been taking place in the hope of improving the development of text mining tools (Fleuren and Alkema 2015).

1.4 Aims and objective of the thesis

This thesis is divided in two parts.

The first aim was the comparison of different methods to identify genetic interactions by taking into account population structure.

The second objective was to try to identify sets of genes enriched for SNP-SNP interactions by investigating two different approaches. The first approach was based on genetic information alone. The second approach was based on functional information using protein-protein interactions.

1.5 Chapter's outline

Chapter 2 details the two GWAS datasets as well as the quality control steps applied to it. These datasets will be used in Chapters 3, 4 and 5.

Chapter 3 explores different ways to account for population structure. Three different models that include covariates into an interaction analysis were used on the same GWAS dataset. Strength and limitations of each method were evaluated.

Chapter 4 assesses interactions within two independent GWAS datasets. By calculating SNPs pair-wise interactions in both datasets, the distribution of interaction p-values is analysed after ranking them by the gene-wide main effects.

Chapter 5 investigates whether protein-protein interactions can be used to identify subsets of genes between which significant interactions are more likely to be present. Several PPI datasets are compared.

Chapter 6 will conclude the thesis with a general discussion of the findings, their implications and limitations.

Chapter 2 - Description of datasets

2.1 Introduction

This chapter describes the ISC (International Schizophrenia Consortium 2008) and CLOZUK (Hamshere et al. 2013) Genome-Wide Association Studies (GWAS) datasets that are used in the thesis and the quality control (QC) tests that were applied to them. It is known that the ability of GWAS to recognize true associated genetic variants relies on the overall quality of the data (Turner et al. 2011). Moreover bad quality samples can lead to the detection of false positive and negative associations in the data (Turner et al. 2011). In addition to this QC is an important step for the reliability of case-control studies (Blomgren et al. 2006).

2.2 Quality controls applied: overview

All of the QC procedures applied to the two datasets are detailed in the section below: the chosen order followed best practice and commonly used QC methods for GWAS (Anderson et al. 2010; Wellcome Trust Case Control Consortium 2007; Turner et al. 2011). The QC steps at the exception of the Principal Component Analysis (PCA) were performed using the genetic analysis tool-kit PLINK (Purcell et al. 2007). The PCA was done using the Eigenstrat software (Price et al. 2006).

2.2.1 Call rates

Call rates were checked at both variant and individual levels for missing or incomplete data in order to exclude them from the analysis. In case-control studies, it is essential to check for significant differences in individual call rates between sub populations to ensure that the combined set will be homogenous (Anderson et al. 2010). Furthermore, markers with

below-average call rates indicate low DNA quality and/or concentration and need to be excluded from the analysis.

Two different thresholds for the variants call rates were compared: 1% and 2%. These threshold values are standardly used and indicate a good coverage of the SNPs. In addition to this, by comparing the different thresholds, it is ensured that there will be a fair balance between having the best genotyping quality possible and dropping a minimum number of samples and SNPs (S. Turner et al. 2011). After comparison of the results, the chosen threshold was applied to each dataset prior to similar investigation of individual call rates.

2.2.2 Minor Allele Frequency (MAF)

The MAF threshold applied varies between studies, usually lying between 1% and 5% (Anderson et al. 2010). By applying such thresholds and removing rare variants from the analysis, the multiple testing and computational burdens are reduced (S. Turner et al. 2011) . This has little impact on studies given that the power of detection of associations for these variants is low (Morris and Zeggini 2010).

Moreover, variants with very low frequency can potentially generate additive effects statistically (Gibson 2012), which could introduce a bias in an interaction analysis if those variants were included.

2.2.3 Heterozygosity

The distribution of mean heterozygosity for all individuals was inspected in order to identify which individuals to remove: samples with unusually high heterozygosity indicate possible sample contamination whereas samples with low heterozygosity indicate samples that do not belong to the population (Anderson et al. 2010). Similarly, the distribution of the inbreeding coefficient was inspected: samples with unusually low inbreeding coefficient indicate contamination. In addition, samples with high inbreeding coefficient also need to be removed because standard case-control analyses assume individuals are independent.

Individuals with outlying heterozygosity rates and outlying inbreeding coefficients were identified and removed from the analysis.

2.2.4 Cryptic relatedness

Relatedness among samples leads to an over-representation of selected alleles and can confound the analysis and the discovery of true associated variants (S. Turner et al. 2011). Related individuals were then excluded from the analysis. The identity by descent (IBD) threshold used was 0.1875 as it is intermediate between second and third degree relative (Anderson et al. 2010).

2.2.5 Hardy Weinberg

Markers with significant deviation from the Hardy Weinberg equilibrium need to be excluded as this could indicate potential genotype errors (Anderson et al. 2010). The threshold was decided at 10^{-4} for both cases and controls.

2.2.6 Principal Component Analysis (PCA)

PCA was used to investigate differences between cases and controls due to ancestry dissimilarities. Following the results of the PCA, a clustering algorithm was used to select cases and controls that couldn't be separated in two different groups and to exclude outlier individuals. This was done using the library Mclust in R (Fraley and Raftery 2002).

2.3 ISC Dataset

2.3.1 Introduction

This GWA study of schizophrenia, performed by the International Schizophrenia Consortium (International Schizophrenia Consortium 2008) is used in Chapter 3 and 4 of the thesis.

2.3.2 Sample details

The initial dataset consists of a total of 3,322 individuals with schizophrenia and 3,587 healthy subjects from 8 different populations (Table 2.1).

Sample	Ancestry	Cases (N)	Controls (N)
University of Aberdeen	Scottish	720	702
University College London	British	523	505
Portuguese Island Collection	Portuguese	347	216
Karolinksa Institutet	Swedish	170	170
Karolinska Institutet	Swedish	390	230
Cardiff University	Bulgarian	528	611
Trinity College Dublin	Irish	275	866
University of Edinburgh	Scottish	369	287
TOTAL		3,322	3,587

Table 2.1: Origin of case and control samples for all individuals (N=6909) by samples in the initial ISC dataset.

All of the cases had a diagnosis of schizophrenia based upon DSM-IV, ICD-10 or ascertainment through hospital records. The general population of each site was used to draw controls often using blood banks. Controls were screened for mental illness in most samples.

2.3.3 Genotyping

DNA was extracted from whole blood. The genotyping was performed at the Broad Institute of Harvard and MIT in Boston, USA. Different chips were used: the Affymetrix Genome-Wide Human SNP 5.0 and 6.0 Arrays. To reduce the batch effect, duplicated, poorly genotyped and contaminated samples were removed from the dataset.

More details on the samples and the GWAS are available in the primary manuscript (International Schizophrenia Consortium, 2008).

2.3.4 Quality controls

Initially, the dataset contained 6,909 individuals (3,587 controls and 3,322 cases) and 739,995 SNPs were genotyped.

2.3.4.1 Call rates

Call rates for both individuals and variants were investigated separately in each sub-population: SNP coverage was investigated before the individuals. After comparing the two different thresholds (Table 2.2), the threshold for analysis was established at 2%: the best balance between the number of variants to exclude and the best genotyping quality possible. After removing SNPs with call rates above this threshold in each sub-population, all call rates were above 0.96 in every sub-population (Figure 2.1). After exclusions, 246,455 SNPs remained. No individuals were excluded.

Pop	ISC1	ISC2	ISC3	ISC4	ISC5	ISC6	ISC7	ISC8
1%	75,608	102,262	101,397	66,942	419,822	420,861	378,894	470,808
2%	42,951	57,612	53,117	35,864	392,891	397,529	366,999	446,025

Table 2.2: Number of SNPs to be removed in each chip using the different cut-off thresholds (1% and 2%) for call rates

2.3.4.2 Heterozygosity

No specific threshold was applied, as the results were acceptable in the first instance: no outlier was identified (Figure 2.2).

2.3.4.3 Cryptic Relatedness

No individual was above the chose IBD threshold.

2.3.4.4 Hardy Weinberg

251 variants were excluded.

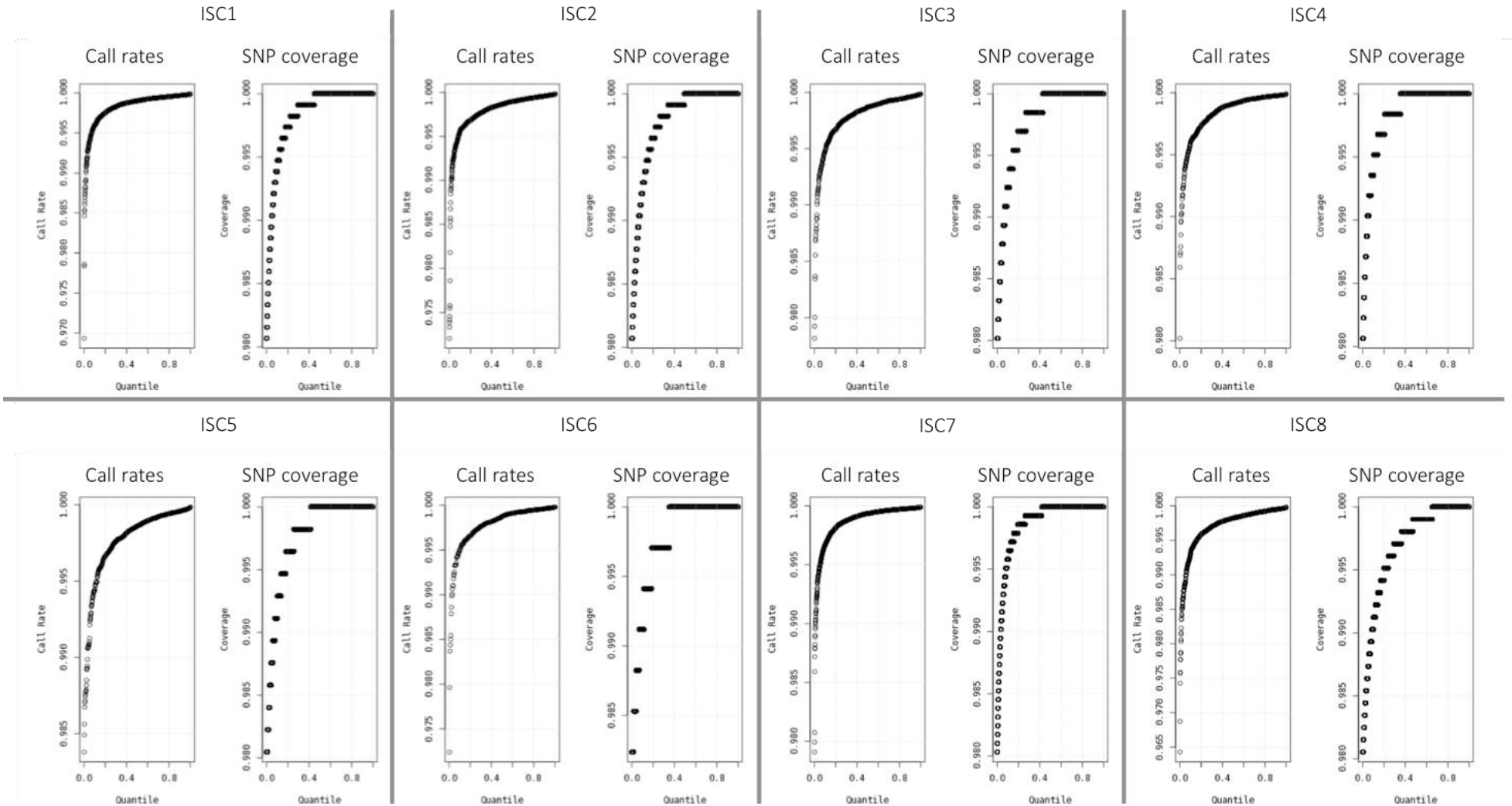


Figure 2.1: Ordered individual call rates and Ordered SNP coverage in each population of the ISC dataset after QC.

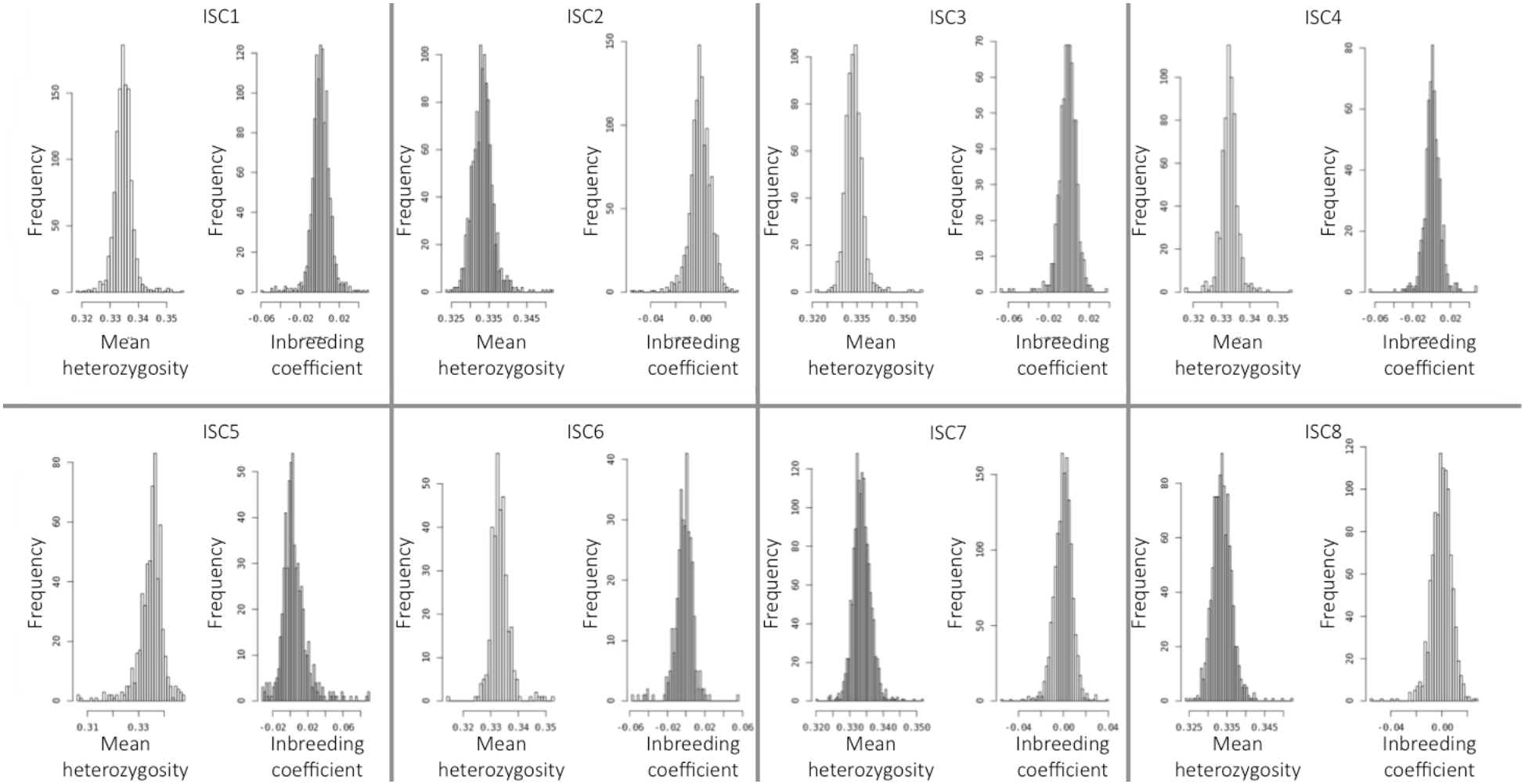


Figure 2.2: Histograms of the mean heterozygosity and of the inbreeding coefficient for the 8 populations of the ISC dataset. No unusual sample is visible as the histograms shows a normal distribution in each chip.

2.3.4.5 Additional Information

For every SNP in the dataset, minor allele frequencies were above 1%, suggesting that rare variants were previously excluded.

2.3.5 Summary

After applying all the various quality control steps mentioned in section 2.2, the final dataset contains 3,322 cases with schizophrenia and 3,587 healthy controls. 246,204 SNPs remained for the analysis.

2.4 CLOZUK Dataset

2.4.1 Introduction

The CLOZUK sample, used in Chapters 4 and 5 of the thesis, was part of a larger GWA study of schizophrenia performed by the Psychiatric Genomics Consortium (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014).

2.4.2 Sample details

The CLOZUK sample (Hamshere et al. 2013) consists of cases ascertained through facilitation with Novartis, the manufacturer of a proprietary form of clozapine (Clozaril). Cases consisted of individuals with a clinical diagnosis of treatment-resistant schizophrenia. Controls were drawn from Wellcome Trust Case Control Consortium 2 (WTCCC2, Wellcome Trust Case Control Consortium 2007). No screening for psychiatric illness was performed.

2.4.3 Genotyping

DNA was extracted from blood. The cases samples were genotyped at the Broad Institute of Harvard and MIT in Boston, USA. The genotyping, described in (Hamshere et al. 2013), was done on two different chips: Illumina HumanOmniExpres-12v1 and OmniExpresssExome-8 that

will be denominated as Combo and Omni. The controls from WTCCC2 were genotyped at the Wellcome Sanger Institute.

After imputation, 7,763 individuals and 10,670,661 SNPs were genotyped on the Combo chip and 4,233 individuals and 10,663,800 SNPs on the Omni chip.

2.4.4 Quality controls

2.4.4.1 Phenotypes

Individuals with missing phenotypes were excluded from both datasets. Originally the Combo chip contained 7,763 individuals including 32 with missing phenotypes. After exclusion, 7,731 individuals remained, corresponding to 4,285 controls and 3,446 cases. The Omni chip contained 4,233 individuals including 111 with missing phenotypes. After exclusion, 4,122 individuals remained, corresponding to 2,014 controls and 2,108 cases.

2.4.4.2 Call Rates

After imputation the Combo chip contained 10,670,661 SNPs and the Omni chip contained 10,663,800 SNPs. Comparison between allowing 1% or 2% of missingness in SNPs call rates was performed (Table 2.3).

Cut-off threshold	Combo Chip	Omni Chip
1%	6,454,563 SNPs (60% of total SNPs)	5,478,552 SNPs (51% of total SNPs)
2%	4,772,338 SNPs (45% of total SNPs)	3,969,836 SNPs (37% of total SNPs)

Table 2.3: Number of SNPs to remove in each chip using two different cut-off thresholds (1% and 2%) for the SNPs call rates.

A cut-off threshold of 2% was applied for the SNP coverage in each chip. After applying this threshold, all individuals call rates were above 0.98 in both chips indicating high genotype reliability (Figure 2.3).

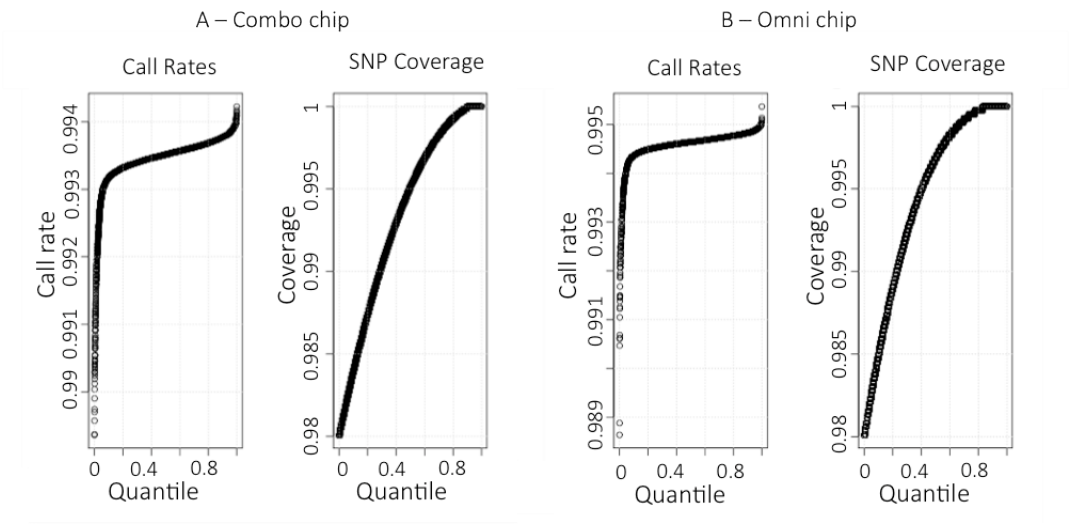


Figure 2.3: Call rate and SNP coverage for the Combo chip (Figure A) the Omni chip (Figure B) after applying the cut-off threshold of 2% for the SNP coverage. For both chips, the call rate and the SNP coverage are above 98%

After applying the 2% cut-off the numbers of SNPs contained in each chip were as follows: 5,898,323 variants in the Combo chip and 6,693,964 markers in the Omni chip.

2.4.4.3 Minor Allele Frequency

Two different MAF thresholds were used to compare the number of SNPs that could be removed: 1% and 5 % (Table 2.4).

MAF threshold	Combo chip	Omni chip
1%	1,030,426 SNPs	1,061,764 SNPs
5%	2,336,133 SNPs	2,491,123 SNPs

Table 2.4: Number of SNPs to remove using different MAF threshold in each chip

A threshold of 1% was chosen, as it was a good balance between the number of SNPs to exclude and the exclusion of rare variants. After applying the control on minor allele frequency, the Combo chip retained 4,867,897 SNPs and the Omni chip 5,632,300 SNPs.

2.4.4.4 Heterozygosity

After checking heterozygosity plots, a few outliers were removed (9 individuals). The final plots showed a normal distribution for both the mean heterozygosity and the inbreeding coefficient without any outliers (Figure 2.4).

After removal of outliers, 7,719 individuals were retained on the Combo chip and 4,107 individuals on the Omni chip.

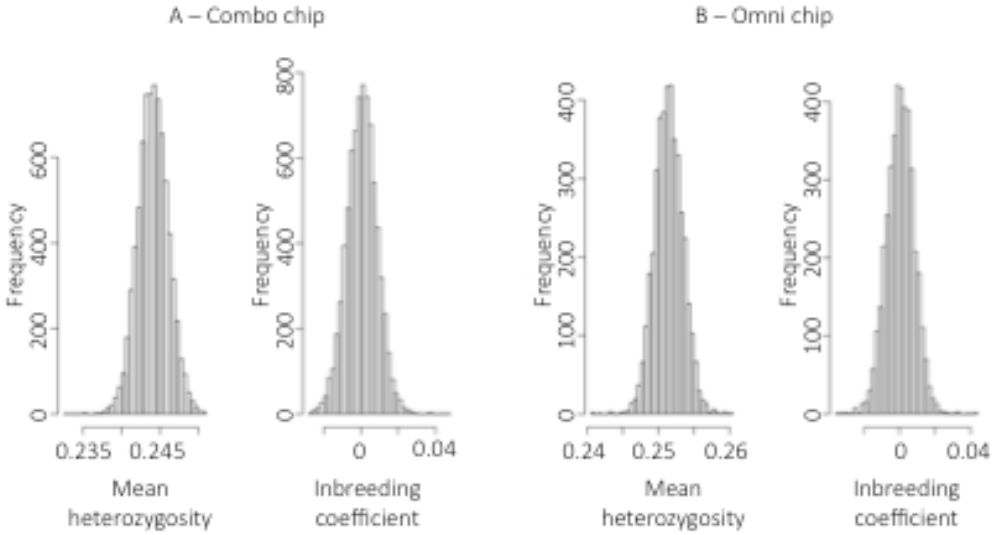


Figure 2.4: Histograms of the mean heterozygosity and of the inbreeding coefficient for the Combo chip (A) and the Omni chip (B). The outliers were previously excluded as those samples could have been contaminated or do not belong to the same population. After the removal of every unusual sample all the histograms shows a normal distribution in each chip.

2.4.4.5 Cryptic relatedness

12 pairs of individuals from the Combo chip were identified as being related, as were 7 pairs of individuals from the Omni chip. One individual from each pair was chosen at random and removed from the data.

2.4.4.6 Hardy Weinberg

All variants passed the Hardy Weinberg thresholds, perhaps indicating that there were removed prior to this analysis.

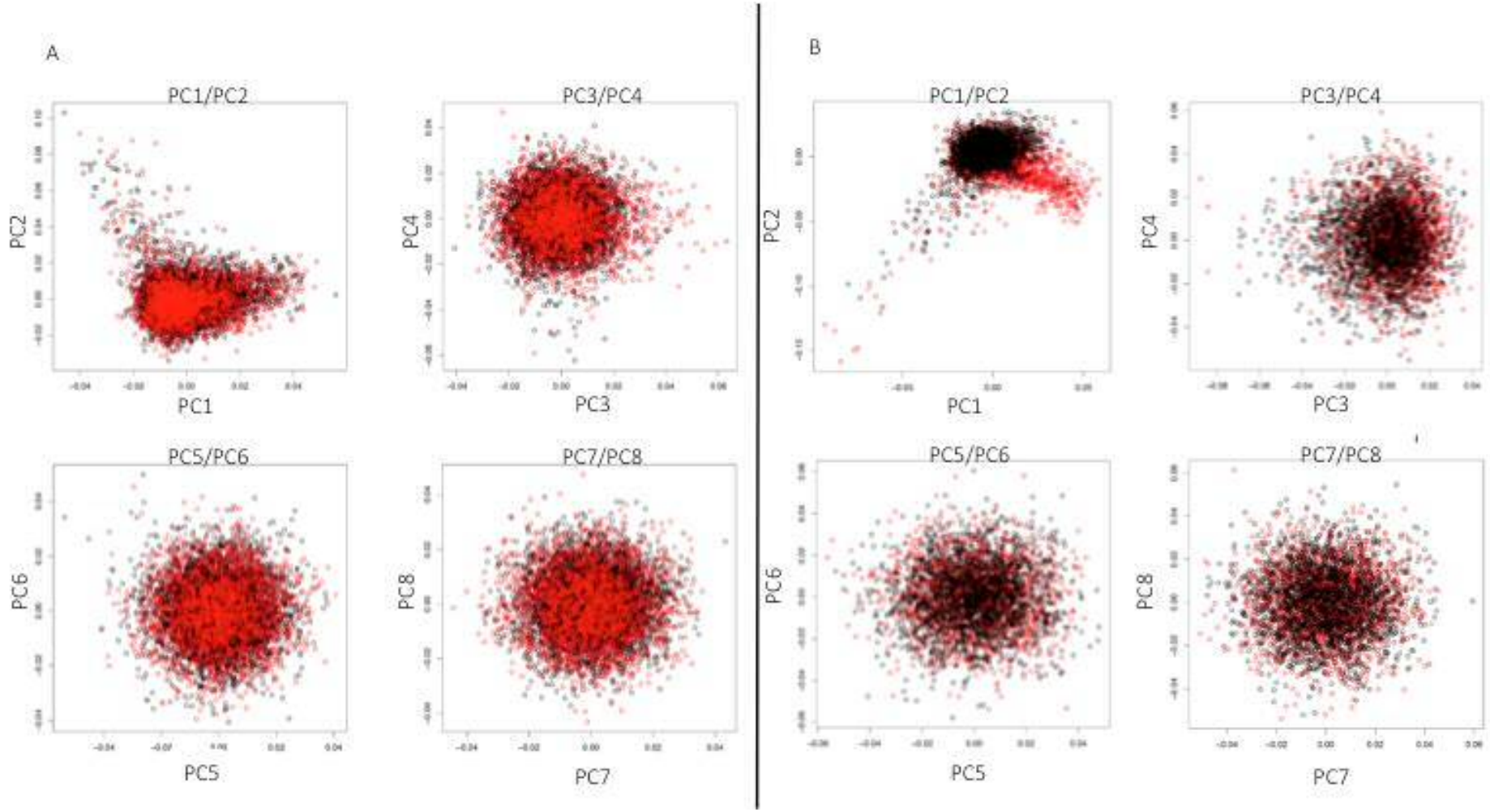


Figure 2.5: PC Plots for the CLOZUK dataset (A: Combo chip, B: Omni chip)

2.4.4.7 Principal Component Analysis

Plots of principal components (PC) against each other were drawn and analysed for each chip (Figure 2.5). The principal component analysis showed some discrepancies. For the Omni chip, cases and controls respectively represented in black and red can be separated visually from each other using the first two PCs (Figure 2.5 Right panel B PC1/PC2). For the Combo chip, a tail of outliers is clearly visible (Figure 2.5 Left Panel B PC1/PC2), indicating non-European ancestry. Following clustering (Section 2.2.6), outliers were then excluded from the analysis: 262 in the Combo chip and 472 in the Omni chip.

2.4.5 Summary

The CLOZUK dataset contains data genotyped using two different chips: Omni and Combo. After QC, the Combo chip retained 7,445 individuals: 3,283 cases and 4,276 controls. The Omni chip retained 3,628 individuals: 1,917 cases and 1,711 controls.

Regarding the number of SNPs in each chip, the Combo chip retained 4,867,897 SNPs and the Omni chip 5,632,200 SNPs that will be used in the analysis of Chapter 3 and 5.

Chapter 3 - Interaction analysis in GWAS dataset with covariates

3.1 Introduction

3.1.1 Background

In complex diseases, gene-gene interaction refers to a phenotype that cannot be explained by adding the independent effects of two loci but can modified the prediction by their joint effect (Carlborg and Haley 2004). This effect can be modelled by adding an interaction term in the analysis.

The most frequently used model to calculate statistical interaction is based on a linear/logistic regression depending on the phenotype distribution. The regression model describes the relationship between one outcome variable and one or several predictor variables, including the interaction term (Cordell 2009). For a case-control study, the interaction model describes the effect of those predictor variables on the log odds of the disease. The method consists of testing the interaction term only and was described in detail in Chapter 1 based on Cordell (2002).

The motivations behind the inclusion of covariates into a statistical model are the improvement of the precision of the model's estimates in the presence of potential confounders and the maximization of the statistical power (Sham and Purcell 2014). By including confounding covariates into an analysis, false positives and false negative associations can be reduced, increasing the power of interaction testing (Sham and Purcell 2014).

When including covariates into an interaction analysis, the most widely used method consists of simply adding the covariate terms into the logistic regression model. However it was demonstrated by Yzerbyt et al. (2004) that in specific situations, to be detailed later,

this could result in the presence of bias. This paper suggested adding not only the covariate terms into the equation but also adding interaction terms between each covariate and each predictor variable to control for covariate's effects.

3.1.2 Aims of the chapter

In this chapter three different models for including sub-population covariates into an interaction analysis were used on the same GWAS dataset. The first model (widely used) consists of adding the covariate terms into the model. The second one follows the recommendation by Yzerbyt et al. (2004): it adds into the model the covariates terms as above and additional interaction terms between each covariates and each predictor variables. In the last model, interactions are calculated independently in each sub-population and a meta-analysis is used to combine the results.

The aim of this chapter is to compare the three methods and assess the strengths and limitations for each of them.

3.2 Material and methods

3.2.1 Data

For this analysis, the International Schizophrenia Consortium (ISC) dataset was used (International Schizophrenia Consortium 2008) as it has a natural population-based structure. It consists of a total of 3,322 individuals with schizophrenia and 3,587 healthy subjects collected from 8 different populations (Chapter 2, Table 2.1). The data and the quality control steps are described in Chapter 2.

3.2.2 SNPs selection and pruning

SNPs outside genes were removed and chromosomes X and Y were excluded from the analysis as well as insertions and deletions, retaining 95,124 genetic variants. Insertions and deletions (indels) were excluded from the analysis due to their rarity as well as the

potential inaccuracy of the calling for indels (Hasan et al. 2015). In order to perform SNP-SNP interactions on X and Y chromosomes, a separate analysis of female and male subsamples would have been needed. This would have resulted in analysing samples of small sizes resulting in an important loss of power for detection of potential interactions as well as an increase number of tests.

When undertaking a pair-wise interaction analysis there is an interest in reducing the number of variants studied (see Chapter 1, section 1.2.3.2). As comprehensive pair-wise SNP-SNP interaction analysis is time consuming and computationally intensive when using data from genome wide association studies. For this reason, the number of SNPs involved in the study needed to be reduced (Moskvina, et al. 2011).

Linkage disequilibrium (LD) based pruning was used to reduce the number of SNPs by using the –clump option in PLINK (Purcell et al. 2007). The following parameters were chosen: a window of 2,000 Mb, the significance threshold p1 of 0.01 and r^2 of 0.1. LD-based pruning also has the advantage of avoiding SNP collinearity. Collinearity happens when a predictor variable in a multiple regression model can be estimated from the others variables with a high level of accuracy. As the regression model takes into account this uncertainty it does not affect the predictive power of the model. However it does increase the standard error and affects calculations regarding each predictor. By using LD-based pruning, the probability of having linear-dependent variables is decreased.

Furthermore, as suggested by Price et al. (2008), the major histocompatibility complex (MHC) region, known as a high-LD region, was excluded prior to clumping.

3.2.3 Testing for interactions

As explained in detail in Chapter 1, statistical interactions are based on a linear/logistic regression model and estimate an outcome variable as a function of predictor variables (Cordell 2009). Here the analysis is based on a case-control binary phenotype, therefore the outcome variable will be the log odds of the disease.

Let and represent two independent variables with three levels (0/1/2) corresponding to the possible genotypes of the markers (aa/Aa/AA).

is the dependent variable that represents the disease status of the individuals: 1 for the presence of the disease in the individual (case) and 0 for the absence of the disease in the individual (control). is a function of the previous variables , and , the final term representing the interaction.

Let C_i be the covariate term for covariate variable *i*. and K be the total number of covariates (for the ISC dataset this corresponds to the number of sub-populations).

Let be the standardized beta coefficients: is the Y-intercept.

Let be the random error component.

When testing for statistical interactions between two independent variables, the effect size of the interaction term between those two variables is tested as shown by the following equation:

(1)

In order to assess the significance of the interaction term, the following hypothesis are tested:

using a log-likelihood ratio test and corresponding to the difference in deviances of the fits of the two models.

However the effect of covariates needs to be accounted for. It has been shown in previous studies that when performing an interaction analysis (Arya et al. 2009), the inclusion of covariates into the model can change the overall results.

The widely accepted method simply adds the covariates into the regression model. However it was suggested by Yzerbyt et al. (2004) that just entering the covariates into the

model did not control properly for the effect those covariates might have on each marker, as it may not only be the main effects that can be influenced by those variables. A demonstration of this bias was described in Yzerbyt et al. (2004). Details on specific situations, which are outlined as in Yzerbyt et al. (2004), were also shown in Keller (2014) and illustrated well the problem for genetic data analysis. One of the main examples cited (Kaufman et al. 2004) investigated if a repeat polymorphism (short/long) at the serotonin-transporter linked region was dependent on childhood maltreatment and social support. In that analysis, ethnicity (African-Americans, biracial and non-African Americans) and sex were taken into account. The main finding was that the interaction between the polymorphic region and childhood maltreatment was significant: the individuals with the s/s allele were found to have significantly higher depression scores. It was also found that African-Americans have a high frequency of the long repeat than non-African Americans. However as the interaction between environment and ethnicity and between the polymorphism and ethnicity were not included, it is not possible to completely eliminate alternative explanations for the finding. For example the detected interaction could have been influenced by ethnicity if due to difference in cultural norms one group does not report depression. In a similar way, this example can be translated to SNP-SNP interactions and can raise similar issues when integrating covariates into interactions studies.

3.2.3.1 Method 1: Inclusion of covariates in the regression analysis

As outlined above, the typically used method includes information about population as covariates. It simply involves adding the covariates into the regression model as additional terms described in Equation 1.

(2)

3.2.3.2 Method 2: Inclusion of covariates and effects between the covariates and the markers in the interaction analysis

This second method controls for all the effects between the covariates and all the markers. Like in Method 1 it adds the covariates into the model but it also adds extra interactions terms between all the covariates and all the markers as suggested by (Yzerbyt et al. 2004).

(3)

3.2.3.3 Method 3: Interactions by means of meta-analysis of sub-populations

This last method analyses separately the 8 populations and takes into account the directionality of effect size with a meta-analysis as shown below:

(4)

3.2.3.4 Implementation of the interaction analysis

The first two methods were run using an R script. The last one used two different software packages: Plink 1.9 (Purcell et al. 2007; Chang et al. 2015) and METAL (Willer et al. 2010). Plink 1.9, a command line tool designed for large-scale genetic analyses, was used to calculate the interactions in each of the 8 samples. METAL, a computationally efficient software for genome-wide scans, was used to combine the results. All the analyses were run on the Cardiff University High Performance Cluster Raven.

3.2.4 Notations

In this chapter, Method 1 will be used to describe the approach adding the covariates as in equation 2. Method 2 will describe the approach where the covariates terms and the interaction terms between each marker and the covariates are included in the regression model as in equation 3. Method 3 will refer to the method using meta-analysis (equation 4).

3.3 Results

Following the quality control steps described in Chapter 2 and the LD-pruning described in section 3.2.2, 5,135 SNPs within 3,105 genes remained. All pair-wise interactions between those variants (13,181,545 interactions) were calculated using the three methods specified above.

In all three methods, there was no interaction effect, which passed the multiple testing correction threshold corresponding to a corrected p-value of 0.05 (Bonferroni threshold $p=3.79e^{-9}$) as shown in Table 3.1.

	Method 1	Method 2	Method 3
Number of interactions with p-value below Bonferroni threshold ($p=3.79e^{-9}$)	0	0	0
Number of interactions with p-value $p \leq 1e^{-5}$	129	138	101
Number of interactions with p-value $p \leq 0.001$	13,586	14,446	10,921
Lowest observed p-values	$7.56e^{-8}$	$4.62e^{-8}$	$8.79e^{-8}$

Table 3.1: Number of interactions with a p-value below different thresholds in each method. For the three methods, no interaction passed the Bonferroni threshold ($p=3.79e^{-9}$). Method 2 detects more interactions below the thresholds 10^{-5} and 0.001. Method 1 has similar results to Method 2, whereas Method 3 detects the least number of interactions below a give threshold.

3.3.1 Computational performance

When comparing the computational time for the three methods (Table 3.2), Method 3 was the fastest with a running time of approximately 2 hours. It used two software packages (PLINK and METAL) that are designed and optimised for genetic data processing and therefore can run these analyses efficiently. The analyses with Method 1 and Method 2 took significantly longer (5 days for Method 1 and 1 week for Method 2) for the calculations. Both of these methods were implemented in R, which in general is not the most efficient statistical software and the running time for these two methods is not directly comparable with the running-time of Method 3.

	Method 1	Method 2	Method 3
Computational methodology	R script	R script	PLINK and Metal
Running time	5 days	1 week	2 hours

Table 3.2: Running time of each method for the interaction analysis.

3.3.2 Distribution of results for the three methods

Visually comparing the distribution of the calculated interactions p-values between the three methods, Q-Q plots from Method 1 and Method 2 look quite similar whereas Method 3 produced slightly less significant results than we would expect to have by chance at the higher significant thresholds (Figure 3.1).

3.3.3 Correlation between the results from the three methods

Correlation coefficients were calculated between the three methods, based on all interaction p-values for each SNP (Table 3.3).

		M1 / M2	M1 / M3	M2 / M3
Pearson	R	0.964	0.793	0.815
	p-value	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$
Pearson ($-\log_{10}P$)	R	0.981	0.847	0.861
	p-value	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$
Spearman	R	0.965	0.793	0.815
	p-value	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$

Table 3.3: Correlation between the three methods. The correlation coefficients were calculated with three methods: Pearson, Pearson using a log transformation of p-values and Spearman.

The correlation between Methods 1 and 2 is very high as can also be seen from Table 3.3 and Figure 3.2 (left panel). When comparing Method 3 with either Method 1 or 2, the degree of correlation is reduced, although still high as is shown in Table 3.3 and Figure 3.2 (middle and right panel). Despite the obvious correlation of p-values with extremely low p-values, Figure 3.2 also indicates that the interactions with low p-value are correlating well.

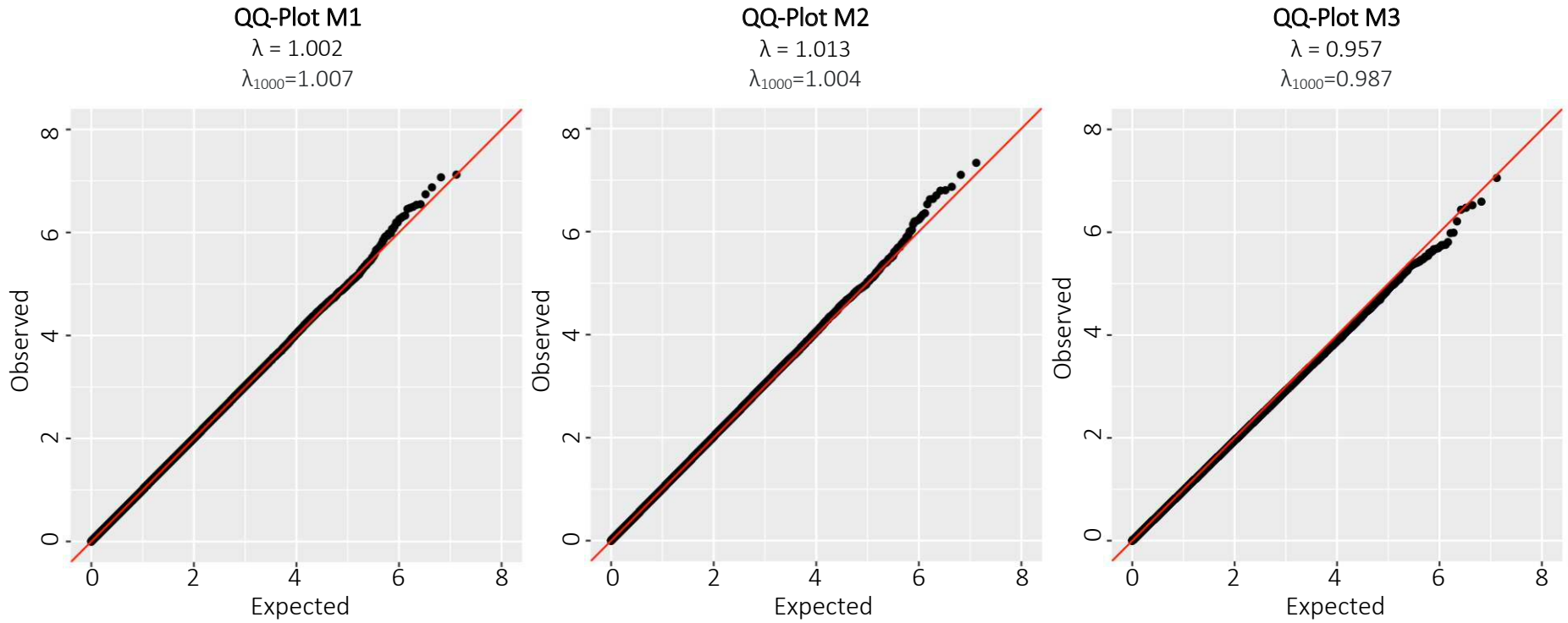


*Figure 3.1: Quantile-quantile (QQ) plots for all pair-wise interactions analysis calculated for each method. The QQ plots show a representation of the deviation of the observed p-values (expressed as the negative log of the p value) from the null hypothesis: the observed P values for each interaction (y axis: Observed) are sorted and plotted against expected values from a theoretical χ2-distribution (x axis: Expected). λ is the genomic inflation factor and λ_{1000} the estimated genomic control (value that would be expected in a study of 1,000 cases and 1,000 controls). The left panel represents Meth*od 1, the middle panel Method 2 and the right panel is Method 3. *The observed p-values in Method 3 are less significant than expected.*

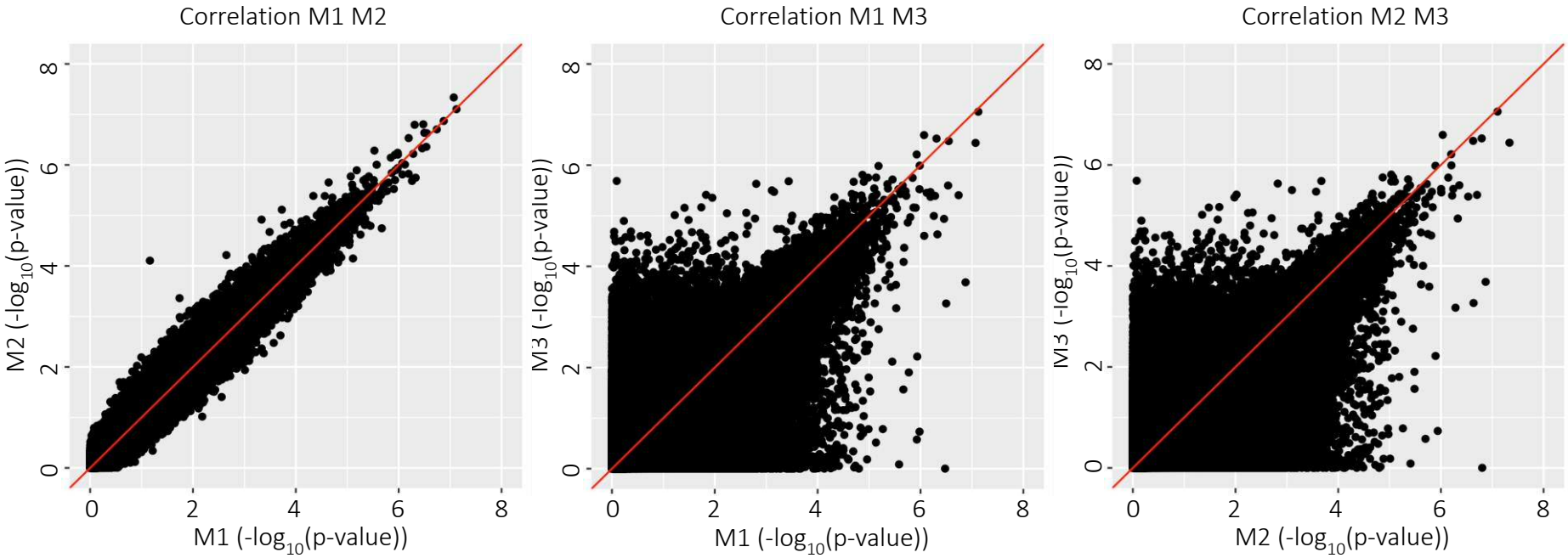


Figure 3.2: Scatter plot of interactions p-values using logarithmic scale for Method 1 against Method 2 (left panel), Method 2 against Method 3 (middle panel) and Method 1 against Method 3 (right panel).

As shown by (Bland and Altman 1986) a strong correlation between two methods does not automatically imply agreement between them: correlation coefficients only measure the strength of the relationship between two variables. For example, a change of scale will not affect the correlation coefficients but will change the agreement.

Bland Altman plots were used to check the agreement between the results of the three methods. This is a simple way of graphically comparing the agreement between two methods. It plots the difference between the results calculated by the two methods as a function of the average of those two results (Bland and Altman 1986). If the points on the Bland Altman plot are randomly distributed, above and below zero, then it is possible to conclude that there is no consistent bias of one method versus the other. But for example if one method always gives a different results e.g. with all points above or below the zero line, it would be possible to conclude that one method over or underestimates the results. However there would be no certainty on which one is responsible for the over/underestimation. The three plots in Figure 3.3 shows the Bland Altman plots respectively for Method 1 against Method 2 (left panel), Method 1 against Method 3 (middle panel) and Method 2 against Method 3 (right panel). The results are above and below zero in all the cases, which tends to indicate that there are no visible biases. However a trend is visible when looking at Method 3 against the other two methods (Figure 3.3 middle and right panel), indicating that the difference between the methods is reduced when looking at high values (very low p-values as the logarithmic scale is used). This would confirm that the overall performance of each method is comparable, with closest agreement for the most significant interactions.

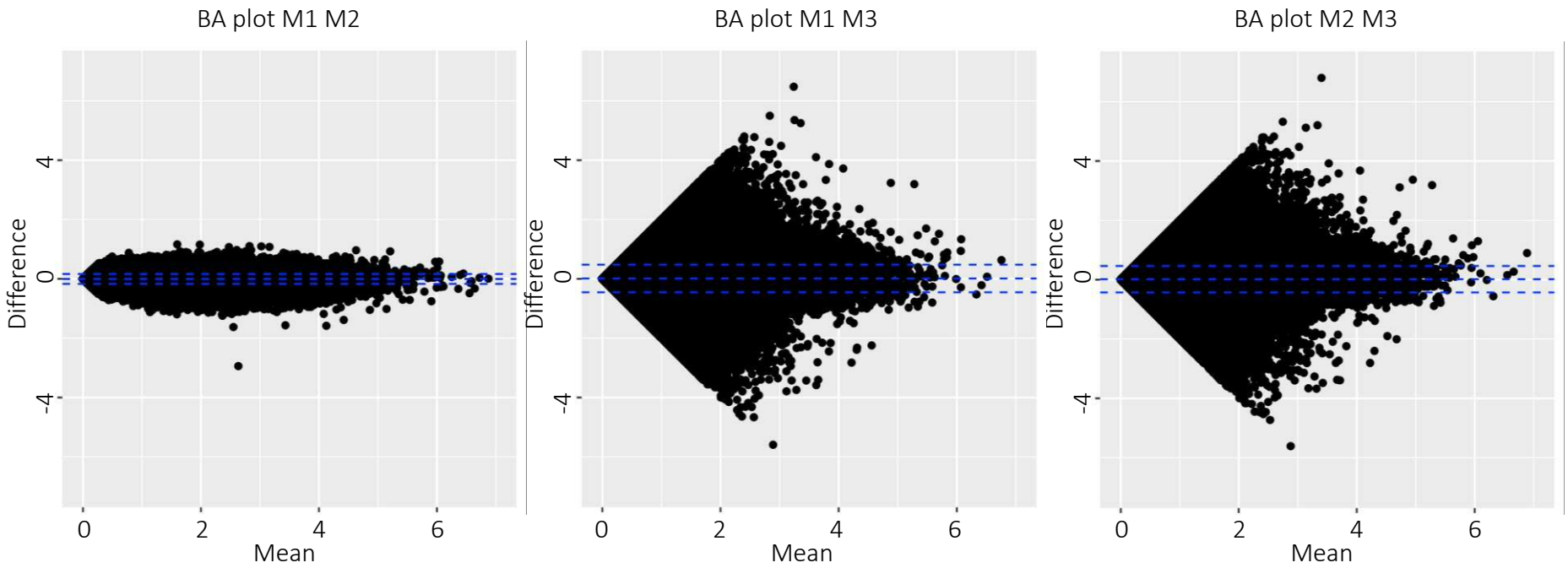


Figure 3.3: Bland Altman plots using logarithmic scale to check agreements between the methods: Method 1and 2 (left panel) Method 1 and 3 (middle panel) and Method 2 and 3 (right panel). Difference refers to the difference between the results calculated for the two methods. Mean refers to the average of those two results. The blue doted lines show the 95 percent confidence intervals.

3.3.4 Interaction results with lower p-values

To further investigate the lower end of the p-value distribution interactions with (both) p≤0.001 in each pair of methods were compared (Table 3.4).

	Methods 1 and 2	Methods 1 and 3	Methods 2 and 3
Number of interactions with p≤0.001 in both methods	11,465	6,733	7,374

Table 3.4: Number of interactions with a p-value p≤0.001 for each pair of methods.

As can be seen from the scatter plots (Figure 3.4), all the methods produced similar result and showed positive correlation. The scatter plot of Method 1 and Method 2 (Figure 3.4, left panel) showed the strongest positive correlation.

The results of the correlation analysis between each pair of methods reflect what is observed in the scatter plots (Table 3.5, Results). The correlation coefficients are obviously lower than the one previously observed when comparing the whole distributions of interactions (Table 3.5, previous results). However when comparing the lower end of the distribution, it shows a very strong positive correlation between Method 1 and Method 2 (Table 3.5, Results). The correlation between Method 3 and Methods 1 and 2 is lower but still shows positive correlation (Table 3.5, Results).

		Previous Results			Results		
		M1 / M2	M1 / M3	M2 / M3	M1 / M2	M1 / M3	M2 / M3
Pearson	R	0.964	0.793	0.815	0.776	0.492	0.546
	p-value	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$
Pearson ($-\log_{10}P$)	R	0.981	0.847	0.861	0.901	0.675	0.721
	p-value	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$
Spearman	R	0.965	0.793	0.815	0.805	0.545	0.595
	p-value	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$	$p<2.2e^{-16}$

Table 3.5: Correlations between the three methods after selecting interactions with a p-value $p \leq 0.001$ in each method*. The correlation coefficients were calculated with three methods: Pearson, Pearson using a log transformation of p-values and Spearman.*

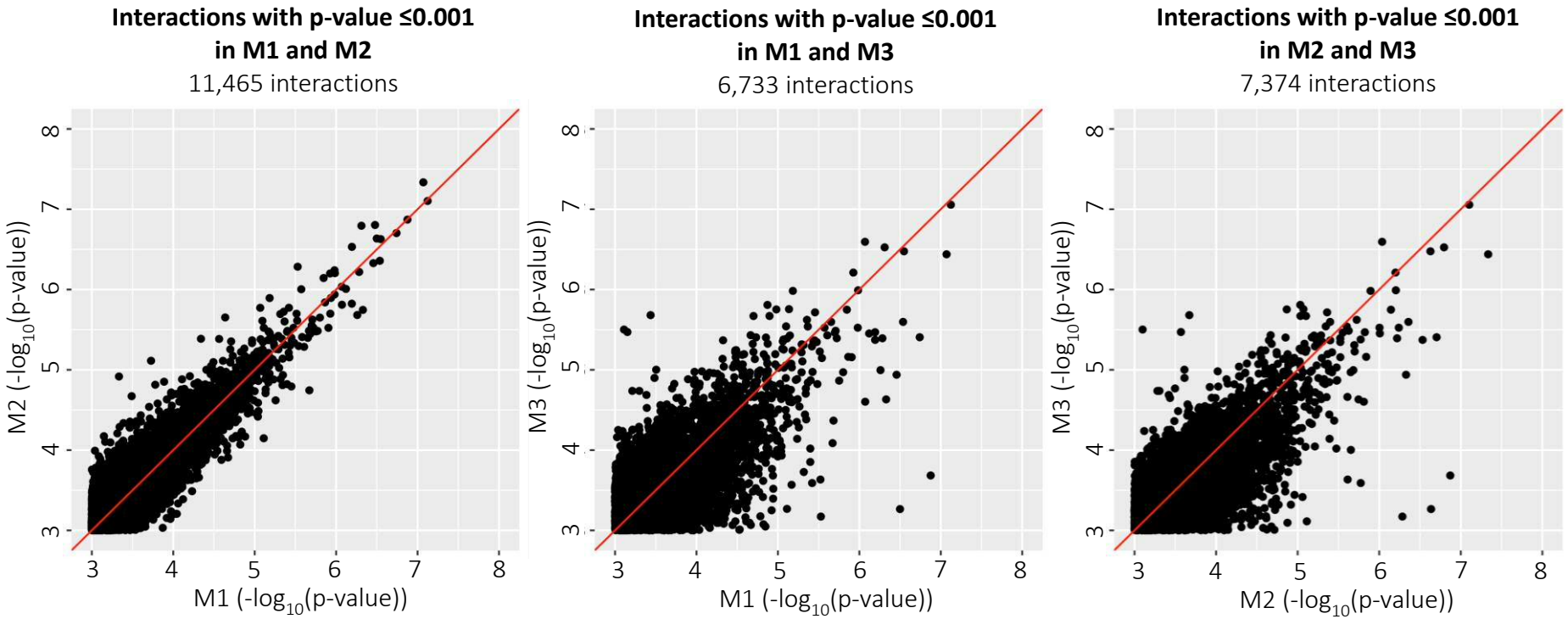


Figure 3.4: *Scatter plots of interactions with p-values (log scale) below 0.001 selected in each pair of methods. Method 1 against Method 2 (left panel), Method 1 against Method 3(middle panel), Method 2 against Method 3 (right panel). All scatter plots show positive correlation.*

3.3.5 Investigation of differences between methods

This section is focused on interactions that are identified as below the threshold (p≤0.001) in at least one method and more specifically on interactions passing that threshold in one method but not in another method. In addition, it explores one of the potential explanations for the differences observed. Other approaches that were not explored due to lack of time will be detailed in the discussion section.

In each pair of methods compared, interactions were selected if in at least one of the two methods the p-value was below the threshold p≤0.001 (Table 3.6).

	Methods 1 or 2	Methods 1 or 3	Methods 2 or 3
Number of interactions with p≤0.001 in at least one method	16,563	17,770	17,989

Table 3.6: Number of interactions with a p-value p≤0.001 in one of each pair of methods.

As previously, scatter plots were drawn to observe the distribution of p-values in each method (Figure 3.5). Compared to previous section (Figure 3.4), interactions with a p-value below the threshold in both methods were removed. Comparison between Method 1 and Method 2 (Figure 3.5, left panel) produced results very similar to those previously observed, with no major differences between the two methods. Comparing Method 3 with the other two methods, the scatter plots (Figure 3.5, middle and right panels) looks different. In some cases, Method 3 is able to identify interactions with reasonably low p-values and the other method does not. In other cases, the opposite is observed.

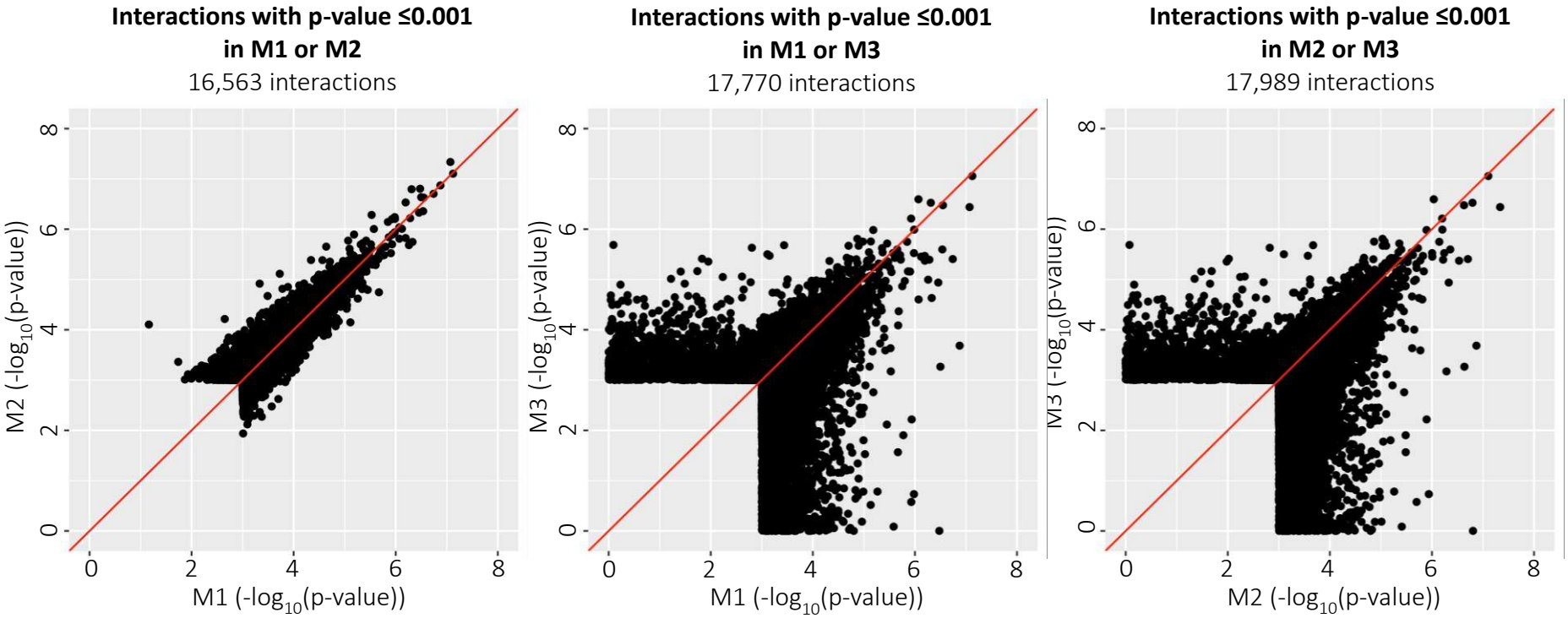


Figure 3.5: Scatter plots of interactions with p-values (log scale) below 0.001 selected in each method. Method 1 against Method 2 (left panel), Method 1 against Method 3(middle panel), Method 2 against Method 3 (right panel). The first scatter plot of Method 1 and Method 2 shows positive correlation. The other scatter plots Method 1 and Method 3 (middle panel) Method 2 and Method 3 (right panel) are more different but interactions with very low p-values have closer results.

3.3.5.1 Role of the direction of effects

The software METAL, used in Method 3, calculated the direction of the effect for each study. Since only interaction terms were meta-analysed, the direction of the main effects was ignored. In the present analysis, as there are 8 different studies that are combined, the number of identical direction of effects varies between 4 and 8. Scatter plots (Figure 3.6) were drawn to investigate whether or not the direction of the interaction effects could explain the observed results (Figure 3.5). This showed that when the directions of effect are identical in four or five out of 8 studies, Method 3's results are different from Methods 1 and 2 (Figure 3.6 A, B, D and E). However when the directions of effect are uniform across the 8 studies, Method 3 produces results that more closely match those of the other methods (Figure 3.6, C and F). This indicates that the direction of the interaction effect might be playing a role in Method 3's results. Perhaps taking into account the SNPs main effects could enhance this method.

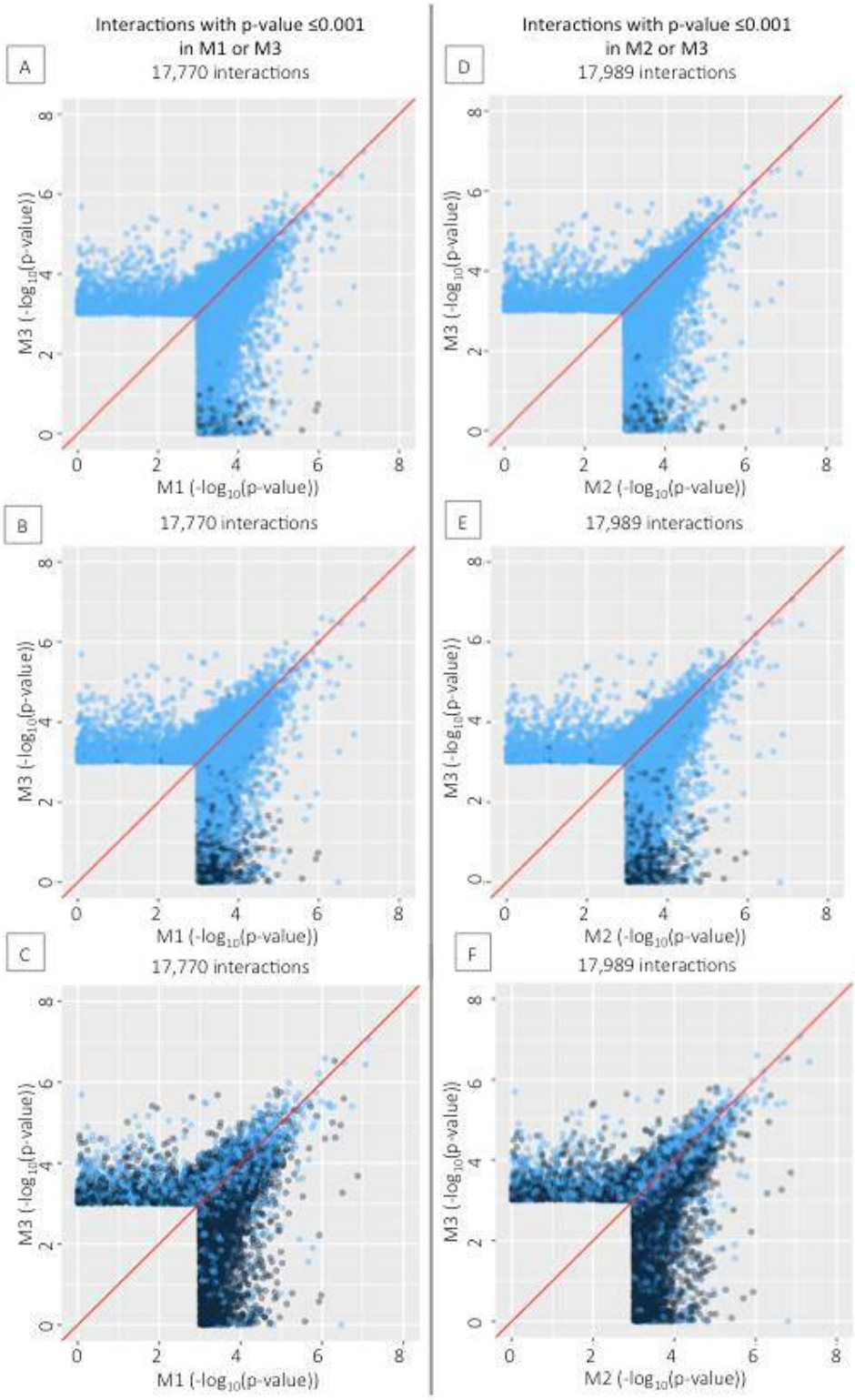


Figure 3.6: Scatter plots of interactions with p-values below 0.001 selected in at least one method. Method 1 against Method 3 (left panel), Method 2 against Method 3 (right panel). In A and D plots, the black dots show interactions where the direction of effects is different in 4 samples out of 8. In B and E plots, the black dots show interactions where the direction of effect is the same in four or five samples out of 8. In C and F plots the blue dots show interactions where the direction of effect is identical in all the samples. All graphs show that the direction of interaction effects might play a role in explaining Method 3's results.

3.4 Discussion

3.4.1 Multiple testing

No interactions survived Bonferroni correction for multiple testing in any analysis performed using the three methods. As the number of calculated interactions is very high, the penalty for multiple testing is severe (Bonferroni threshold $p=3.70e^{-9}$).

One possible explanation for the absence of significant interactions could be the sample size: the ISC dataset contains 3,322 cases and 3,587 controls. A bigger sample would increase the statistical power as shown by (Gauderman 2002) and as will be detailed in the next Chapter.

3.4.2 Correlation between the results of the three methods

Methods 1 and 2 were found to have extremely similar result: the correlation coefficient between the results of the two methods is close to the perfect linear association (correlation almost equal to 1) indicating a strong positive relationship between both approaches (Table 3.3). Method 1 only adds the covariates into the primary equation (see Equation 2). Method 2 differs from Method 1 by the fact it deals with the bias explained by Yzerbyt et al. (2004): it takes into account the possible effect between the covariates and each marker by adding interactions terms between covariates and markers into the equation (see Equation 3). As the presented results between those two methods are comparable, it seems that the bias mentioned above does not play a large role, at least in the analysed dataset. However, a large-scale simulation study would be needed in order to definitely prove this assertion.

The correlation between Method 3 and the other methods (Method 1 and Method 2) appears to be quite high as well (Table 3.3). It also showed that interactions with low p-value (i.e. greater significance) seem to be moderately well correlated (Figure 3.2).

Scatter plots and Bland Altman agreements plots between the results of the three methods showed very strong relationship between Method 1 and 2, and less strong but good association between Method 3 and the other two methods. In addition, the Bland Altman agreements plots do not show evidence of bias of one method versus the other as the observed values tend not to be consistently above or below zero. This confirmed the results observed by looking only at the correlation coefficients (Table 3.3).

3.4.3 Interactions with lower p-values

As the interest in this analysis was focused on significant results, the lower tail of the distribution of p-values was examined in more detail.

When selecting interactions with a p-value below the threshold $p \leq 0.001$ in each method, it was observed that Method 1 and 2 again produced similar results: the correlation coefficients were high (Table 3.5) and the scatterplot shows a strong positive relationship between the two approaches (Figure 3.4, left panel). Some differences were visible between Method 3 and the other two methods (Figure 3.4, middle and right panel): Method 3, which uses the meta-analysis, did not identify as many interactions with a p-value $p \leq 0.001$ as the other methods (Table 3.4). However, the calculated correlation coefficient between Method 3 and Methods 1 and 2 showed positive correlation (Table 3.5).

3.4.4 Differences observed at the lower end of the distribution

When investigating interactions with a p-value below the threshold $p \leq 0.001$ in either method, two main observations were made. First, there is no major difference between Methods 1 and 2. Secondly, when comparing Method 3 with the other two methods, differences were observed. In some cases, Method 3 was able to identify interactions with reasonably low p-values while the other method (Method 1 or Method 2) was not, and vice versa.

Several explanations can be given for this observation. Firstly, differences in the direction of interaction effect between populations could be a factor influencing the model used by Method 3. When the direction of the interaction effect was investigated, it was found that when the direction of effect was identical in only four or five out of eight studies, Method 3's results tended to differ most from those of Methods 1 and 2.

A second hypothesis is that this observation is due to an over-estimation of the results by Method 3. Method 3 uses METAL for the meta-analysis but only meta-analyse the interaction term, ignoring the two SNP main-effects. In contrast, Method 1 and 2 take into account the main effects. It is possible that not including the main effects results into an over-estimation of the interaction effect size in Method 3, as it may lead to the interaction term capturing some of the main SNP effects. To investigate this, a multivariate approach taking into account the covariance between the SNPs main effects and the interaction effects could be used. Further work would include the application of a different meta-analysis method, such as the one described in Van Houwelingen et al. (2002), to account for this and to see if this ameliorates the results.

3.4.5 Covariates, power and epistatic model

The power of interaction detection in statistical analysis can be influenced by many factors such as the increase of sample size or the inclusion of covariates (Sham and Purcell 2014). Often the inclusion of appropriate covariates into linear regression models leads to improved power of detection (Sham and Purcell 2014). However there is one interesting exception for case control studies in complex diseases: including covariates into a logistic regression model can result in a loss of power of detection when the disease prevalence is low (Pirinen et al. 2012). Potentially this can be translated to gene-gene interactions and raises the question of the need to adjust for covariates. However if the covariates are well known confounds, such as population structure, there is a need to introduce them to properly control for their effects and thus accepting the loss of power.

Furthermore the choice of scale is crucial. The scale depends on a link function that models the relationship between the phenotype and the predictors (Frånberg et al. 2015): in this thesis the logistic regression was used. The choice of link function has important implications: it can impact on the definition and interpretation of epistasis (Clayton 2012). Indeed, Clayton (2012) argued that the logistic regression model with no statistical interaction between genes is quite strongly epistatic.

Ultimately the choice of scale is problematic as it depends on the data and its underlying biological model, which is unknown (Frånberg et al. 2015). For this reason it is possible to choose a scale that would reveal interaction despite having a biological model that does not present any interaction (Clayton 2012). On the other hand, a true interaction effect would be weakened by an inappropriate choice of scale (Frånberg et al. 2015).

For this reason further work could include the use of different link functions in order to investigate this issue: assuming true interactions in the biological model, the detection of such interactions should not depend on the choice of scale (Frånberg et al. 2015; Knol and VanderWeele 2012).

3.4.6 Summary of the chapter

This chapter compared three different ways to include covariates in an interaction analysis. The comparison presented showed that the overall performance of the three methods is similar. No evidence of bias in one method versus the other was observed as seen through the Bland Altman and scatter plots.

Efficiency and running time is a strong factor in the choice of the method to use. Method 3 is the simplest and most efficient method by far compared to the other two methods. In addition, Method 3 is extremely memory efficient making it the easiest and fastest method to carry out interaction analysis for thousands of SNPs. This supports the use of Method 3 to perform SNP-SNP analyses in Chapter 4 and 5. The possible influence of the direction of

interaction effects in the samples will not be addressed in the following chapters as work exploring that effect was done a posteriori.

Chapter 4 - Interactions in GWAS datasets

4.1 Introduction

4.1.1 Background

Genome-wide association studies (GWAS) have demonstrated that there are hundreds of loci associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). In addition CNVs and rare variants studies have also shown evidence of association with the disease (Rees et al. 2014). However the cumulative effects of these findings only account for a minority of the heritability of schizophrenia (Sullivan et al. 2003): there are indications that some of the unexplained heritability may be attributed to interactions between loci (Zuk et al. 2012).

Evidence for genetic interactions has been reported in many model organisms including for the prediction of complex traits (Bloom et al. 2013; He et al. 2010; Mackay 2013; ÁLvarez-Castro et al. 2012). Many methods have been developed to identify genetic interactions in human GWAS data (Wei et al. 2014): several large-scale interaction studies have now evaluated evidence for interactions genome-wide for several complex disorders (Wei et al. 2012; Moskvina et al. 2011; Prabhu and Pe'er 2012). Among those studies, only a few interactions have reached an appropriate level of significance and have been replicated (Hemani et al. 2014; Chu et al. 2014) .

Genome-wide interaction studies have been quite unexplored in schizophrenia. A few studies have provided evidence supporting a role for gene-gene interactions in schizophrenia (Nicodemus et al. 2010; Burdick et al. 2008). In this chapter the focus is drawn on disease associated pair-wise interactions through exhaustive genome-wide testing.

4.1.2 Aim of the chapter

The main aim of this chapter is to assess SNP-SNP interactions in two non-overlapping GWAS dataset (ISC and CLOZUK). Two interaction analyses will be presented: one using SNPs specific to each dataset and the other one will be restricted to the SNPs common to both dataset.

The second aim of this chapter is to investigate whether genetic information alone could be used to identify sets of genes enriched for interactions.

4.2 Material and Methods

4.2.1 Data

The datasets used were the CLOZUK dataset (Hamshere et al. 2013) and ISC studies (International Schizophrenia Consortium 2008). CLOZUK consists of a total of 5,200 individuals with schizophrenia and 5,987 healthy subjects. ISC consists of 3,322 cases and 3,587 controls. The data, including the quality control steps applied to it, are described in Chapter 2.

4.2.2 Method and hypothesis

The method used to calculate SNP-SNP interactions was described in detail in Chapter 3: Method 3. SNP-SNP interactions are calculated independently in each sub-datasets using PLINK 1.9 (Purcell et al. 2007; Chang et al. 2015) and the meta-analysis tool METAL (Willer et al. 2010) combines the results. Only the interaction terms are meta-analysed: interaction p-values are independent from the main effects.

This chapter investigates if genetic information (in the form of SNP main effects) can help to identify group of genes enriched for 'significant' interactions. Pairwise SNP-SNP interactions were calculated then ranked by the gene wide (main effect) significance of the genes involved: the highest ranking SNP-SNP interactions will be those linked to genes highly associated with schizophrenia and the lowest ranking SNP-SNP interactions will be

those linked to genes least associated with the disease. The distribution of interaction p-values in high- and low-ranking gene pairs was compared to assess the evidence for an enrichment of SNP-SNP interactions within gene pairs most highly associated with the disease.

4.2.3 SNP selection

Firstly, only SNP inside genes were kept. The present work is focusing on interactions between SNPs: insertions and deletions (in-dels) were excluded from the analysis due to their rarity as well as the inaccuracy of their calling (Hasan et al. 2015).

Secondly, the analysis focuses on autosomal chromosomes: all variants on X and Y chromosomes were removed as a separate analysis of male and female would have been needed. In addition, a separate analysis would have led to a smaller sample size and resulted in a loss of power for detection.

For the common-SNP comparison, a supplementary step was added: the selection of SNPs present in both datasets.

4.2.4 Clumping procedure

As detailed in Chapter 1, section 1.2.3, getting enough power to identify significant interactions between correlated variants is unlikely: the inclusion of every variant in the interaction analysis would massively increase the multiple testing burden (under simple Bonferroni type correction) (Moskvina et al. 2011). Moreover it is important to prevent collinearity problems that can arise when SNPs are highly correlated (see Chapter 3, section 3.2.2)

For these reasons linkage-disequilibrium (LD) pruning was used to reduce the number of SNPs involved in the analysis: the most highly associated SNPs were selected and any other SNPs in LD with them were removed.

In addition all the SNPs within the major histocompatibility complex (MHC) region were excluded due to extensive LD in that specific region (Price et al. 2008).

In the independent analysis where SNPs specific to each dataset are used, ISC and CLOZUK were clumped separately. The same parameters were chosen for the window and the R-square threshold (Table 4.1). However due to the high difference of number of variants after QC in each dataset (Chapter 2), different parameters were used for the significance threshold p1 (Table 4.1) to obtain a similar number of SNPs within each dataset.

For the common-SNPs analysis, only the CLOZUK dataset was clumped, as it has a larger number of cases and controls than the ISC dataset. The parameters used were the same as previously indicated (Table 4.1).

Clumping parameters	Independent analysis		Common-SNPs analysis	
	ISC	CLOZUK	ISC	CLOZUK
Window w in kb	2,000	2,000	NA	2,000
R-square threshold r2	0.1	0.1	NA	0.1
Significance threshold p1	0.1	0.01	NA	0.01

Table 4.1: Clumping parameters used in the two analyses: window, R-square threshold and significance threshold. The Independent analysis refers to the interaction analysis performed using SNPs specific to each dataset. The common SNPs analysis refers to the interaction analysis that used SNPs common to both datasets: the clumping was performed only on the CLOZUK dataset. NA: Not applicable.

4.2.5 SNP-SNP interaction analysis

In this chapter, statistical interactions were calculated using a logistic regression model that estimates the disease status of individuals as a function of independent predictor variables corresponding to the genotypes of the considered markers. The method used to calculate SNP-SNP interactions was previously described in details in Chapter 3 section 3.2.3.3. Briefly, the 8 different sub-populations in the ISC dataset as well as the two chips in CLOZUK were analysed separately using PLINK 1.9 (Purcell et al. 2007; Chang et al. 2015). The results were then combined by means of inverse variance meta-analysis for each

dataset using METAL (Willer et al. 2010). The interaction p-values are independent from the main effects as only the interaction terms are meta-analysed.

4.2.6 Ranking of the interactions

Gene wide significance for each gene was calculated using Brown's method (Brown 1975). This approach is derived from Fisher's statistics and allows combining p-values from dependent results. In order to calculate the gene-wide significance for each gene, the method combines the association p-values of every SNP for that gene taking into account the number of SNPs and the degree of LD between them (Moskvina et al. 2011). The resulting p-value reflects the overall degree of association between SNPs in this gene and disease.

SNP-SNP interactions were then ranked using the gene-wide p-values for the corresponding genes. If the interaction involved two SNPs in the same gene, the gene-wide p-value of this gene was used. When the interaction involved two SNPs from two different genes, Fisher's method was used to combine the gene-wide p-values from the two genes as shown in the following equation:

with representing the gene wide p-values for the gene and is the number of genes being combined (here k = 2). As a result, the highest-ranking SNP-SNP interactions are those corresponding to pairs of genes most highly associated with the disease, whereas the lowest ranking SNP-SNP interactions are those linked to genes least associated with the disease.

To check for potential confounding factors, the interactions were also ranked by gene-wide p-values from other psychiatric disorders: Alzheimer disease (Lambert et al. 2013), Parkinson disease (Nalls et al. 2014) and Bipolar disorder(Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011). The Alzheimer disease GWAS (Lambert et al. 2013)

was a large, two-stage meta-analysis of individuals from European ancestry: it used genotyped and imputed data for 17,008 cases and 37,154 controls in stage one and genotyped data in 8,572 cases and 11,312 controls for stage 2 and identified 19 variants associated with the disease. The Parkinson disease GWAS (Nalls et al. 2014) was a meta-analysis of 13,708 cases and 95,282 controls using genotyped and imputed variants: it identified 28 independent loci linked to the disease. The Bipolar Disorder GWAS (Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011) used 11,974 cases and 51,792 (combining discovery and replication set) and identified 34 SNPs associated with the disease, among which 18 replicated.

In addition the interaction results from one dataset were ranked by the gene-wide significance from the other dataset: the CLOZUK interactions were ranked using the ISC gene-wide p-values and vice versa. Finally interactions in both dataset were also ranked by gene-wide p-values calculated from the Psychiatric Genomics Consortium 2 GWAS SNP association statistics (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014).

4.2.7 Testing for enrichment

4.2.7.1 Testing the full interaction distribution

To test the hypothesis that interactions are more likely to occur within genes associated with schizophrenia, a Spearman rank correlation test was performed between the interaction p-values and the corresponding (combined) gene-wide significance p-value.

To further investigate the hypothesis, a linear regression was used as described in the following equation.

Let be the Fisher's combined gene-wide p-values (-log10) as described in section 4.2.6 .

Let be the interaction p-values (-log10).

Let be the number of variants in the genes involved in the interaction.

Let be the length of the genes involved in the interaction.

Let the standardized beta coefficients: is the Y-intercept.

Let be the random error component.

The linear regression model will assess whether SNPs from genes with greatest disease effects (largest values) tend to have more significant interactions (larger) while taking into account the number of SNPs and the length of each gene.

4.2.7.2 Testing with prior selection of interactions

To further compare interactions between associated genes to those between non-associated genes, a one-sided Wilcoxon Mann Whitney test was used on the ranked SNP-SNP interactions: the top N% of SNP-SNP interactions ('top' referring to interactions involving genes most highly associated with schizophrenia) was tested against the bottom N% of SNP-SNP interactions ('bottom' referring to interactions involving genes that are least associated with schizophrenia), with N varying from 1% to 50%. This test compares the distribution of SNP-SNP interaction p-values between SNPs in 'top' genes (highly associated with schizophrenia) to the distribution of SNP-SNP interaction p-values between SNPs in 'bottom' genes (those least associated with disease).

To further quantify the observed effect, ratios were calculated of the number of SNP-SNP interactions with p-value $p < \alpha$ in 'top' genes to the number of interactions with p-value $p < \alpha$ in 'bottom' genes for a range of significance thresholds α. This provided more detail on the level of significance driving the effect. A ratio above one would indicates an excess of SNP-SNP interactions with $P < \alpha$ in the set of 'top' genes.

4.2.8 Network of SNP-SNP interactions

To further assess interactions between associated genes to those between non-associated genes, networks were drawn to visualise and compare links between the two groups of ranked interactions: the top N% of SNP-SNP interactions ('top' referring to interactions involving genes most highly associated with schizophrenia) against the bottom N% of SNP-SNP interactions ('bottom' referring to interactions involving genes that are least associated with schizophrenia), with N varying from 1% to 50%. In both groups, SNP-SNP interactions with a p-value $p<0.01$ were selected.

While many SNP-SNP interactions with a p-value $p < 0.01$ are likely to be false positive, it is possible that a proportion of SNPs could potentially contribute to multiple true interactions and therefore be of interest.

The network comparison was done at both the SNP and the gene level. In figures illustrating these networks, each SNP (or gene) is represented by a node and interactions with $p < 0.01$ between two SNPs (or genes) are represented as a link between the corresponding nodes. To further assess the networks, the Database for Annotation, Visualization and Integrated Discovery (DAVID) (Sherman et al. 2007) was used to investigate shared biological processes (GO terms) between genes (Ashburner et al. 2000).

4.2.9 Functional annotation of significant interactions

Genes involved in interactions with an interaction p-value $p < 10^{-4}$ were further examined (genes sizes were not taken into account) The biological functionalities of those genes were analysed using the DAVID (Sherman et al. 2007). The background list of genes was built with the genes from the interaction that did not pass the selection threshold. For the functional annotation, it was decided to use the summarised version of the Biological processes in the Gene Ontology (Ashburner et al. 2000). A modified Fisher's exact test was used to determine whether genes involved in interactions with an interaction p-value $p < 10^{-4}$ were enriched for GO terms compared to a background list of genes. An enrichment

threshold was chosen for p=0.05: an annotation category is considered of interest at that threshold (Sherman et al. 2007). The analysis was performed on the results for the independent comparison as well as on the ones form the common-SNPs comparison.

4.3 Results

4.3.1 Overall assessment of the interaction analyses

For the independent analysis, each dataset was studied separately. Out of the SNPs present in all 8 subpopulations of the ISC dataset, 2,898 SNPs within 2,080 genes were left after linkage disequilibrium pruning (Table 4.3). All pair-wise interactions were assessed separately for each population using logistic regression analysis and then combined by meta-analysis (Figure 4.1). No interaction survived the multiple test correction (Table 4.3). Similarly after LD pruning, 3,318 SNPs within 2,418 genes remained out of the selected SNPs in the CLOZUK dataset (Table 4.2). All possible pair-wise interactions were calculated separately for each chip and combined as explained previously. Once again, none of the interactions passed a Bonferroni corrected p-value threshold of 0.05 (Table 4.3).

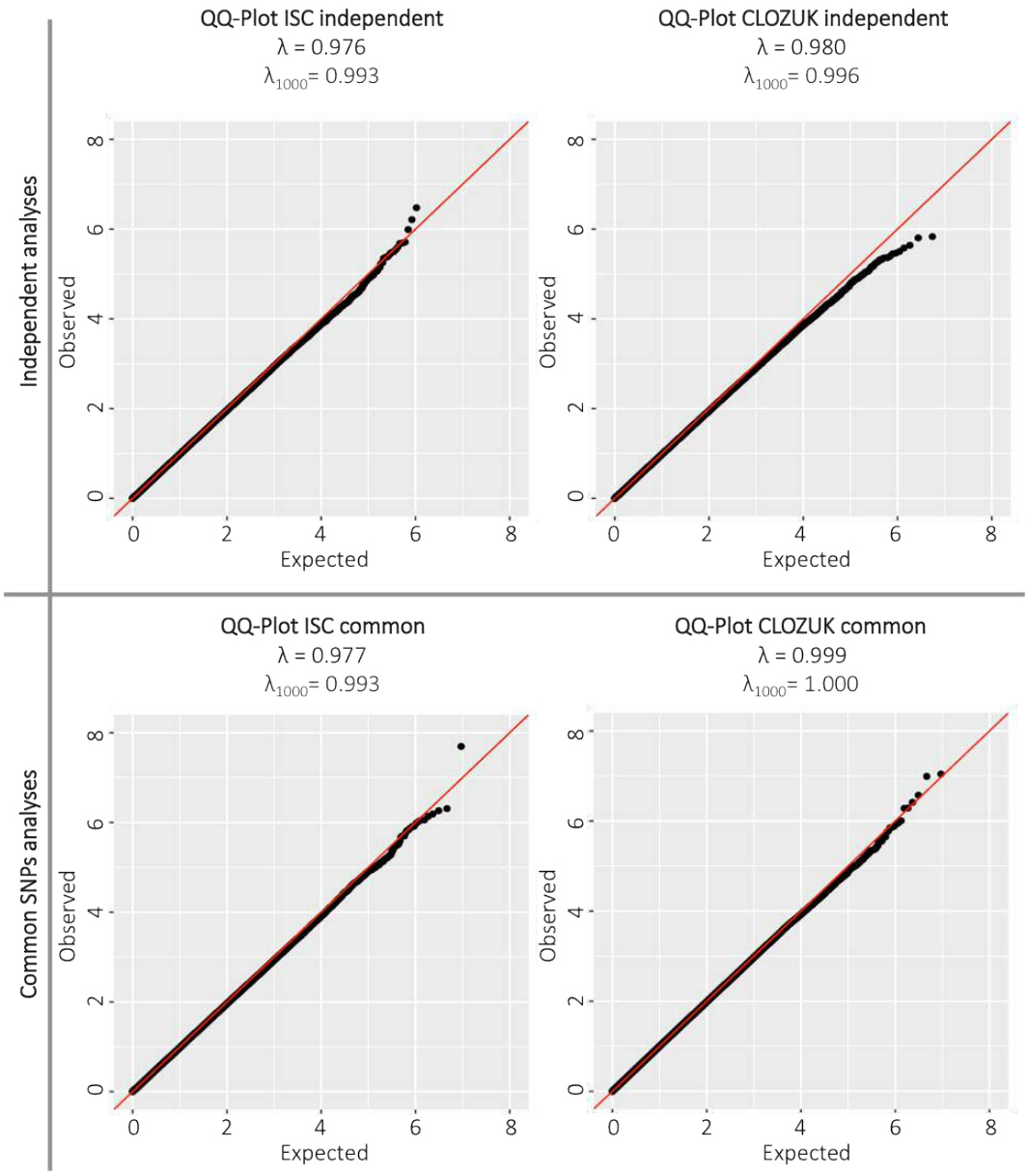


Figure 4.1: Quantile-quantile plots for all pair-wise interactions analysis calculated for each dataset. The top panels show the independent SNPs analysis for the ISC (top left) and the CLOZUK (top right) datasets. The bottom panels show the common-SNPs analysis for the ISC (bottom left) and the CLOZUK (top right) datasets.

	Independent analysis		Common SNPs analysis	
	ISC	CLOZUK	ISC	CLOZUK
Number of SNPs before LD pruning	92,860	1,751,178	71,038	71,038
Number of SNPs after LD pruning	2,898	3,318	4,318	4,318
Number of associated genes	2,080	2,418	2,791	2,791
Number of calculated interactions	4,197,753	5,502,903	9,320,403	9,320,403

Table 4.2: Number of SNPs (before and after LD pruning), number of associated genes and number of calculated interactions in the two analyses (Independent and common-SNPs) for each dataset.

For the common-SNPs analysis, after selecting the common variants between the two populations, 71,038 SNPs remained. LD pruning was performed on the larger sample (CLOZUK) resulting in the selection of 4,318 SNPs within 2,791 genes that were used to perform the pair-wise analysis (Table 4.2). Out of the calculated interactions, none survived Bonferroni correction in either dataset (Table 4.3).

	Independent analysis		Common SNPs analysis	
	ISC	CLOZUK	ISC	CLOZUK
Smallest observed p-value	$6.15e^{-7}$	$1.47e^{-6}$	$2.02e^{-08}$	$9.06e^{-08}$
Bonferroni Threshold	$1.19e^{-8}$	$9.09e^{-9}$	$5.36e^{-09}$	$5.36e^{-09}$
Number of interactions below the Bonferroni Threshold	0	0	0	0

Table 4.3: Smallest interaction p-value, Bonferroni threshold and number of calculated interactions below that threshold in the two analyses (Independent and common-SNPs) for each dataset. None of the interactions passed the Bonferroni threshold correction.

4.3.2 Ranking the interactions

4.3.2.1 Ranking by gene-wide p-values linked to each dataset

As detailed in section 4.2.6, interactions were first ranked by gene-wide significance calculated in the same dataset.

4.3.2.1.1 Spearman ranked correlation

Spearman ranked correlation coefficients were calculated between interaction p-values and the combined gene-wide significance of the genes involved in each interaction (Table 4.4). While both independent analyses showed a significant relationship between interaction p-values and gene-wide significance, the correlation was very small (Table 4.4) and probably only detected due to the very high number of degrees of freedom. For the common-SNPs analyses the correlation coefficient was either extremely low (CLOZUK) or low and negative (ISC) with non-significant p-values, indicating no relationship between the interaction p-values and gene-wide significance.

		Independent analysis		Common-SNPs analysis	
		ISC	CLOZUK	ISC	CLOZUK
Spearman ranked correlation	R	$2.36e^{-3}$	$5.14e^{-3}$	$-1.63e^{-04}$	$2.72e^{-04}$
	P-value	$1.36e^{-6}$	$2.2e^{-16}$	0.618	0.406

Table 4.4: Spearman ranked correlation calculated for each dataset (ISC and CLOZUK) in both analyses (Independent and common-SNPs) between the interaction p-values and the gene-wide p-values.

4.3.2.1.2 Linear Regression Model

To assess whether SNPs from genes with greatest disease effects tend to have more significant interactions, a linear regression model taking into account the number of variants per genes and the length of the gene was used. For the independent analysis, the adjusted R-square (Table 4.5) showed that a small part of the variance in the gene-wide significance variable can be explained by the predictor variables (SNP-SNP interaction p-values, number of SNPs and length of each gene). In addition it was found that the interaction p-values were significant predictors but did not play a big role. However for the

common SNPs analysis, the adjusted R-square are very low and the interaction p-values are not significant predictors (Table 4.5).

		Independent analysis		Common-SNPs analysis	
		ISC	CLOZUK	ISC	CLOZUK
Adjusted R-squared		0.107	0.0726	$1.131e^{-3}$	0.0436
Coefficient for interaction p-values	Estimate	$4.37e^{-3}$	$6.88e^{-3}$	$1.05e^{-3}$	$5.07e^{-04}$
	P-value	$1.23e^{-3}$	$3.15e^{-10}$	0.141	0.435

Table 4.5: Adjusted R-squared, beta coefficient for interaction p-values in the linear regression.

4.3.2.1.3 Ranking test

As explained in section 4.2.7 a one-sided Mann Whitney Wilcoxon ranking test was used to quantify the over-representation of small interactions p-values within the most highly schizophrenia associated genes compared to interactions within genes least associated with the disease.

The comparison was done in both analyses (Figure 4.2): the independent analysis and the common SNPs analysis. In the independent analysis, the over-representation p-values were significant in both ISC and CLOZUK datasets (Figure 4.2, top panel) indicating an excess of significant interactions within the genes that are most highly associated with the disease. The effect was much more pronounced for the CLOZUK dataset than it was for the ISC dataset (Figure 4.2, top panel). In the common-SNPs analysis the over-representation p-values were not significant in both ISC and CLOZUK datasets (Figure 4.2, bottom panel)

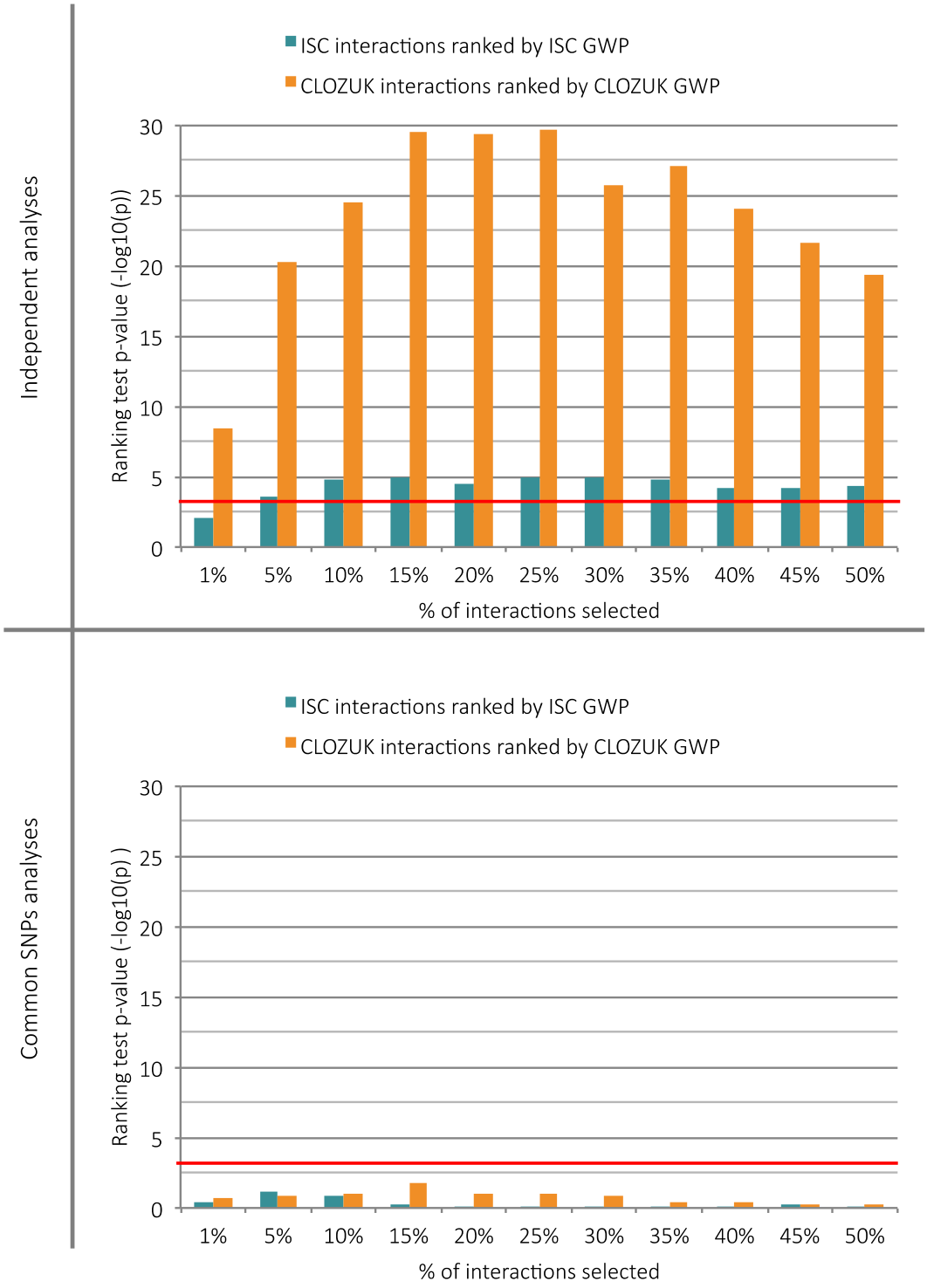


Figure 4.2: Histograms of the ranking test (-log10(p-value)) for the independent analyses (top panel) and the common SNPs analysis (bottom panel) for both datasets (ISC and CLOZUK). This test compares the distribution of N% SNP-SNP interaction p-values between SNPs in genes highly associated with schizophrenia to the distribution of N% SNP-SNP interaction p-values between SNPs in genes least associated with disease. GWP refers to gene-wide p-value. The red line shows the multiple test significance thresholds for a corrected p-value of 0.05 (Bonferroni correction p-value p=0.0022). An excess of significant interactions within the genes that are most highly associated with the disease is observed for ISC and CLOZUK in the independent analysis (top panel).

4.3.2.2 Ranking by gene-wide p-values from the other dataset and the PGC

In this section, interactions were ranked by gene-wide significance of the corresponding genes as calculated in the other dataset (CLOZUK interactions ranked by gene-wide p-values calculated in the ISC dataset and vice versa) and from the Psychiatric Genomics Consortium 2 (PGC2). The ISC and CLOZUK datasets are both part of the PGC2 dataset. (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). As in the previous section ranking was done for both analyses (independent and common-SNPs) to compare the results (Figure 4.3).

In the independent analysis when ranking interactions from the CLOZUK dataset using the gene-wide p-values calculated for the ISC dataset (and vice versa), the enrichment for significant interactions disappeared in both cases (Figure 4.3, top panel) However when ranking by the gene-wide p-values calculated for the PGC2 dataset (which contains both ISC and CLOZUK data), in both ISC and CLOZUK dataset an enrichment is detected.

In the common-SNPs analysis, only when ranking interactions from the ISC dataset using gene-wide p-values from CLOZUK was enrichment observed in some cases (35%, 40% and 45% of the distribution) (Figure 4.3, bottom panel).

Furthermore, comparison of correlation coefficients between the ISC, CLOZUK and PGC2 gene-wide significance p-values (Table 4.6) revealed low correlation between the ISC and CLOZUK. The strongest observed relationship is between the CLOZUK and PGC2 gene-wide significance (R=0.294).

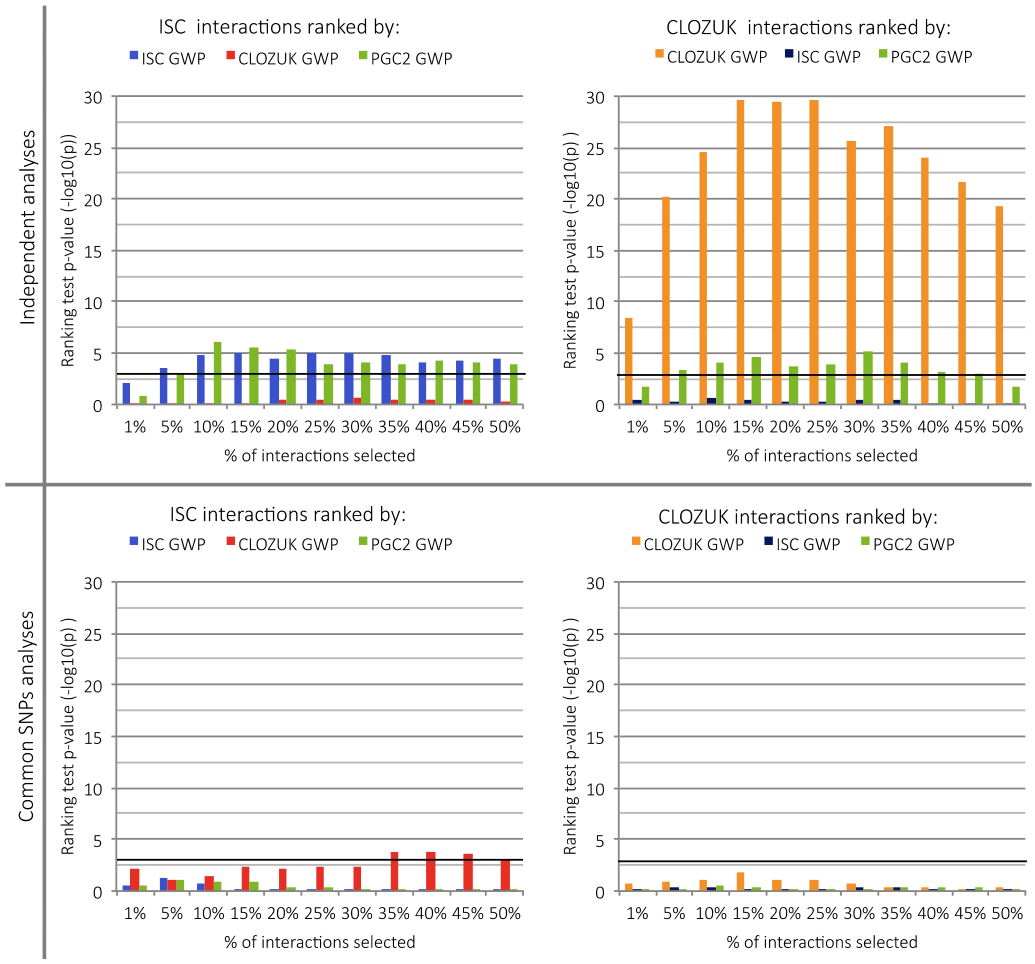


Figure 4.3: Histograms of the ranking test ((-log10(p-value)) for the independent analyses (top panel) and the common SNPs analysis (bottom panel) for both dataset (ISC and CLOZUK). The ISC interactions were ranked by ISC, CLOZUK and PGC2 GWP. The CLOZUK interactions were ranked by CLOZUK, ISC and PGC2 GWP. GWP=gene-wide p-value. The black line shows the multiple test significance thresholds for a corrected p-value of 0.05 (Bonferroni correction p-value p=0.0015). An excess of significant interactions within the genes that are most highly associated with the disease is observed for ISC and CLOZUK in the independent analysis (top panel) when ranking by the GWP (gene-wide p-value) from the dataset (ISC for ISC and CLOZUK for CLOZUK) or from the PGC2.

Correlation between:		ISC and CLOZUK GWP	ISC and PGC2 GWP	CLOZUK and PGC2 GWP
Pearson correlation	R	0.049	0.109	0.272
	P-value	$5.20e^{-9}$	$<2.2e^{-16}$	$<2.2e^{-16}$
Spearman rank correlation	R	0.049	0.128	0.294
	P-value	$6.85e^{-9}$	$<2.2e^{-16}$	$<2.2e^{-16}$

Table 4.6: Calculated correlation coefficients (Pearson and Spearman) between the gene-wide significance p-values (GWP) of ISC, CLOZUK and PGC2.

4.3.2.3 Ranking by SNP p-values

To investigate whether SNP-SNP interaction p-values were influenced by the SNP p-values, a similar ranking by the SNP p-values was performed (Figure 4.4). In the independent analysis, when ranking interactions in the CLOZUK dataset by SNP p-values (Figure 4.4, top right) an enrichment was consistently detected. In the ISC dataset, enrichment was only observed when comparing 35 and 40% of interactions in high ranked genes with interactions in low ranked genes (Figure 4.4, top left). In the common-SNPs analysis, no enrichment was detected in either dataset (Figure 4.4, bottom).

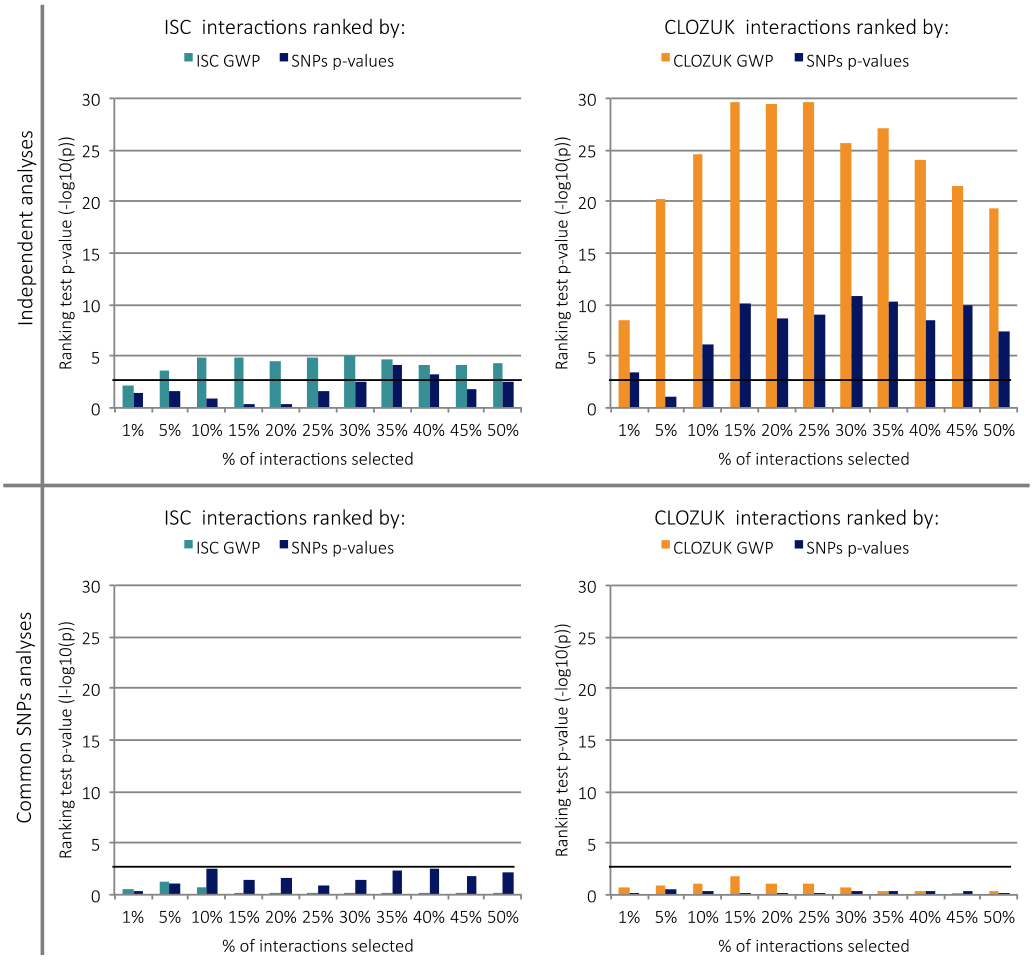


Figure 4.4: Histograms of the ranking test (log scale) for the independent analyses (top panel) and the common SNPs analysis (bottom panel) for both dataset (ISC and CLOZUK). GWP refers to gene-wide p-value. The ISC interactions were ranked by ISC GWP and the SNP association p-values. The CLOZUK interactions were ranked by CLOZUK GWP and the SNP association p-values. GWP=gene-wide p-value. The black line shows the multiple test significance thresholds for a corrected p-value of 0.05 (Bonferroni correction p-value p=0.0015).

4.3.2.4 Ranking by gene-wide p-values from other disorders

The ability of gene-wide p-values to identify genes enriched for nominally associated interactions suggests that disease-relevant interactions are indeed more likely to occur between genes themselves associated with the disorder but also that some interactions are specific to each dataset. To investigate whether this result could instead be driven by confounding factors (e.g. differences in SNP density or genotyping quality) present in GWAS data but unrelated to disease, interactions were re-ranked using gene-wide p-values from other neuro-psychiatric/-degenerative disorders: Alzheimer Disease (Lambert et al. 2013), Parkinson Disease (Nalls et al. 2014) and Bipolar Disorder (Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011). Of these, only Bipolar Disorder is known to have a strong genetic overlap with schizophrenia (Lichtenstein et al. 2009). Assuming the effect to be genuine, Bipolar Disorder gene-wide p-values might potentially have the ability to identify genes enriched for schizophrenia-associated interactions.

In the independent analysis, when ranking the interactions results of the ISC dataset using the gene-wide association p-values from Alzheimer Disease and Parkinson Disease, no enrichment was observed (Figure 4.5, top left). However, there was an enrichment of significant interactions when ordering the interactions using gene-wide p-values for Bipolar Disorder (Figure 4.5, top left). The observed effect was the strongest when N=10% (p-value=$3.72e^{-4}$). When ranking interactions calculated on the CLOZUK dataset, no ranking test passed the multiple correction threshold for a corrected p-value of 0.05, the smallest p-value occurring when ranking by the Parkinson gene-wide p-value (p-value=$2.21e^{-3}$ when N=35%).

In the common-SNPs analyses, when ranking the interaction from the ISC dataset using gene-wide p-values form other psychiatric disorders, no test was significant; the same was observed in the CLOZUK dataset (Figure 4.5, bottom). Consequently this shows that the observed results are probably not due to confounding factors inherent to GWAS data.

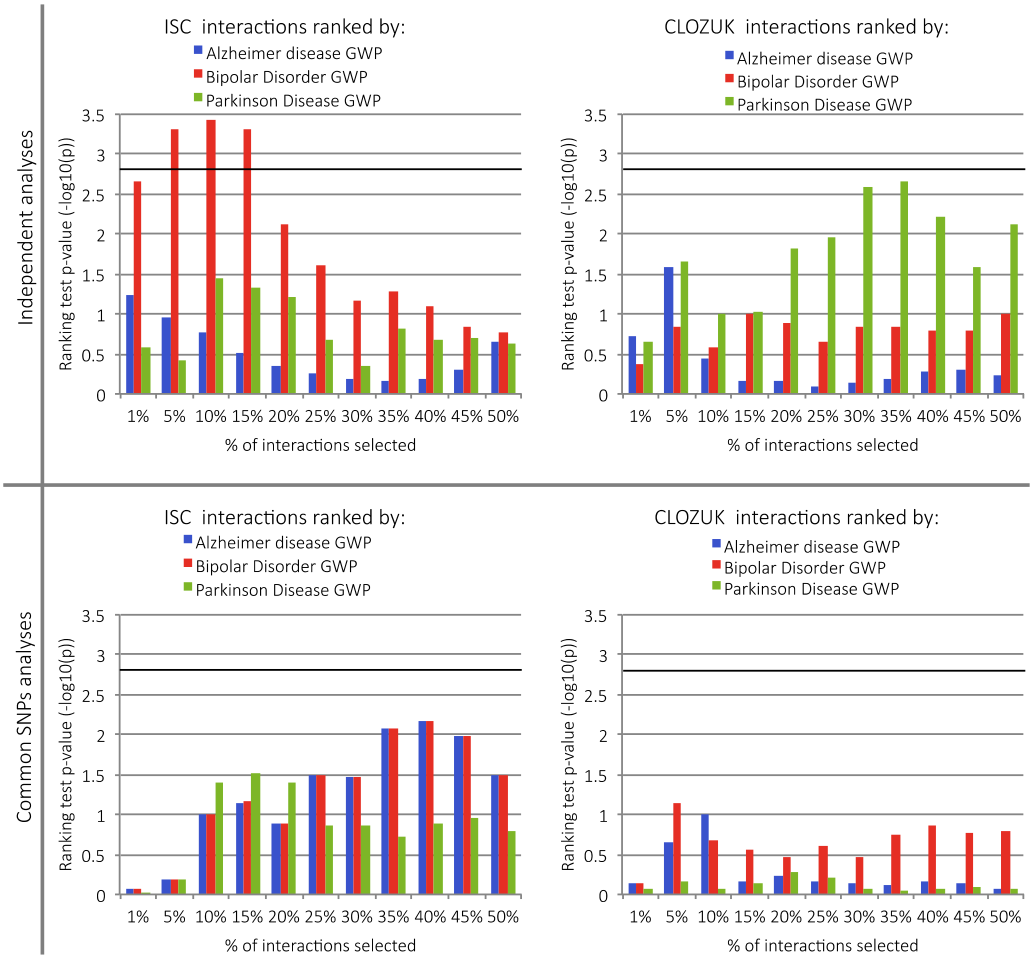


Figure 4.5: Histograms of the ranking test (log scale) for the independent analyses (top panel) and the common SNPs analysis (bottom panel) for both dataset (ISC and CLOZUK). The interactions were ranked by Alzheimer disease, Bipolar disorder and Parkinson Disease GWP. GWP refers to gene-wide p-value. The black line shows the multiple test significance thresholds for a corrected p-value of 0.05 (Bonferroni correction p-value p=0.0015). For the independent analysis, using the ISC dataset, an excess of significant interactions within the genes that are most highly associated with the Bipolar Disorder is observed.

4.3.2.5 Ratios

To further quantify the enrichment effect observed in the independent analysis, ratios of the number of interactions below a significance level α were computed. As for the ranking test, the ranked SNP-SNP interactions were divided in two groups: N% of the 'top' SNP-SNP interactions (top referring to interactions involving genes most highly associated with the disease) and N% of the 'bottom' SNP-SNP interactions (bottom referring to interactions

involving genes least highly associated with schizophrenia), N varying from 1% to 50%. Then, interactions with a p-value below the significance level α were counted in each group in order to calculate the ratio between them. This provided more detail on the level of significance driving the effect: a ratio above 1 indicates an excess of interactions with a p-value below the significance level α within interactions in high ranked genes compared to interactions in low ranked genes (Figure 4.6).

When comparing interaction ratios in both the CLOZUK and ISC datasets, there was an excess of interactions below the significance level α in high ranked genes (i.e. ratio >1) for almost all α. The greatest excess occurred when α is between 0.01 and $1e^{-4}$. The highest ratio observed for the CLOZUK dataset is 1.531 (N=1%, $\alpha=10e^{-3}$) and for the ISC dataset is 1.571 (N=10%, $\alpha=10e^{-4}$).

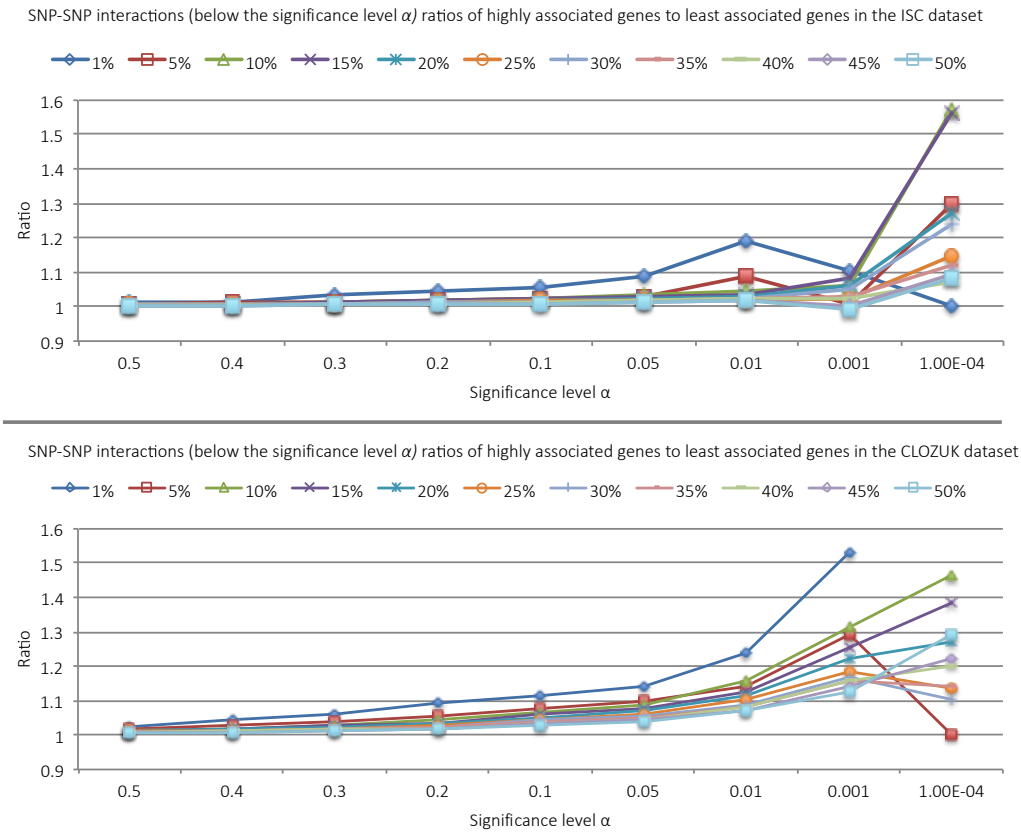


Figure 4.6: Calculated ratios for the ISC (top panel) and CLOZUK (bottom panel) datasets of SNP-SNP interactions with a p-value below the significance level α (from 0.5 to $1e^{-4}$) in two groups: interactions involving genes most highly associated with schizophrenia and interactions involving genes that are least associated with the disease.

4.3.2.6 Networks

As explained in section 4.2.8, it is possible that a proportion of disease-relevant SNPs (or genes) will contribute to multiple interactions. Two sets of the ranked interactions were plotted as network: 1% of SNP-SNP interactions involving genes most highly associated with the disease (highly ranked) and 1% of SNP- SNP interactions involving genes least associated with schizophrenia (lowest ranked). This particular threshold was used as it gives the clearest visualisation, however the same phenomenon is observed at different thresholds with a more complex graph structure.

Each node represents a SNP (or a gene) and the link joining two SNPs (or genes) represents the existence of a pairwise interaction between them with $p<0.01$.

This was done on the independent comparison for both the CLOZUK and the ISC datasets.

	ISC		CLOZUK	
	HRG	LRG	HRG	LRG
Number of interactions with p<0.01	435	366	512	414
Number of SNPs	427	218	478	354
Number of genes	368	302	437	306

Table 4.7: Number of interactions, number of SNPs and genes in each dataset after selecting 1% of SNP-SNP interactions involving genes most highly associated with the disease (highly ranked) with a p-value below 0.01 and 1% of SNP- SNP interactions involving genes least associated with schizophrenia (lowest ranked) with a p-value below 0.01. (HRG: high ranked genes and LRG: low ranked genes)

When plotting the network graphs (Figure 4.7 for CLOZUK and Figure 4.8 for ISC), different patterns can be observed. For the CLOZUK dataset in the network of interactions in high ranked genes (Figure 4.7A) for both the SNPs and the genes network, interactions are

clearly concentrated around a small number of hub SNPs (or genes). This is not observed in the networks of interactions in low ranked genes (Figure 4.7B). The same phenomenon was also observed in the ISC dataset (Figure 4.8). This indicates that a handful of SNPs are responsible for driving the interactions in the top of the distribution (Table 4.7).

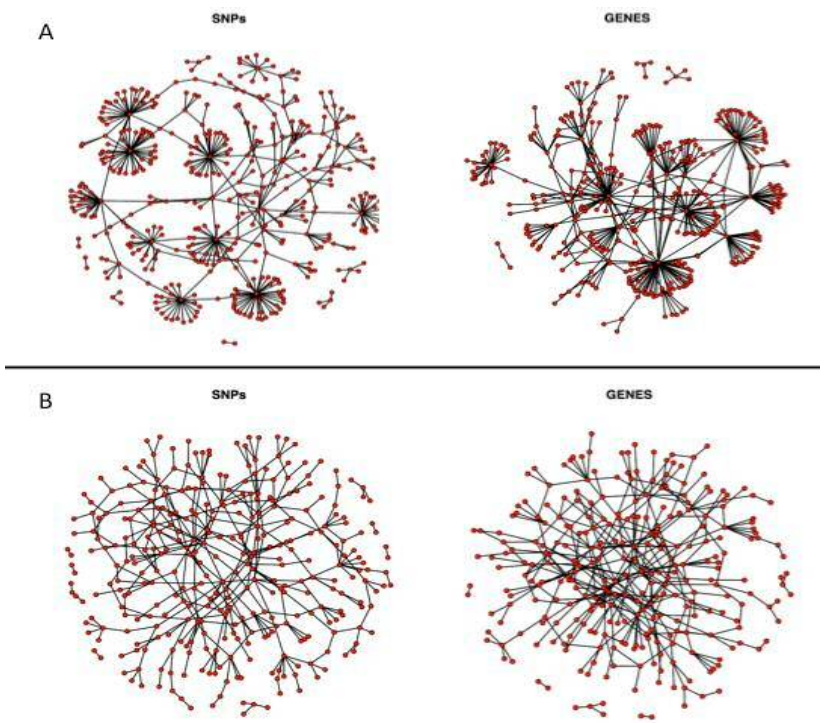


Figure 4.7: Networks of SNPS and Genes interactions in the CLOZUK dataset. Interactions were ranked by gene-wide significances. 1% of interactions was selected with the highest ranking (A) (genes highly associated with schizophrenia) and with the lowest ranked interactions (B) (genes least associated). In both groups, interactions with p< 0.01 were selected to drawn the network. For interactions in high ranked genes, the network appears to present hubs both in SNPs and Genes networks (A).

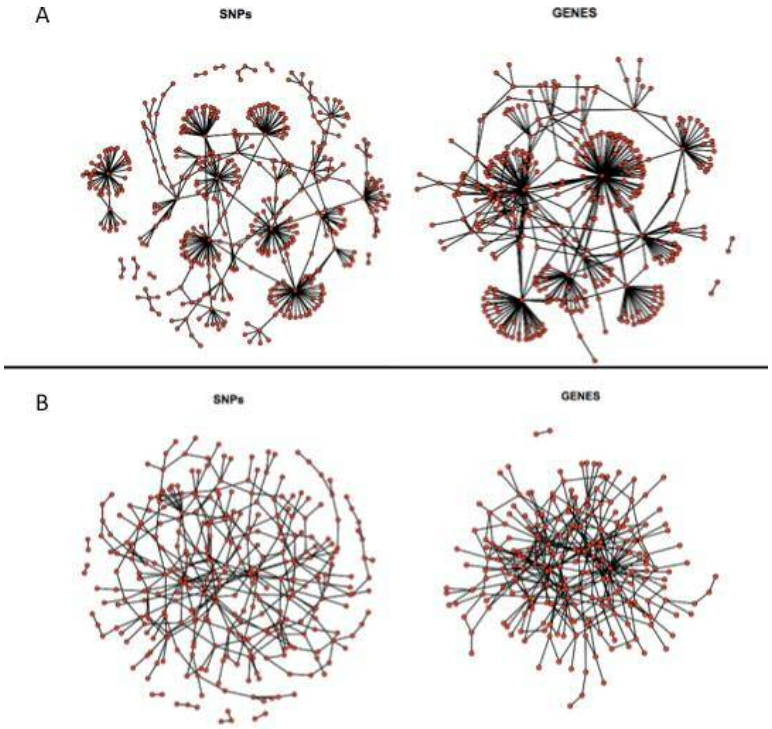


Figure 4.8: Networks of SNPS and Genes interactions in the ISC dataset. Interactions were ranked by gene-wide significances. 1% of interactions was selected with the highest ranking (A) (genes highly associated with schizophrenia) and with the lowest ranked interactions (B) (genes least associated). In both groups, interactions with p< 0.01 were selected to drawn the network. For interactions in high ranked genes, the network appears to present hubs both in SNPs and Genes networks (A).

The hubs were further examined to assess whether additional biological information could be derived from them. The top 5 hubs (the hubs with the highest degree) in each dataset were selected. The DAVID database (Sherman et al. 2007) was used to assess whether any biological process (GO term) was shared between genes connected to the same hub compared to other genes present in the network. This GO analysis looked for evidence supporting a link between the observed hubs and biological processes. In both datasets, no enrichment for GO term was detected.

In order to assess whether the observed hub-effect was a statistical artefact, I investigated the relationship between the gene degree (number of link) in each network (Figure 4.7 and Figure 4.8) and the gene length. As showed by the correlation table (Table 4.8), there is no relationship between the gene length and the gene degree in the high ranked genes network. The opposite is observed for the lowest ranked genes network.

Correlation between number gene degree and gene length		ISC		CLOZUK	
		HRG	LRG	HRG	LRG
Pearson correlation	R	$1.03e^{-2}$	0.513	0.0625	0.280
	P-value	0.843	$< 2.2e^{-16}$	0.262	$1.53e^{-6}$
Spearman rank correlation	R	$-5.27e^{-3}$	0.399	$-4.72e^{-2}$	0.222
	P-value	0.919	$1.27e^{-10}$	0. 3976	1.57e-4

Table 4.8: Calculated correlation coefficients (Pearson and Spearman) between the number of gene degree and the gene length in the hub network for both the ISC and CLOZUK dataset. HRG=High Ranked Genes. LRG=Least Ranked Genes.

Similarly, the relationship between the gene degree in the hub network and the number of SNPs per gene was investigated. The correlation table (Table 4.9) showed a strong relationship between the gene degree and the number of SNPs per gene in both high ranked genes and low ranked genes networks.

Correlation between number of gene degree and number of SNPs		ISC		CLOZUK	
		HRG	LRG	HRG	LRG
Pearson correlation	R	0.479	0.784	0.458	0.622
	P-value	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$
Spearman rank correlation	R	0.599	0.608	0.435	0.408
	P-value	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$1.10e^{-13}$

Table 4.9: Calculated correlation coefficients (Pearson and Spearman) between the number of gene degree and the number of SNPs per gene in the hub network for both the ISC and CLOZUK dataset.

4.3.3 Overall assessment of the functional annotation

SNP-SNP interactions with a p-value below $1x10^{-4}$ were further analysed using the DAVID database to investigate if biological processes were shared among them (Table 4.10). I compared the results obtained for the CLOZUK dataset with those for the ISC dataset. This purely descriptive analysis was performed both for the results obtained from the independent comparison (using different SNPs for both datasets) and for those from the common-SNP comparison. Due to lack of time, this analysis only aims to give a brief description of preliminary results that would need further investigation to be complete. Annotation category with a p-value equal or smaller than 0.05 was considered as potentially relevant (Sherman et al. 2007).

	Independent analysis		Common-SNPs analysis	
	ISC	CLOZUK	ISC	CLOZUK
Number of genes	502	600	1,036	1,102
Number of genes used as background list	2,080	2,418	2,791	2,791

Table 4.10: Number of genes into the main and the background list for the analysis in David.

In the independent analysis, 14 GO-terms were identified as significant for the ISC dataset and 6 GO-terms for the CLOZUK dataset (Figure 4.9, top). No shared process between the two was found.

In the common SNPs analyses, 7 GO-terms were found to be enriched in the ISC dataset and 5 in the CLOZUK dataset (Figure 4.9, bottom). One shared process was identified between the two dataset: GO:0007156 homophilic cell adhesion via plasma membrane adhesion molecules. Other similar functions are shared: transmembrane transport and calcium ion transmembrane transport.

Taking into account multiple test correction (FDR rate), no GO terms was found to be significant in either analyses.

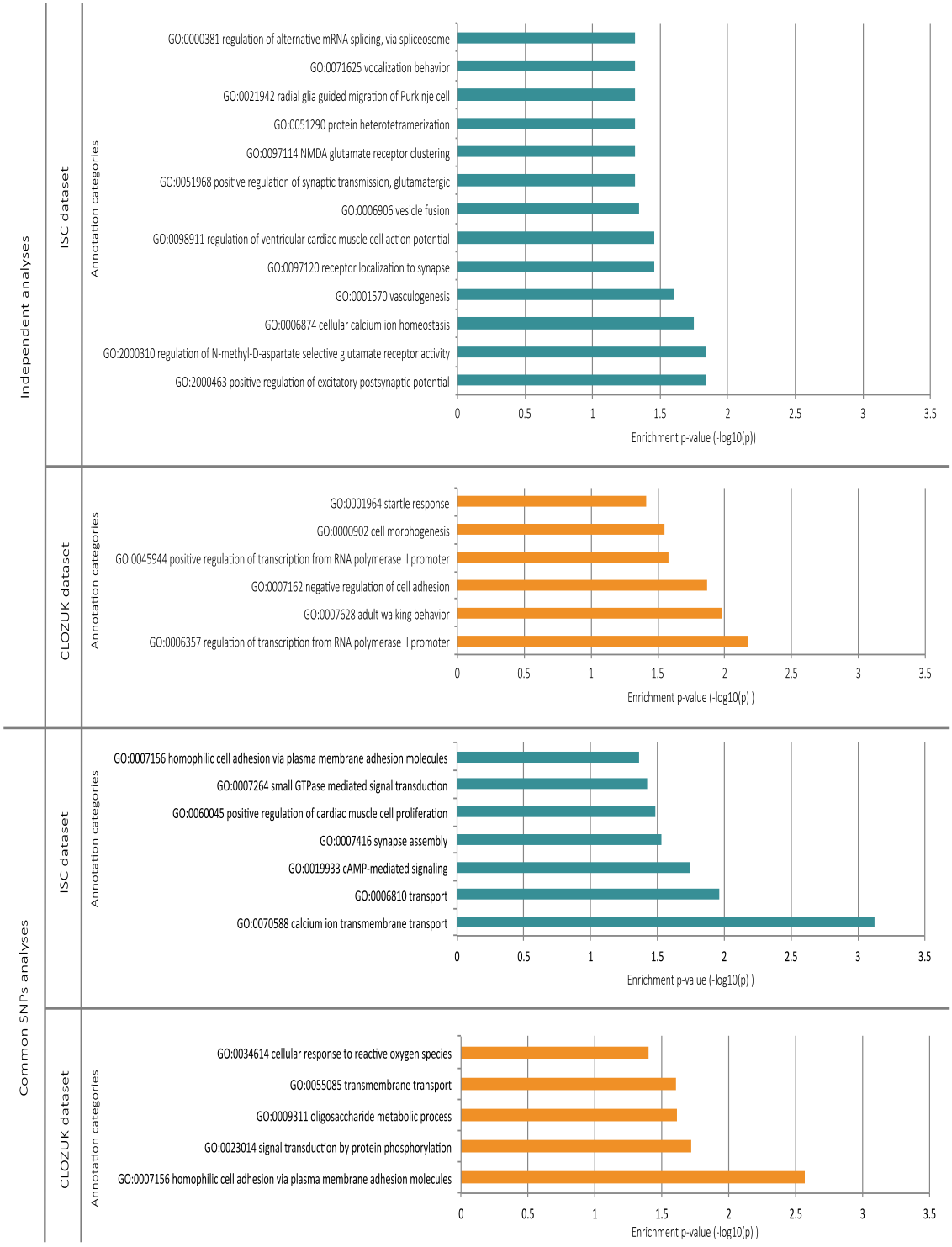


Figure 4.9: GO-terms analysis for the independent and same-SNPs comparison in both ISC and CLOZUK datasets.

4.4 Discussion

4.4.1 Towards a better understanding of gene-gene interactions in schizophrenia

A detailed study of genetic interactions was performed on two non-overlapping schizophrenia GWAS datasets. Analyses were first performed independently in each dataset; the SNPs used being different and specific to each dataset. To investigate whether results were sensitive to differences in the SNP content of the genotyping chips used in each study, and to make the analysis more directly comparable, the procedure was re-run after selecting for common variants present in both datasets. In every analysis, no interaction survived correction for multiple testing. This is unsurprising as the sample sizes are still insufficient to reliably detect individual interactions statistically.

To investigate whether genetic information alone could be used to identify sets of genes enriched for interactions, SNP pair-wise interactions were calculated and ranked by the gene wide significance of the genes involved in each interaction. First the comparison of Spearman correlation coefficients between the interaction p-values and the gene-wide significance p-values revealed a weak relationship between the two in the independent analysis (for both CLOZUK and ISC datasets). However this was not the case for the common-SNPs analyses.

The distribution of interaction p-values in high ranked genes (for which there was greatest gene-level evidence of main effects contributing to disease) was then compared to that in low ranked genes. Analysing samples independently (selection of SNPs being specific to each datasets), there was evidence for an enrichment of significant interactions amongst genes with the highest gene-level evidence of association when compared to interactions involving the least associated genes. The observed effect was stronger for the CLOZUK than the ISC dataset: this is probably due to an insufficient power to detect the effect in the smaller set of genes in the ISC dataset (the dataset being smaller). Furthermore, as

indicated by the ratio comparison the observed effect was the strongest when comparing the distribution of interaction p-values between the 1% of SNP-SNP interactions between SNPs in genes highly associated with schizophrenia with the 1% of SNP-SNP interactions between SNPs in genes least associated with the disease. This is the first consistent evidence that genes contributing to schizophrenia risk also interact.

Investigating interactions with a p-value <0.01 in those groups, it was found that several SNPs acted as hubs for the interactions within the genes most highly associated with schizophrenia. When looking at the interactions within the other group (genes least associated with schizophrenia) no SNP hub was evident. Further assessment of the genes involved in the top five hubs did not reveal any shared biological process in the CLOZUK dataset.

Potential statistical artefacts were investigated. The relationship between the degree of each gene (number of links) and the gene length was assessed: for the low ranked genes networks in both the ISC and CLOZUK datasets the correlation coefficients showed a stable relationship whereas the opposite was observed for the high ranked genes networks. As high ranked genes tend to be larger in size than low ranked genes, it does not appear to create a bias in the high ranked genes networks as no correlation is observed between the degree and the length of each gene. A second relationship was assessed between the degree of each gene and the number of SNPs per gene. For both the high ranked genes and the low ranked genes networks a strong correlation between the variables was observed. As this effect is present in every network it does not seem to play a role in the hub pattern observed.

When performing the common-SNP comparison the previously detected enrichment effect of significant interactions within highly associated genes was no longer evident. One explanation for this would be that common SNPs between the CLOZUK and the ISC datasets

are contributing less to the interactions effects than the ones identified with the independent comparison. It is also possible that interaction effects are more sensitive to sample differences than main effects. Other factors could also explain this difference and will be detailed in the following section.

To further investigate the observed effect, different rankings were performed on the interaction data. The ranking by the gene-wide significance calculated on the PGC2 dataset showed similar effect to those previously observed: enrichment was detected in both ISC and CLOZUK datasets in the independent analysis but not in the common-SNPs analysis. However when the ranking was done using the gene-wide significance from the other dataset (using ISC gene wide significance to rank CLOZUK interactions and vice versa), no effect was observed in both the independent and the common-SNPs analyses. Several factors could contribute to these observations. Firstly the sample size, the ISC dataset is significantly smaller: perhaps there is a lack of power to evaluate the gene-wide significance in this dataset. This would explain why when using PGC2 gene-wide significance to rank ISC interactions, a bigger enrichment of 'significant' interactions in high ranked genes is detected. However this does not explain why there is enrichment in ISC interactions ranked by ISC gene-wide significance. If it is simply power that is needed to identify gene-wide association, then using the CLOZUK gene-wide significance to rank ISC interactions should be better than ranking the same interactions by ISC gene-wide significance, which is not the case. Nevertheless, PGC2 dataset is closer in size to CLOZUK dataset than ISC dataset but the ranking by PGC2 seems to produce a stronger signal in the ISC dataset than in the CLOZUK dataset. Secondly the chip type, the CLOZUK dataset uses more recent and more similar genotyping chips than many of the PGC2 samples. In addition, the ISC dataset uses a mix of Affymetrix chips. Perhaps the chip quality influences the power to detect true effects and having to correct for multiple chips, this decreases their power to identify interactions. Thirdly the sample homogeneity, the CLOZUK dataset

is a more homogeneous sample with less population sub-structure than the ISC (or the PGC2) dataset. Perhaps by using the CLOZUK dataset, population-specific effects are picked up whereas the ISC (or the PGC2) dataset will pick up more generic effects. This could explain the similarity between ISC and PGC2 and their difference to CLOZUK. The disease biology, as cases in the CLOZUK dataset are treatment-resistant it may be that a large proportion of the interactions picked up are specific to treatment-resistance. Finally looking at gene-wide p-value between the three datasets, ISC and CLOZUK are more highly correlated with PGC2 than with each other. This would explain why ranking with PGC2 gene-wide p-values gives more similar results than ranking ISC with CLOZUK gene-wide significance (and vice versa). Also it is worth noticing that the correlation between the two largest datasets (CLOZUK and PGC2) is moderate (R~0.3) indicating that gene-wide significance are still quite variable in samples of these sizes.

Following those results, I further investigated if some of the interactions results could be driven by confounding factors present in GWAS data but not related to disease. In order to assess it, all the interactions were re-ranked using gene-wide p-values from others neuro-psychiatric/-degenerative disorders. Regarding the independent comparison, the results of this cross-disorder analysis showed no evidence that the interaction results were driven by such confounds. Indeed no enrichment was detected when ranking interactions by gene-wide significance from GWAS from genetically unrelated (as far as it is known) disorders. In addition, it was found that Bipolar Disorder gene-wide p-values could be used to identify genes enriched for schizophrenia-associated interactions in one dataset, which could be explained as Bipolar Disorder and Schizophrenia share genetic factors (Lichtenstein et al. 2009) if it is a genuine effect. No replication was observed in CLOZUK, indicating that perhaps the effect is not genuine. When performing the cross-disorder investigation on

the common-SNPs comparison, no enrichment was observed in either the CLOZUK or the ISC datasets.

Finally interactions with a p-value below $1x10^{-4}$ were further investigated in order to assess if biological processes were shared among them.

The preliminary results from this descriptive analysis showed that some interesting biological processes (GO terms) involving synapses or ion transport are involved among 'significant' interactions in both datasets. In the common-SNPs analysis some processes are common between CLOZUK and ISC. In addition DAVID does not take into account gene size or the number of semi-independent variants. The larger the gene, the more likely it is to have multiple semi-independent SNPs and the more SNPs, the more likely it is to find a 'significant' interaction by chance. As brain genes tend to be larger, gene size can influence both probability of finding an interaction and the functional annotation. Further work would need to use an enrichment test that accounts for those differences.

4.4.2 Detection and replication of results: challenges and conclusion

The main challenge in a genetic interaction study has often been the detection of significant results (McCarthy et al. 2008). Indeed, very few genome-wide epistasis studies have discovered significant interactions (McCarthy et al. 2008). In this chapter, none of the individual interactions studied in the four analyses survived correction for multiple testing. Indeed, the sample sizes are still insufficient to detect individual interactions statistically. This indicates a lack of power and that individual interactions contributing to disease risk are likely to have extremely modest effect sizes. This problem could be overcome by using bigger sample size. For comparison, one of the first GWAS to detect a genome-wide significant main effect in schizophrenia required 3,322 cases and 3,587 controls (Purcell et al. 2009). For interactions, the sample size needed would be bigger (Wang and Zhao 2003). The sample size required to detect an interaction effect is inversely proportional to the

square of the effect size (Zuk et al. 2012): for N loci with similar effects, the sample size to detect N^2 interactions scales with N^4.

Reducing the number of SNPs in the interaction analysis would also contribute to diminish the multiple testing burden and make the threshold for significance less stringent. However if the effect sizes of the interactions are lower than the size of the main effects then it will become extremely unlikely to detect such effects. In addition, it could be argued that by reducing the number of SNPs in the analysis, a increasing number of the interactions passing the multiple correction thresholds could be in fact false positives. For example, we can see a parallel with early candidate genes studies where most of the significant results obtained were in fact false positive. As a result a threshold of $5e^{-8}$ has become the significance standard in GWAS (International HapMap Consortium 2005). In this thesis, a stringent Bonferroni correction is applied and sets the significance threshold as 0.05 divided by the number of tests. As the number of interactions tested is substantial, the threshold is highly conservative: the cost of avoiding too many type I error can then result in a loss of power (Musani et al. 2007). Perhaps best practice should be that no matter of the number of variants used in an interaction analysis, an appropriate genome-wide significance threshold should be used for pairwise interactions (Musani et al. 2007).

In addition, to deal with population stratification, the analysis was done separately on the eight populations of the ISC dataset. Similarly the two chips of the CLOZUK data have been analysed independently and the results have then been combined for analysis. Perhaps true interactions have been lost in that process.

Another challenge for gene-gene interaction studies is the replication of results (Hemani et al. 2014; Chu et al. 2014). In this chapter, some results are partially replicated. For example, when performing an independent comparison, when the SNPs are specific to each dataset, there is an excess of significant interactions within genes that related to schizophrenia: this effect is strongly observed in the CLOZUK dataset as well as the ISC

dataset. The difference of sample size in term of both number of cases and controls as well as number of SNPs between the CLOZUK and the ISC dataset might explain the difference. In addition as gene-wide p-values are quite variable between the datasets, the ranking of interactions is not stable: large samples are certainly needed to overcome this.

Supposing that there are multiple reasonably distinct sets of biological pathways contributing to schizophrenia, it may be that one sample is (by chance) more enriched for individuals with perturbations of one set of pathways and the other sample enriched for a different set of pathways. Then the SNP/gene main effects and interactions within each dataset would be drawn from the same set of genes but there would be much less overlap between datasets. A bigger sample, similar to PGC2, would be needed to identify signals in both sets of pathways at the same time. In addition ranking by gene-wide p-values in this better powered dataset should reveal evidence for interactions in the smaller datasets.

4.4.3 Summary of the chapter

In this chapter, I investigated SNPs-SNPs interactions in two independent and non-overlapping GWAS datasets. When performing the independent comparison (markers specific to each dataset), an enrichment of significant interactions within genes that are most highly associated with schizophrenia was detected. However that effect was not observed when using the common SNPs between the datasets.

Chapter 5 - Interaction in GWAS datasets using data from protein-protein interactions

5.1 Introduction

5.1.1 Background

Finding gene-gene interactions in human genetic data has proven to be a difficult task (Gilbert-Diamond and Moore 2011). As illustrated in section 4.2.5 of Chapter 4, many factors contribute to the explanation of a lack of success in interaction detection: low statistical power, severe multiple testing correction, small effect size, lack of large sample size (Moskvina et al. 2011).

To help overcome such problems, the prioritisation of the SNPs used in the interaction analysis is a key factor (Moskvina et al. 2011). In Chapter 4, SNPs within genes were selected using LD pruning. In this chapter I investigate whether functional information such as protein-protein interactions can be used to prioritise genes.

Protein-protein interactions (PPI) are physical contacts between proteins that allow the formation of more complex functional units. As protein rarely acts alone, such units are particularly important as many molecular processes depend on it. This explains the growing interest in the study of such units: for example one study (Pawson and Nash 2000) have shown that the specificity in signal transduction relies on PPI.

As detailed in section 1.3.3 of Chapter 1, large-scale studies of PPI have allowed the identification of thousands interactions as well as the creation of several public PPI databases such as BIND (Bader et al. 2003), HPRD (Mishra et al. 2006) or Intact (Kerrien et al. 2007).

Furthermore some efforts have been made to obtain comprehensive screens (which investigate all potential interactions within a particular set of proteins) with PPI thoroughly

verified. However while comprehensive screens are more likely to contain good quality PPI, they are still far from complete and a large proportion of all potential interactions have yet to be investigated.

Nevertheless, newly discovered PPI are mostly found in small studies rather than in large comprehensive screenings. Furthermore their number keeps increasing exponentially with the progress of the technologies allowing their detection. As a result, the majority of newly discovered PPI will be available mostly in the literature, which renders the task of extracting all that information very difficult.

5.1.2 Aim of the chapter

The aim of the chapter is to investigate the relationship between biological interactions such as PPI and statistical interactions. PPI from both databases and literature curation will be used to assess the relationship.

The first part of the chapter reviews three text-mining tools extracting PPI from abstracts and compare their performances. The best tool was then used to identify a set of PPIs through analysis of PubMed abstracts.

The second part of the chapter will assess whether significant SNPxSNP interactions are enriched within sets of PPI from different source.

In this chapter, screen PPI will refer to PPI discovered through high throughput screens (e.g. yeast two hybrid assays). Literature PPI will refer to PPI discovered through the analysis and curation of the literature (i.e. the results of numerous small-scale interaction studies), using text-mining tools for example.

5.2 Materials and Methods

5.2.1 GWAS data

For this analysis, the CLOZUK dataset was used (Hamshere et al. 2013). It consists of a total of 5,200 individuals with schizophrenia and 5,987 healthy subjects from 8 different

populations. The data, including the quality control steps applied to it, was described in Chapter 2.

The CLOZUK dataset was preferred to the ISC dataset (International Schizophrenia Consortium 2008) due to its larger sample size: higher number of cases and controls. A meta-analysis of both datasets would have also solved the sample-size issue but the number of common variants between the two datasets was deemed too low to proceed.

5.2.2 Text Mining

5.2.2.1 Background

As discussed in Chapter 1, advances in Natural Language Processing (NLP) techniques have allowed the development of tools for extracting information such as PPIs from text.

After reviewing the available literature on the different PPI extraction tools, three were selected for further comparison (Table 5.1): ODIN (Rinaldi et al. 2014), PPIInterFinder (Raja et al. 2013) and @Note (Lourenço et al. 2009). The selection criteria for this review included the availability of the tools as well as the described performance.

Name	Type	Limitations	Strengths
PPIInterFinder	Web based	Limited number of abstracts to be processed (10)	Three different options to upload the input
ODIN	Web based	Not fully publicly available for personal use	Processed on the entire PubMed
@note2	open source	Limited number of abstracts to be processed (100)	Very user friendly

Table 5.1: Limitations and strengths of the different PPI extraction tools

ODIN is a web service that allows the extraction of PPI from abstracts. It uses the annotation standard BioC (Comeau et al. 2013). BioC is a data format based on the XML language: it facilitates the data exchange for systems that process biological texts. In addition ODIN has been presented at the Bio Creative challenge II.5 (Krallinger et al. 2011), where the results proved it to be an efficient and competitive tool. Its high ranking (best

results for the detection of protein-protein interactions) and its availability (not every tool participating to the challenges were public available) made it a strong candidate for the analysis.

PPIInterFinder is a web-based tool that extracts human PPIs from abstracts of the biomedical literature. This tool uses the co-occurrence of protein names combined with a search for relational keyword in order to identify PPI candidate (Raja et al. 2013). The upload was limited to 10 abstracts at one time, making it quite hard to use for a large-scale analysis.

@Note is a platform for Biomedical Text Mining that processes abstracts as well as full texts and retrieves PPI information within them. The software can be easily downloaded and has a very user-friendly interface (Lourenço et al. 2009). However the analysis was limited by an upload of 100 abstracts at one time.

5.2.2.2 Corpora

In order to evaluate the different text-mining tools and to assess their performance, each tool was tested on two selected corpora. The first corpus consisted of a collection of 100 abstracts, created by a simple query in PubMed using the following Mesh Terms: protein-protein interactions. Every abstract was manually checked to insure that the corpus contained a sufficient number of PPI. The second corpus was a sample of 100 abstracts randomly selected from the ACT corpus used in a well-known text mining event: the BioCreative Challenge (Krallinger et al. 2011).

5.2.2.3 Indicators for performance assessment

The approach used here consisted of testing each PPI extraction tool against the two corpora defined above. Three main indicators are commonly used to evaluate the performance of text mining tools: the recall, the precision and the F-measure. Those indicators are defined using the number of correct interactions detected (true positives

TP), the number of missed interactions (false negatives FN), the number of false interactions detected (False positives FP).

The recall is the true positive rate or sensitivity: it measures the true positives that are identified and takes into account the number of those missed. It is defined as the following:

The precision sometimes known as positive predictive value measures the number of correctly retrieved PPI by a method. It is defined as the following:

It has been showed that these two measures are often negatively correlated (Rebholz-Schuhmann et al. 2012): if the precision is increased, the recall measure can decreased. The best tool is the one that manage the best balance between those two measures: the F-measure. The F-measure is the harmonic mean of precision and recall and provides a single indicator:

5.2.3 Protein-protein interaction data

5.2.3.1 Background

Finding PPI has been an important challenge in the last decade to identify complex molecular mechanisms involved in a cell. As a result, the identification of thousands interactions has enabled the creation of numerous resources collecting the data together such as PPI databases as defined in Chapter 1, section 1.3.3. In addition, comprehensive screens of PPI have allowed the emergence of good quality and thoroughly verified PPI resource.

However both types of resource present some limitations and challenges to overcome as seen in Chapter 1, section 1.3.3. For example comprehensive screens are highly dependent

on the number of proteins covered by the screen: in particular they might not cover synaptic PPI.

For those reasons, I chose to use PPI issued from both resources: data from the Human Interactome Project (Rolland et al. 2014), a comprehensive screen of PPI, and data from two databases: the String database (Szklarczyk et al. 2015) and the synaptic set of PPI within the Intact database (Orchard et al. 2014). Differences between the sets will be briefly assessed in this chapter and I will investigate whether those sets are enriched in statistical interactions.

5.2.3.2 Human Interactome Project

The Human Interactome Project (HIP) aims to build a reference map of the human PPI network by describing all of the physical interactions between each protein. The approach consists of using yeast two-hybrid assay (Y2H) experiments in Human Embryonic Kidney cells (HEK293) to obtain high quality interactions. As HEK293 is a cell line derived from kidney cells, it is indeed possible that this dataset contains tissue specific PPIs but it will also contains non-tissue specific interactions that are of interest. In addition, PPIs have rarely been identified in the context of distinct cell types but a few studies are aiming to correct this (Yeger-Lotem and Sharan 2015).

Five datasets of PPI are available on the HIP website which were used to form two sets of PPI.

The two proteome-scale efforts HI-I-05 (Rual et al. 2005) and HI-II14 (Rolland et al. 2014) were grouped with two other maps containing high quality interactions identified during the development of the protocol: Venkatesan-09 (Venkatesan et al. 2009) and Yu-11 (Yu et al. 2011). Hi-I-05 screened a space containing over 7,000 genes and identified over 2,700 binary PPI of high quality. HI-II-14 generated over 14,000 PPI. Venkatesan-09 contains over 200 high-quality Y2H PPI and Yu-11 over 1,200 interactions.

The second set of PPI drawn from the HIP consisted of PPI obtained by extraction of information from known PPI databases such as BIND, BioGrid or DIP: Lit-BM-13 (Rolland et al. 2014). Out of those preliminary results, HIP selected only the PPI that were supported by at least two pieces of evidence from the literature (Rolland et al. 2014).

To conclude, in the following analysis, two sets of PPI were derived using the HIP data: one supported experimentally and the other one supported by the literature.

5.2.3.3 Intact Database

The Intact database (Orchard et al. 2014) is an open source database containing interaction data curated from the literature and also from direct depositions. Out of the computationally maintained datasets of the database, one consisted of PPI with proteins known to have a role in the pre-synapse. This set of PPI was selected for the analysis.

5.2.3.4 String Database

The String Database (Szklarczyk et al. 2015) contains PPI obtained from three different sources: experimentally obtained PPI (drawn from the following databases: BIND, DIP, GRID, HPRD, IntAct, MINT, and PID), computationally predicted PPI and PPI extracted from the literature by a co-occurrence method. Two large sets of interactions were drawn from the database: One containing PPI experimentally obtained and the other of PPI extracted from the literature. Predicted PPI were excluded from this study as PPIs prediction tools are not sufficiently precise: many datasets are highly skewed containing many non-interacting PPIs (Browne et al. 2010).

5.2.4 Interaction analysis

5.2.4.1 Workflow

Figure 5.1 represents the workflow of the interaction analysis performed. Each step will be further detailed in the section below.

5.2.4.2 Gene Selection and PPI sets

Four gene sets were selected using the different sets of PPI previously detailed in this chapter, section 5.2.3.

The first one, SET1, contained genes involved within the PPI obtained from text-mining using the results obtained by the best text mining tool.

Two sets of genes were drawn from the Human Interactome Project: one from the set generated through high throughput screening of PPIs (SET2A) and the set curated from the literature (SET2B). The PPIs from the two groups were combined, forming the SET 2. Using SET 2 set of interactions, it was possible to compare the enrichment for the same interactions in the subset SET 2A and the subset SET 2B. In addition, SET 2A was analysed on his own in order to assess whether enrichment could be observed within that good quality dataset only.

The third set of genes, SET 3, consists of the PPI drawn from the synaptic dataset of the Intact database.

The last set of genes, SET 4, was drawn from the String database, which contains a very high number of interactions: PPI from screening origin (SET 4A) and PPI extracted from the literature (SET 4B). As for the second set of genes, the analysis was carried out after joining the two sets for comparison.

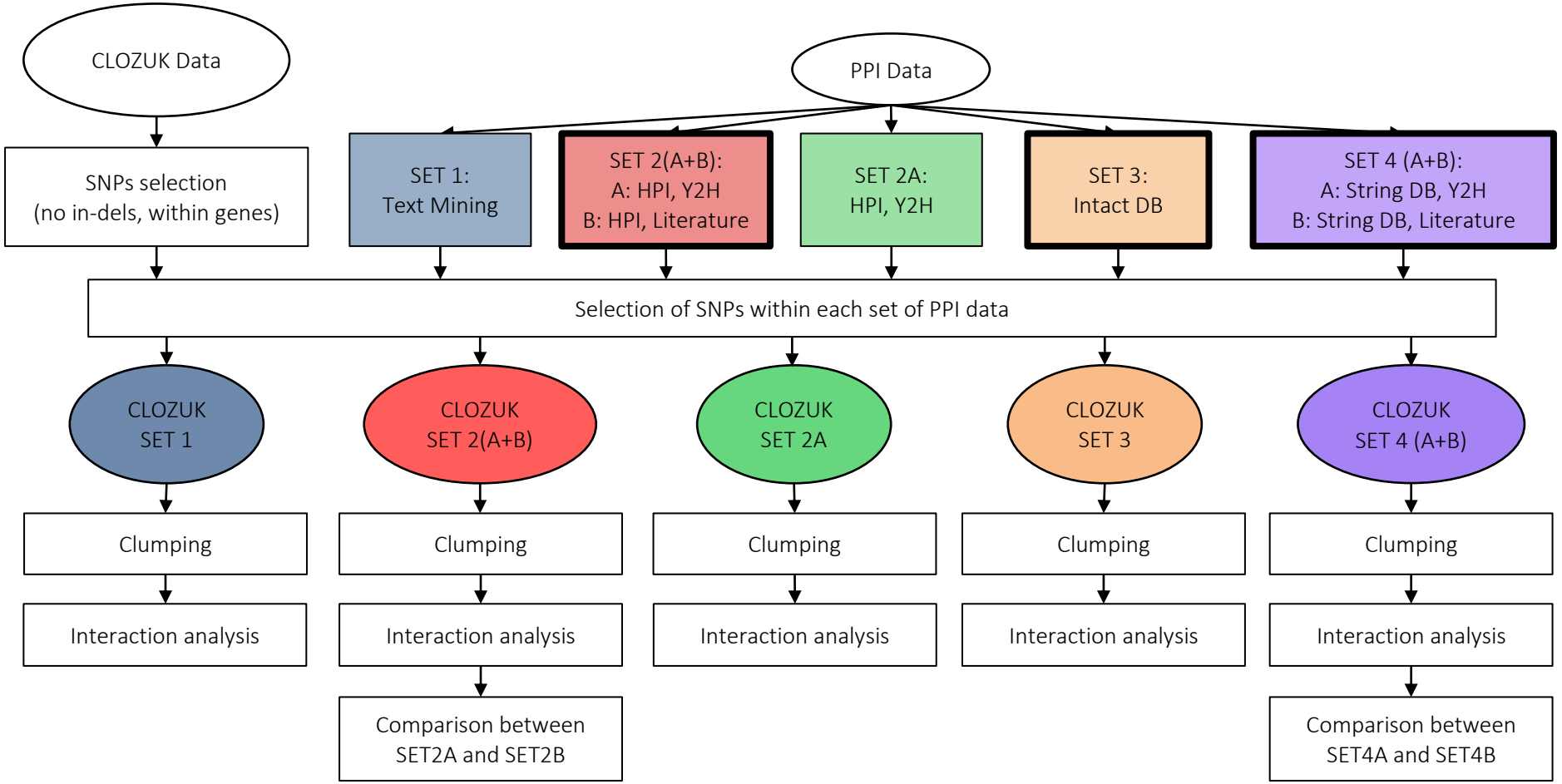


Figure 5.1:Workflow of the analysis. PPI data and CLOZUK data are merged in order to select the genes common to both.

5.2.4.3 SNPs selection and clumping

Firstly, only SNPs inside chosen genes were kept. The analysis focussed on autosomal chromosomes: all variants on X and Y chromosomes were removed as a separate analysis of male and female would have been needed resulting in smaller sample size and loss of power for detection. In addition, all the insertions and deletions were excluded from the analysis. The same clumping procedure described in Chapter 3 and 4 was used to restrict the number of SNPs using linkage-disequilibrium pruning. Only the most highly associated SNPs were kept using the option –clump in PLINK (Purcell et al. 2007). The following parameters were chosen: a window of 2,000 kb and r^2 of 0.1 and p1 threshold of 0.01. When further investigation was needed, different p1 thresholds were tested (p1=0.05 and p1=0.1) in order to increase the number of SNPs used in the analysis. By increasing the number of variants used, it was then possible to investigate how stable the result was.

5.2.4.4 Pair-wise SNP interaction

As explained in detail in section 3.2.3 of Chapter 3 and section 4.2.5 of Chapter 4, interactions between all the possible pairs of clumped SNPs were calculated using PLINK 1.9 (Purcell et al. 2007; Chang et al. 2015). As argued in section 4.2.3.3 of Chapter 4, the results obtained from the different chips were combined by means of a meta-analysis using the software METAL (Willer et al. 2010).

5.2.4.5 Analysis of the results

The analysis of the results relied onto the calculation of every possible pair of interactions between the clumped SNPs within genes previously identified in section 5.2.4.2.

It also relied onto the comparison between the distribution of SNP interaction p-values for gene pairs linked to PPIs and the distribution of SNP interaction p-values for gene-pairs not linked by PPIs within the same set of genes. To assess if any enrichment of statistical

interactions could be detected, a Mann Whitney Wilcoxon rank test was used to compare the interactions in genes linked to PPI with the interactions in genes not involved with PPI. Then, a chi-square test was used to evaluate if there was an excess of genetic interactions with a p-value below two thresholds (p 0.05 and p 0.01) in interactions linked to PPIs compared to interactions not linked by PPIs within the same set of genes. When the count of interactions below the given threshold did not qualify for a chi-square test, Fisher's exact test was used.

5.3 Results

5.3.1 Comparison of the three text-mining tools

The three text-mining tools identified in section 5.2.2: ODIN, PPIInterFinder and @Note were tested against the two corpora described in section 5.2.2.2.

The first corpus tested was the 100 abstracts sampled from the ACT corpus. @Note performed best in term of recall with a score of 0.92. However when looking at the precision, ODIN performed better with a score of 0.672. Taking into account the F-Score (measure that combines precision and recall indicators), ODIN obtains the higher score with a F-measure of 0.754 (Figure 5.2).

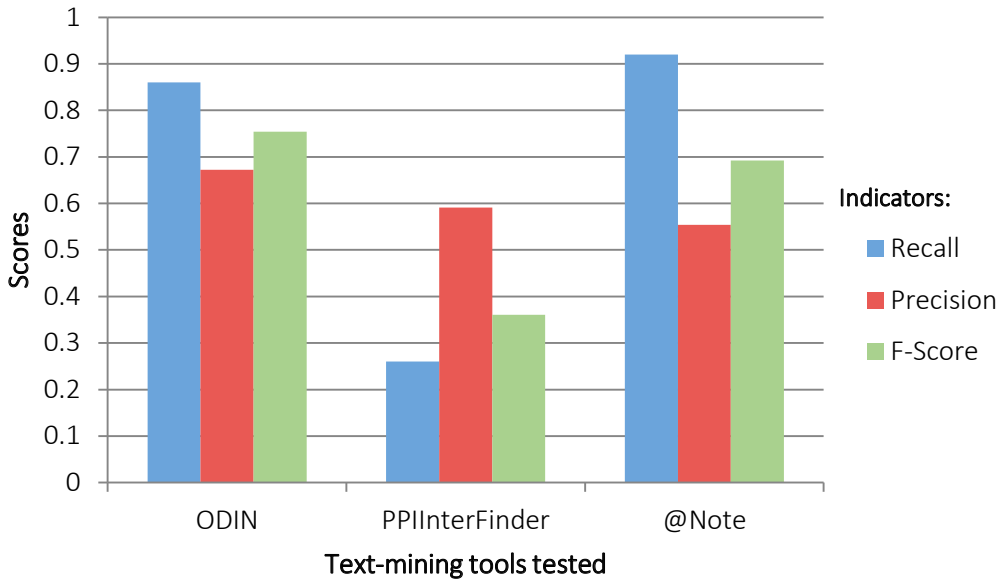


Figure 5.2: Result of the three tests performed (Recall, Precision and F-score) using the three text mining tools (ODIN, PPIInterFinder and @Note) against 100 abstracts randomly selected from the ACT corpus.

The second corpus was a collection of abstract from PubMed, carefully selected using Mesh Terms. As in the previous test, @Note performed the best for the recall with a score of 0.915 but a lower score for precision. PPIInterFinder obtained the best precision score with a score of 0.667. However, ODIN with a F-measure of 0.679 obtains the best overall score (Figure 5.3). The F-measure is a balance between the recall and precision scores: it was chosen as the main indicator of performance for each tool. Having the highest F-measure score, ODIN was identified as the best text-mining tool.

In 2015, ODIN was run across all PubMed in the search for PPI. ODIN ranks the interactions it finds, using a scoring system to assess the likelihood of each interaction being a true PPI (Rinaldi et al. 2012). The calculated score is based on a machine learning process that can select interesting articles based on the previously classified articles (Rinaldi et al. 2012). Only interactions with a confidence score > 100 were kept for further analysis in order to select highly ranked interactions only (Rinaldi et al. 2012).

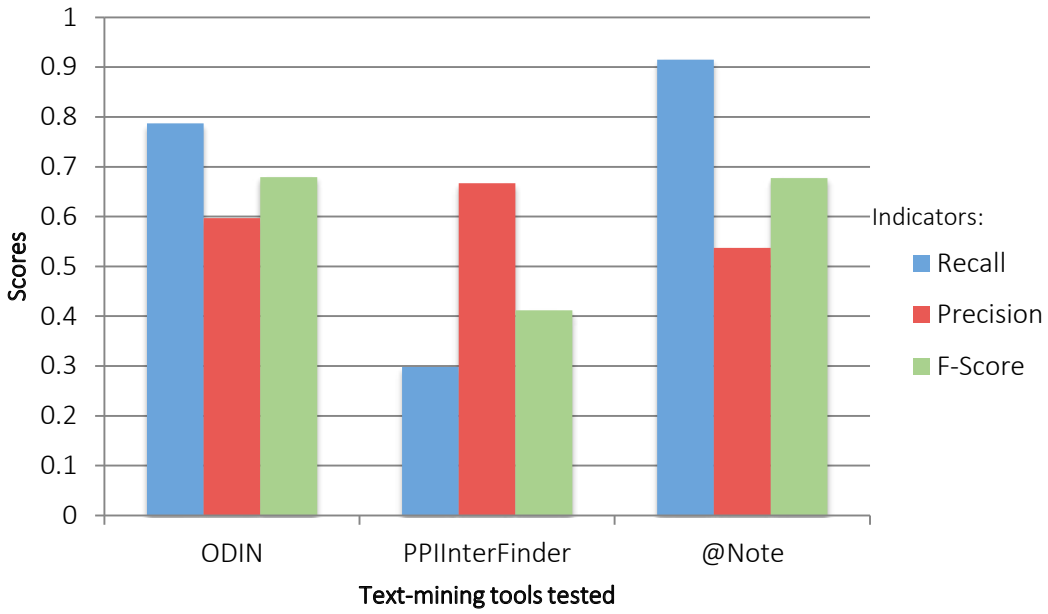


Figure 5.3: Result of the three tests performed (Recall, Precision and F-score) using the three text mining tools (ODIN, PPIInterFinder and @Note) against 100 abstracts selected from PubMed using MeshTerms.

5.3.2 Summary of the different PPI sets used

PPI Set	Description	Number of gene pairs	Number of genes
SET 1	PPI extracted from the literature by the text mining tool ODIN	21,525	4,001
SET 2A	Screen PPI from the Human Interaction Project	15,254	4,480
SET 2B	Literature PPI from the Human Interactome Project	10,183	5,295
SET 3	Synaptic PPI from the Intact Database	3,453	1,720
SET 4A	Screen PPI from the String database	2,511,242	14,392
SET 4B	Literature extracted PPI from the String database	5,744,559	16,137

Table 5.2: Description of the different PPI sets analysed: number of gene pairs and number of genes

Table 5.2 details the number of gene pairs and the number of genes involved in each PPI set or subset. As outlined in section 5.2.4.1 and showed by Figure 5.1, each PPI set was analysed independently. Both SET 2 and SET 4 contain PPI from screening origin as well as PPI extracted from the literature for comparison purpose between two sources of PPI. The

subset SET 2A was also analysed separately in order to assess interactions in a high quality PPI dataset. The same clumping parameters were used for each set.

Table 5.3 describes the different samples with the number of genes in each set as well as the number of SNPs used for the interaction analysis.

PPI Set	Description	Number of Genes	Number of SNPs prior to Clumping	Number of SNPs after Clumping (p1=0.01)
SET 1	PPI extracted from the literature by the text mining tool ODIN	4,001	398,134	865
SET 2	PPI from the Human Interaction Project Subset A: Screen interactions Subset B: extracted from literature	7,879	775,445	1,583
SET 2A	Screen PPI from the Human Interaction Project	4,480	357,087	830
SET 3	Synaptic PPI from the Intact Database	1,720	211,503	452
SET 4	PPI from the String database Subset A: Screen interactions Subset B: extracted from literature	16,192	1,570,015	2,849

Table 5.3: Description of the different PPI sets analysed: number of genes, number of SNPs before and after clumping.

5.3.3 Assessment of the enrichment of statistical interactions within PPI sets

A Mann Whitney Wilcoxon rank test was used to compare the distribution of genetic interaction p-values for gene pairs linked by PPIs with the distribution of p-values for interactions within gene-pairs not linked by PPIs within the same set of genes. This test assessed if there was any enrichment of statistical interactions within gene pairs linked by PPIs.

The results using the clumping parameter p1=0.01 (Figure 5.4) show that enrichment was only detected within the SET 3: the synaptic PPI from the Intact database: p=0.00031 (Bonferroni threshold p=0.0071).

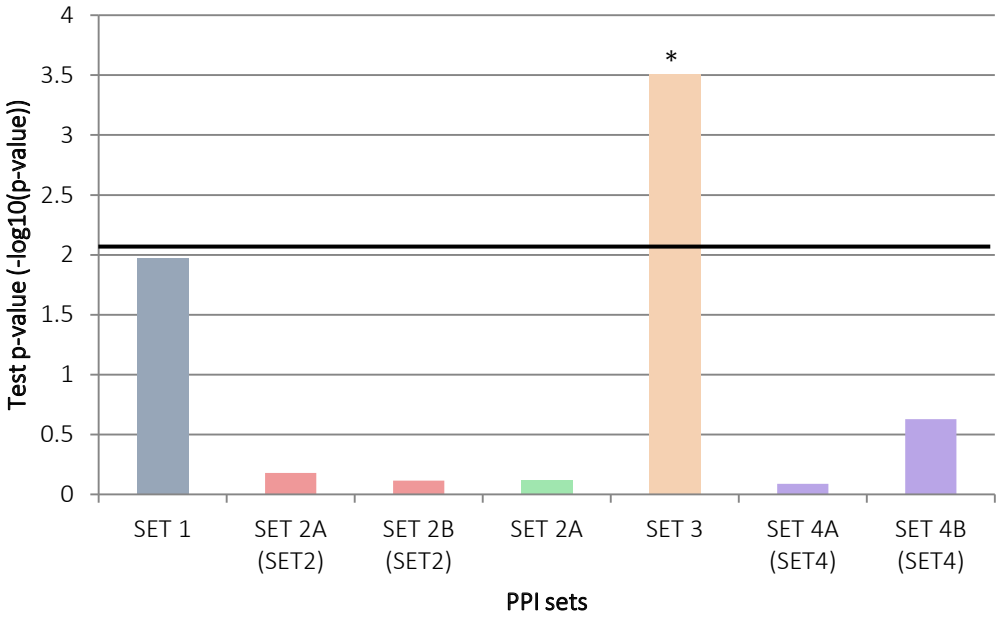
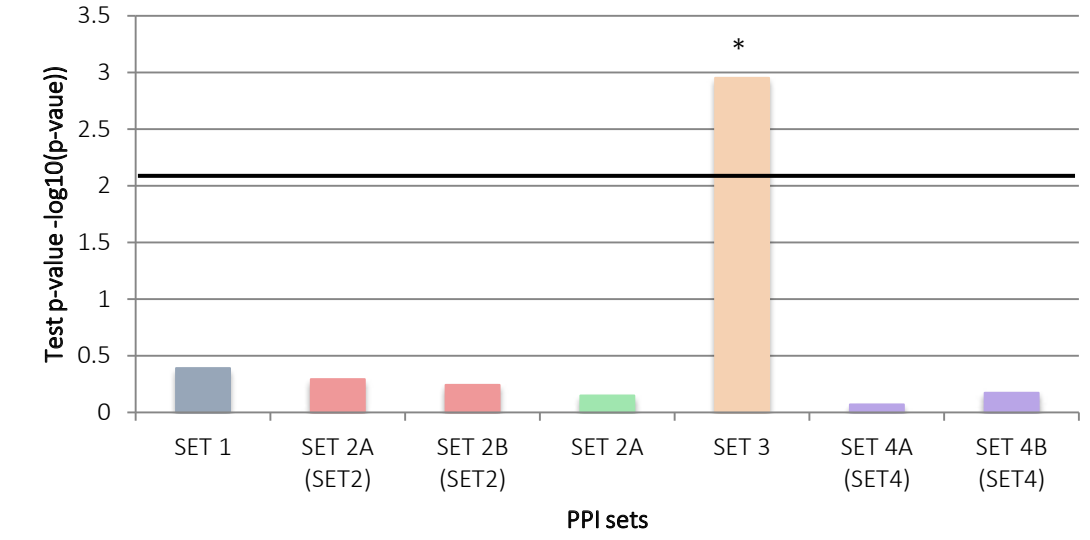


Figure 5.4: Ranking test for the enrichment of statistics interactions with low p-values performed on the different sets of PPI. The colour for each set corresponds to the ones in Figure 1.1. The black bar indicates the Bonferroni threshold (p=0.0071). Only SET 3 shows enrichment.

In addition a chi-square test was used to see if an enrichment of interactions below the thresholds (p 0.05 and p 0.01) could be detected (Figure 5.5). When looking at interactions with p-values under the 0.05 threshold, only the synaptic PPI set (SET 3) shows an enrichment (p=0.0011, Figure 1.5A). However when performing the same test using a different threshold (p 0.01, Figure 1.5B) no enrichment is detected in any SET of PPI.

A



B

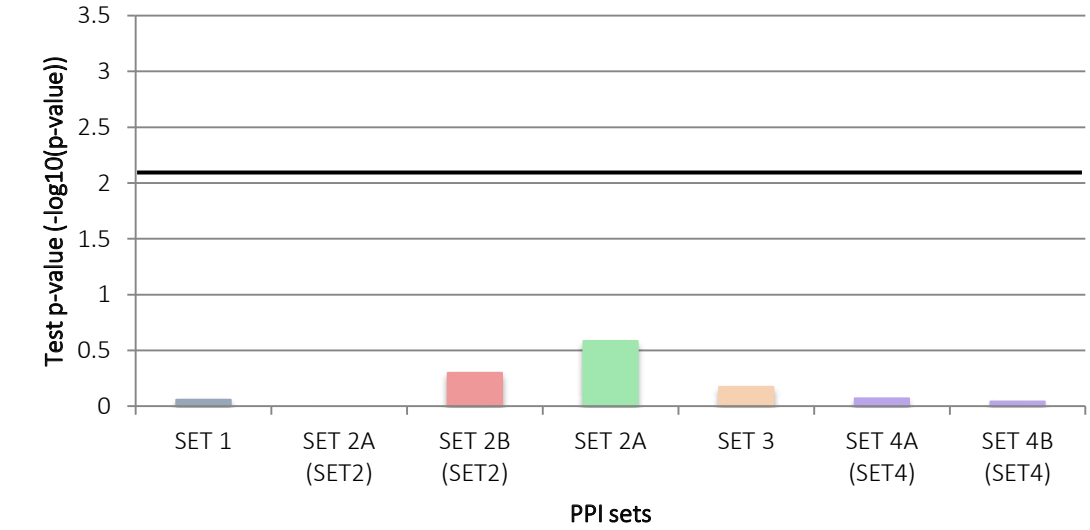


Figure 5.5: Chi-square test results for each set tested when selecting interactions with p-values below 0.05 (A) and below 0.01(B). The colour for each set corresponds to the ones in Figure 1.1. The black bar indicates the Bonferroni threshold (p=0.0071). Only SET 3 shows enrichment when looking at interactions with p-values below 0.05 (A).

Following the detection of enrichment of statistical interactions with low p-values within SET 3, further analysis were performed away. The number of SNPs used in the analysis (452 SNPs) was quite small: the clumping parameters were altered in order to increase this number (p1=0.01 and p1=0.05, Table 5.4) and the same analysis was performed.

	Number of SNPs in SET 3
Prior to clumping	211,503
After Clumping (p1=0.01)	452
After Clumping (p1=0.05)	1,439
After Clumping (p1=0.1)	2,435

Table 5.4: Number of SNPs in SET 3 prior to clumping and after clumping using three different parameters: p1=0.01, p1=0.05 and p1=0.1.

Figure 5.6 shows the results of the Mann Whitney Wilcoxon rank test: the only enrichment detected remains the one observed initially. By increasing the number of SNPs in the analysis, no additional enrichment of interactions is detected in SET 3.

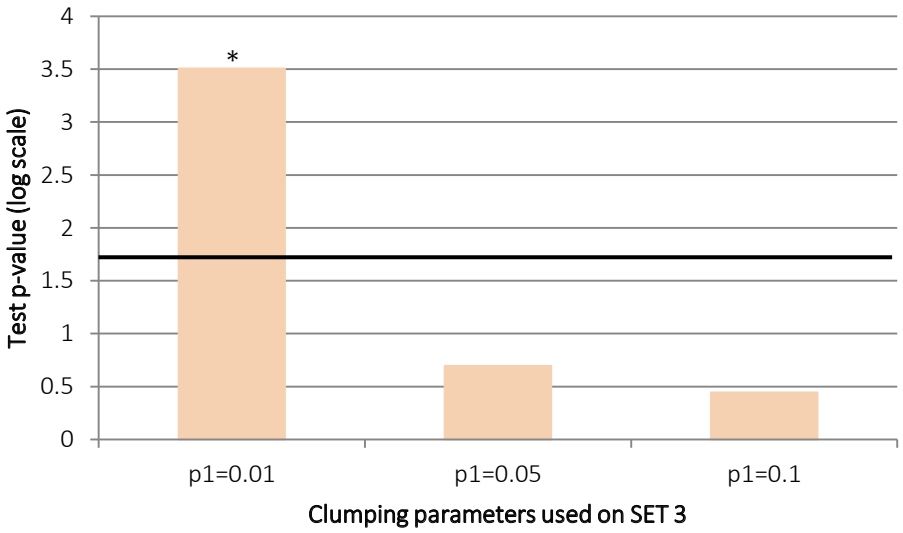


Figure 5.6: Ranking test for the enrichment of statistics interactions with low p-values performed on SET 3 using three different clumping parameters (p1=0.01, p1=0.05 and p1=0.1). . The black bar indicates the Bonferroni threshold (p=0.017). The same enrichment detected is the same as previously observed in Figure 5.4. No additional enrichment is observed by increasing the number of SNPs into the analysis.

5.4 Discussion

5.4.1 Perspective on the extraction of PPI by text-mining tools

Three different tools were compared to assess the feasibility of PPI extraction among a collection of abstracts. Three tools were selected for further comparison: ODIN PPIInterFinder and @Note. The F-score being a balance between the precision and the recall, it was the indicator of choice to compare the tool's performance. Out of the three tools, ODIN was the one that performed the best, achieving a F-score of 0.754 and 0.679 on the two corpora tested. In addition, ODIN was able to process the whole of PubMed, making it a superior candidate for this analysis.

Despite those promising results, PPI detection from abstracts is not extremely precise and reliable: there are still place for improvement for PPI-extraction tools. Perfect PPI detection from text is not yet possible but current tools are likely to have an impact on simplifying the process of PPI article selection (Krallinger et al. 2011). One of the main challenges to be face by PPI-extraction tools is the recognition and extraction of novel interaction from text (Krallinger et al. 2011). Furthermore there is also an interest in being able to distinguish direct and non-direct interaction usually involving contact between more than two proteins (Krallinger et al. 2011). The current development of machine learning techniques such as graph kernel approaches (Airola et al. 2008) or support vector machines methods (Yang et al. 2010) should help to tackle those issues. Machine learning methods have the advantage of deriving information from their training dataset (Krallinger et al. 2011), making it interesting candidates for PPI-extraction tools.

However with the rise of projects such as the Human Interactome Project (Rolland et al. 2014) the potential interaction space is better covered by high throughput screens. As a result, in the future there may well be less need for tools that collect PPI information from small-scale studies.

5.4.2 Perspective on PPI and statistical interactions

To investigate whether PPI could help to identify sets of genes enriched in significant statistical interactions, I selected PPI sets from four different sources. The first subset was generated from ODIN's results: PPI extracted from abstracts of PubMed. Two subsets of PPI were drawn from The Human Interactome Project: one from comprehensive pairwise screens and the second from PPI supported by the literature. From the Intact database a third set was created that specifically contained synaptic PPI data. The final set was formed of PPI from the String database with two subsets: PPI supported by Y2H experiments and PPI extracted from the literature.

Using a ranking test, I tested for an enrichment of significant SNPxSNP interactions within sets of genes linked to PPIs. The PPI set from the Intact database, containing the synaptic PPI was the only significant result that was detected. Using a chi-square test, this enrichment was further investigated in order to assess whether interactions with a p-values below 0.05 and 0.01 were enriched in each set of PPI. In the first case, an enrichment was detected in the synaptic set of PPI (SET 3). However when examining interactions with p-values below 0.01, the effect disappear, probably due to the small number of interactions selected that results in a loss of power.

Furthermore different clumping parameters were used on the synaptic PPI set in order to increase the number of variants used in the analysis. No enrichment was detected when performing the same analysis with a higher number of variants.

The genes involved in this PPI set (SET 3, synaptic dataset from the Intact database) are brain-related genes linked with synaptic functionalities. As schizophrenia is a psychiatric disorder it is very likely that such genes would be relevant for the disease and hence be enriched with significant interactions.

However the analysis compared SNP-SNP interaction p-values for gene pairs linked by PPIs to gene pairs not linked by PPIs within the same set of synaptic genes. As a result the analysis controls for a general enrichment of significant interactions between synaptic genes. The result found here is a potentially interesting finding that would need to be followed-up in a larger sample. A bigger dataset composed of the ISC and CLOZUK datasets combined could have been used to that extend. The sample size issue would have been tackled but I choose not to use it: the number of common variants between the two datasets was low and I would have had to exclude many PPI pairs from my analysis. Imputing the ISC dataset might have solved that issue but it was not available at the time of the analysis.

Furthermore many studies showed a link between schizophrenia and synaptic genes (Glessner et al. 2010; Fromer et al. 2014). For example the gene NRXN1 that encodes a membrane protein involved in the formation of synaptic contacts, has been found to increase the risk in schizophrenia (Kirov, Rujescu, et al. 2009). In addition schizophrenia was linked with *de novo* mutations in the ARC (activity-regulated cytoskeleton-associated protein) and NMRAM (N-methyl-D-aspartate receptor) complexes that are involved in synaptic functions (Kirov et al. 2012). These studies' findings reinforce the interest in the main finding of this chapter: the link between PPI from a synaptic dataset with the interactions analysis on a schizophrenia dataset.

5.4.3 Summary of the chapter

In this chapter, I investigated whether PPI could be used in order to detect significant interactions. After reviewing the literature, I compared three text-mining tools able to extract PPI from abstracts. The best tool was selected to identify a set of PPIs through analysis of PubMed abstracts. In addition, I also used other PPI sets from available databases and high throughput screens in order to assess whether significant SNPxSNP

interactions were enriched within those sets of PPI. Only the set of synaptic PPI from the Intact database presented an enrichment of significant interactions. This result is potentially interesting given that several studies showed a link between schizophrenia and synaptic genes (Fromer et al. 2014) and it needs to be follow-up in a larger sample.

Chapter 6 - Discussion

6.1 Summary and implication of results

6.1.1 Taking into account the population structure in an interaction analysis

Regression-based methods are the most frequently used approach in an interaction analysis and show the effect of predictor variables on the disease. In a case-control analysis, this method consists of testing the interaction term only using a logistic regression model as described in Chapter 1. However, interaction analysis studies do not often account for population structure. In Chapter 3, I explored three different ways to account for population structure in an interaction analysis. The three models were tested on the same GWAS dataset where cases and controls belonged to 8 different sub-populations. This allowed the inclusion of the sub-population information as a covariate in the interaction analysis. When dealing with covariates, the most broadly used method consists of adding the covariates as extra terms into the equation. Following recommendations by Yzerbyt et al. (2004), the second method takes into account the possible effect between the covariates and each marker by adding interaction terms between covariates and markers into the equation. Indeed adding the extra interaction terms between covariates and markers allows to control for the effect those covariates could have on the main effect (Yzerbyt et al. 2004; Keller 2014). Finally in the last tested model, interactions were calculated independently for each sub-populations and a meta-analysis was used to combine the results. The methods produced similar results overall, indicated by a good correlation between them. The first two methods were found to have extremely similar result. Therefore while interactions between markers and covariates are possible, they do not seem to play a big role in practice and the bias mentioned by (Yzerbyt et al. 2004) is not observed in the analysed dataset. To definitely prove this hypothesis a large-scale simulation study is needed. It was interesting to assess the differences between these two

methods and the one using the meta-analysis. The produced results of the meta-analytic method were slightly more divergent from the other two methods but there was no evidence for any systematic differences. Such differences are likely due to increased variance for each individual study when analysing studies separately and results in a loss of power.

In terms of running time and memory efficiency of the different methods, the meta-analytic approach outperformed the other two, making it the easiest and fastest method to carry out interaction analyses for thousands of SNPs.

Further comparisons between the three methods focused on the lower tails of the p-value distributions. The method adding covariates and the one suggested by (Yzerbyt et al. 2004) that adds covariates and the interaction terms between covariates and each marker showed good correlation. However, when comparing those two methods with the meta-analytic approach some differences were observed. In some cases the meta-analytic method was able to identify interactions with reasonably low p-values while the other two methods were not. The opposite was also observed. Upon investigation of the direction of effect, the meta-analytic approach differs most from the other methods when the direction of effect is identical in only four or five out of 8 studies. In addition it is possible that the meta-analytic approach over-estimate the results, which could explain the differences observed. Indeed this method does not include the main effects: only the interaction terms are meta-analysed possibly resulting in the interaction term capturing some of the main SNP effects. Further investigation of this issue should include the application of a multivariate meta-analysis such as the one described in Van Houwelingen et al. (2002), to see if it improves the results. Finally it is also possible that a loss of power of detection is observed in the other two methods. By adding extra terms in the model the number of degree of freedom is increased and can cost precision in the estimation of parameters.

6.1.2 Identifying sets of genes enriches for SNPs-SNPs interactions

6.1.2.1 Approach based on genetic information

In Chapter 4 I aimed to determine whether genetic information could be used to identify set of genes enriched for disease-relevant statistical interactions. In order to investigate this, I assessed interactions in two independent and non-overlapping GWAS datasets. The first analysis was an independent one where the SNPs were specific to each dataset. The second analysis was performed after selecting the SNPs common to both datasets. LD clumping was performed in each dataset to reduce the number of variants in the analysis in order to avoid collinearity problems and diminish the multiple testing burden. All pair-wise SNP interactions were then calculated.

As expected, in every analysis, no interaction survived correction for multiple testing: the sample sizes were not large enough to detect individual statistical interactions. Interactions were ranked using gene-wide significance p-values. Spearman ranking correlation was calculated between the interactions p-values and the gene-wide significance. It showed a small positive relationship between the two in the independent analysis, meaning that there is slightly greater evidence for SNP interactions between genes more highly associated with schizophrenia. This was not observed for the common-SNPs analysis.

I then investigated whether an enrichment for more significant interaction p-values was observed among the subset of genes that were most highly associated with schizophrenia when compared to genes that were least associated.

When performing the independent comparison between the two datasets with variants specifically selected in each one, there was evidence for enrichment of more highly associated interactions amongst genes that are most highly associated with schizophrenia compared to interactions in the least associated genes. The observed effect was strongest for the CLOZUK dataset, which is the bigger dataset. In addition, the effect was the strongest when comparing the ranked distribution of SNP-SNP interactions p-values

between the 1% of SNP-SNP interactions involving genes most highly associated with the disease and the 1% of SNP-SNP interactions involving genes that are least associated with schizophrenia (i.e. after ranking interactions by gene main-effects).

However, when performing the common SNP comparison, no evidence for an enrichment of associated interactions was observed. When ranking by the gene-wide significance of the PGC2 dataset, the same non-enrichment pattern was observed.

Looking at correlation between gene-wide significance p-values ISC and CLOZUK are better correlated with PGC2 than with each other: this would explain why ranking CLOZUK SNP-SNP interactions with PGC2 gene-wide p-values gives more similar results than ranking CLOZUK with ISC gene-wide significance (and vice versa). Also the best observed correlation between CLOZUK and PGC 2 (the two largest datasets R~0.3) was not high: this indicates that the gene-wide significances are quite variable from one samples to another. As a result, the ranking per gene-wide significance might not be extremely stable.

On further investigating interactions with a p-value <0.01 in the independent analysis, it was found that several SNPs acted as hubs in the network formed by genetic interactions between high ranked genes. Potential statistical artefacts were explored (relationship between gene degree and length of gene and between gene degree and number of SNPs). Similar effect is observed when comparing gene degree and number of SNPs per genes. When comparing gene degree and gene length, good correlation was found between the two variables for the network involving the lowest ranked genes (network without the hub pattern). One explanation for this observation could be linked to the largest size of brain genes: as it is possible that brain genes are highly associated with schizophrenia it is unlikely that there are found into the lowest ranking genes network. As a result, smaller genes are probably involved in the lowest ranking genes network thus creating a bias.

Several factors could further explain the observed differences between the two datasets. Cases in the CLOZUK dataset were treatment-resistant which was not true for the ISC

dataset where we would only expect roughly one third to be treatment resistant. This could lead to have slightly different genes associated with each phenotype and perhaps the interactions picked up could be specific to treatment-resistance. Furthermore the population of the two datasets is different: British population in the CLOZUK dataset and a mix of 8 different populations (British, Swedish and Bulgarian) in the ISC dataset. As a result, if interactions are enriched between the most associated genes and if a different set of genes is driving the association signal in each sample (due to phenotypic differences or variability of the population between the two datasets), interactions seen in one dataset might not be seen in the other. In addition, the presence of noise due to the small sample size could also explain the failure to identify enrichment within schizophrenia associated genes. Different genotyping chips were used in the two studies and the CLOZUK dataset uses more recent chips. Perhaps the chip quality influences the detection of true effects and having to correct for multiple chips, the power of detection is lowered.

The results of the cross-disorder analysis within the independent comparison did not provide evidence that the observed effect was driven by confound inherent in GWAS data but unrelated to disease. Furthermore it is interesting to note that interactions within bipolar disorder genes are enriched for significant interactions using the ISC dataset. However this did not replicated in the CLOZUK dataset.

The final step of this chapter was the investigation of interactions with a p-value below 0.01. This consisted of a preliminary analysis of the biological functions shared among such interactions after selecting terms with $p<0.05$ which shows a small degree of enrichment. Some GO terms were shared between CLOZUK and ISC for the common-SNP analysis. However none of the terms identified passed multiple correction testing.

It will need further work in order to see if biological process could help to identify groups of genes between which an enrichment of significant interactions is observed.

6.1.2.2 Approach based on the use of protein-protein interactions as functional information

The aim of Chapter 5 was to investigate if protein-protein interactions could be used to identify sets of genes between which there is an enrichment of significant statistical interactions.

Efforts have been made in order to create exhaustive protein-protein interactions databases. However such databases present some limits. Curating interactions form the many small-scale studies to be found in the literature is difficult as it is time consuming: as a result many interactions data present in the literature do not appear in databases. The quality of data is variable and working out what is of good or bad quality can became an enormous task. In addition databases overlap each other and the overlap shows difference in annotations due to the difference of interpretation by biologists (Mathivanan et al. 2006).

Thorough and comprehensive screens of PPI have emerged such as the Human Interactome Project. These resources are very valuable as they contain good quality PPI however it is also dependent on the coverage of the search space: the sets of proteins between which PPIs have been evaluated. In addition, the literature contains newly discovered PPI found in small studies and its number is also increasing. There is an interest in developing methods that will allow PPI information to be extracted from scientific published articles.

In Chapter 5, I tested three text-mining tools capable of extracting PPI from abstracts. Following a comparison of their performances using standard indicators (precision, recall and F-measure) I was able to obtain a set of PPI extracted from PubMed using the best tool: Odin.

I also selected PPI sets from four other sources. The Human Interactome Project data was classified into two subsets of PPI: one from comprehensive pairwise screens and the

second from PPI extracted from abstracts. A third set was formed by the synaptic set of PPI from the Intact database. The last set was comprised of PPI from the String database, which contains PPI backed up by Y2H experiments and PPI extracted from the literature. Using each set of PPI, I tested whether significant SNP-SNP interactions were enriched within such sets of PPI by using a ranking test. Only the set of PPI from the Intact database, containing synaptic PPI presented an enrichment of interactions with low p-values. However when including more SNPs into the analysis (by using different clumping parameters), this effect disappeared. Also the positive result does not necessary imply that it is specifically physically-interacting brain genes that are enriched for evidence of interactions, as opposed to brain-expressed genes in general. Indeed schizophrenia is a brain disorder so significant results could be explained by the selection of brain-related genes for the analysis: the genes are more likely to be relevant for the disease so perhaps they are also more likely to be enriched with significant interactions (in line with the evidence from previous chapters). This hypothesis could be tested by comparing PPI pairs form the Intact database with randomly selected pairs of genes known to be involved with brain functions. In addition using a bigger dataset is needed in order to fully investigate if the effect observed is real. A bigger dataset composed of the ISC and the CLOZUK datasets combined could have been used to that extent.

6.2 Strengths and limitations

6.2.1 Strengths

In this thesis I assessed statistical interactions within two independent GWAS datasets. The independent analysis in both datasets showed an enrichment of significant interactions between genes most highly associated with schizophrenia when comparing to interactions between least-associated genes. This provides consistent evidence that interactions could contribute to schizophrenia.

In addition, the methodology comparison of the three models gave a better insight for interactions analysis when dealing with covariates. This work could be useful for future analysis using the recommendation made here.

I also performed a thorough analysis on the use of PPI data in order to identify sets of genes between which an enrichment of statistical interactions can be observed. Despite the various origin of the PPI set used (from comprehensive screens, databases or PPI extracted from the literature) no enrichment was detected except in the synaptic set of PPI. Considering many studies have showed a link between schizophrenia and synaptic genes (Glessner et al. 2010; Fromer et al. 2014; Kirov, Rujescu, et al. 2009; Kirov et al. 2012) this is a potentially interesting finding that needs to be follow-up in a bigger sample.

6.2.2 General limitations

One of the major limitations in an interaction analysis is the lack of power to detect statistically significant interactions. Indeed individual interactions contributing to disease risk are likely to have modest effect sizes due to the sample size used (Manolio et al. 2009). Furthermore the high number of tests performed adds a supplementary challenge with regards to Bonferroni multiple testing correction. As the number of interactions tested is high, the threshold is highly conservative and result in a loss of power. Perhaps an appropriate genome-wide significance threshold should be used for pairwise interactions (Musani et al. 2007) such as the one used for the detection of variants in GWAS ($p=5e^{-8}$). On the other side, the power of detection could greatly be improved by using bigger sample sizes.

In this study the chosen SNPs for every interaction analysis were within autosomal genes as the main interest was the focus on gene-gene interactions. In addition LD clumping was use to restrict the number of SNPs in the analysis: analysing all pair-wise interactions without restricting the number of variants is computationally difficult. However it is important to be aware that some information may be missed by the use of stringent

thresholds. Given the lack of computational power to fully explore the search space of SNP-SNP interactions, there is a need for a trade-off between computational capacities and having sufficient number of markers for the analysis.

Regarding the different PPI sets used in Chapter 5, there are also some limitations. PPI available in database such as String or Intact needs to be better evaluated. In addition, while thorough screens contain better quality PPI data, the search space is not completely mapped, which can introduce bias in the analysis. For example, the data from the Human Interactome Project used here covers only 42% of the search space. With the efforts towards the search of the whole proteome, better data will be available and allow us to perhaps detect some effects. In addition, the investigation was only performed on the CLOZUK dataset and needs to be follow-up on a bigger sample. Imputing the ISC dataset, and combining it with the CLOZUK dataset could have resolved that issue but imputing data were not available at the time of the analysis.

6.2.3 Methodology considerations

In Chapter 3, I compared three different methods to take into account covariates in an interaction analysis. The meta-analytic method was used in Chapter 4 and 5 to perform the interaction analysis. This method could be improved: only interaction terms were meta-analysed and perhaps taking into account the direction of the main effect could help to improve the accuracy of the method.

It was argued in Chapter 3 that including covariates in a logistic regression model could result in a loss of power of detection when the disease prevalence is low (Pirinen et al. 2012). As the method used to calculate interactions is based on a logistic regression model, perhaps by adjusting for covariates the power of detection is lowered. However if the covariates are well known confounds (as it is the case in this thesis), it Is necessary to control for their effects and thus to accept the loss of power.

In the method used to calculate interactions in this thesis, a logistic function was used as a link function to model the relationship between the phenotype and the predictors. This choice has important implications: it could be argued that no statistical interaction between genes is epistatic (Clayton 2012). However an inappropriate choice of scale could result in weakened interaction effects impossible to detect (Frånberg et al. 2015). To investigate this, different link functions could be use on the same datasets (Frånberg et al. 2015; Knol and VanderWeele 2012).

6.3 Future work

Regarding the method comparison in Chapter 3, further investigation could be done to develop the potential of the meta-analytic approach. For example, the approach used in this study meta-analysed the interaction terms only. It would be interesting to see if the accuracy of the method can be improved by taking into account the main effects.

The three methods could also be applied to a different dataset in order to see if the correlations between methods replicated. Due to lack of time and computational capacities, the three methods weren't tested on the CLOZUK dataset, which could be used for this follow-up analysis.

In Chapter 4, I presented preliminary results of the analysis of interactions with a low p-value to assess if biological processes were shared among them. However this analysis stays purely descriptive and a thorough investigation needs to be performed in order to evaluate if it is possible to identify groups of genes between which an enrichment of significant interactions is observed. In addition, DAVID (Sherman et al. 2007) does not take into account the gene size or the number of semi-independent SNPs into account. Further work could include an enrichment test that accounts for those differences.

The main issue in current gene-gene interaction studies is the lack of sample size in order to detect significant interactions. It would be interesting to use a simulation study under a

range of plausible disease models in order to better estimate what sample size would be needed to detect interactions in schizophrenia.

Replication is another major challenge in interaction analysis. It would be interesting to use two similar datasets (in term of population and phenotype) and to reproduce the analysis presented in Chapter 4. Using similar but larger datasets would allow better control over the genetic variability and would perhaps help with the replication of the results.

The thesis is focused on two-way gene-gene interactions but three-way interactions could also be investigated.

Furthermore and despite obvious computational issues, it would also be of interest to look at interactions outside genes: other functional regions of the genome could also play a role. Studies have shown that changes in regulatory regions (such as promoters or enhancers regions) influence the expression pattern of genes. For short-range interactions, extending the gene boundaries would allow to capture interactions between short-range enhancers and their genes. However long-range interactions present another challenge due to the looping property of the DNA.

In addition, genes represent only a small portion of the genome and it would be interesting to further investigate the non-coding part of the DNA. As expected, computational issues would rise as the number of variants to include in such analysis would increase drastically. There is a need to prioritize the number of SNPs in the analysis, perhaps using information regarding the 3D structure of the DNA.

The recent development of genome wide chromosome conformation capture (Hi-C) has permitted the study of chromatin interactions within the nucleus. The resulting interaction maps contain information on loop within the structure of the DNA and possible contacts points between enhancers and promoters. This information could perhaps help to narrow down the number of variants to test in an interaction analysis. Furthermore, the resulting interaction maps show that genomes can be divided into large local chromatin domains

termed Topologically Associated Domains (TADs) (Dixon et al. 2012). Within TADs, the genome appears to be organized to favour strong internal chromatin interactions rather than external interactions with neighbouring TADs. It has also been suggested that TADs might help to delineate basic genomic functions such as gene regulation. Further work could assess whether significant statistical interactions are enriched within TADs.

6.4 Conclusion

To conclude, in the first part of this thesis, I analysed different methods to account for population structure in an interaction analysis.

In the second and third part I investigated two different approaches in order to identify sets of genes between which an enrichment of significant interactions can be observed: one based on the genetic information and the second one based on functional information using protein-protein interactions. Using genetic information, the independent analysis of the two GWAS dataset suggested that gene-gene interactions might play a role in schizophrenia. In addition there might be some enrichment for interactions amongst genes most highly associated with schizophrenia. Further work using GWAS dataset with bigger sample size would be needed in order to improve the power to detect interactions. When investigating protein-protein interaction datasets the only evidence for enrichment of significant interactions was observed in the synaptic dataset. This potentially interesting finding needs to be further investigated.

Chapter 7 - References

Ackermann, M. and Beyer, A. 2012. Systematic detection of epistatic interactions based on allele pair frequencies. *PLoS genetics* 8(2), p. e1002463.

Airola, A. et al. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics* 9 Suppl 11, p. S2.

Alako, B.T.F. et al. 2005. CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC bioinformatics* 6(1), p. 51.

ÁLvarez-Castro, J.M. et al. 2012. Modelling of genetic interactions improves prediction of hybrid patterns - A case study in domestic fowl. *Genetics Research* 94(5), pp. 255–266.

Anderson, C.A. et al. 2010. Data quality control in genetic case-control association studies. *Nature protocols* 5, pp. 1564–1573.

Arya, R. et al. 2009. Effects of covariates and interactions on a genome-wide association analysis of rheumatoid arthritis. *BMC proceedings* 3 Suppl 7, p. S84.

Ashburner, M. et al. 2000. Gene Ontology: tool for the unification of biology. *NATURE GENETICS* 25(1), pp. 25–29.

Bader, G.D. et al. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic acids research* 31, pp. 248–250.

Ben-Hur, A. and Noble, W.S. 2005. Kernel methods for predicting protein-protein interactions. *Bioinformatics (Oxford, England)* 21 Suppl 1, pp. i38-46.

Bergen, S.E. and Petryshen, T.L. 2012. Genome-wide association studies of schizophrenia: does bigger lead to better results? *Current opinion in psychiatry* 25(2), pp. 76–82.

Berggård, T. et al. 2007. Methods for the detection and analysis of protein-protein interactions. *Proteomics* 7(16), pp. 2833–2842.

Bland, J.M. and Altman, D.G. 1986. *Statistical methods for assessing agreement between two methods of clinical measurement.*

Blomgren, K.J. et al. 2006. Interviewer variability - Quality aspects in a case-control study. *European Journal of Epidemiology* 21(4), pp. 267–277.

Bloom, J.S. et al. 2013. Finding the sources of missing heritability in a yeast cross. *Nature* 494(7436), pp. 234–237.

Brown, M.B. 1975. A method for combining non-independent, one-sided tests of significance. *Biometrics* 31, pp. 987–992.

Browne, F. et al. 2010. From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions. *Advances in Artificial Intelligence* 2010, pp. 1–15.

Burdick, K.E. et al. 2008. Elucidating the relationship between DISC1, NDEL1 and NDE1 and the risk for schizophrenia: Evidence of epistasis and competitive binding. *Human Molecular Genetics* 17(16), pp. 2462–2473.

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 - Seminars in Medical Genetics* 97(1), pp. 12–17.

Carlborg, O. and Haley, C.S. 2004. Epistasis: too often neglected in complex trait studies? *Nature reviews. Genetics* 5(8), pp. 618–625.

Ceol, A. et al. 2010. MINT, the molecular interaction database: 2009 update. *Nucleic acids research* 38, pp. D532–D539.

Chang, C.C. et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4(1), p. 7.

Chu, M. et al. 2014. A genome-wide gene-gene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese. *Carcinogenesis* 35(3), pp. 572–577.

Clayton, D. 2012. Link functions in multi-locus genetic models: Implications for testing, prediction, and interpretation. *Genetic Epidemiology* 36(4), pp. 409–418.
Combarros, O. et al. 2009. Replication by the Epistasis Project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer's disease. *Journal of neuroinflammation* 6, p. 22.
Comeau, D.C. et al. 2013. BioC: A minimalist approach to interoperability for biomedical text processing. *Database* 2013.
Cordell, H.J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics* 11(20), pp. 2463–2468.
Cordell, H.J. 2009. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics* 10(6), pp. 392–404.
Cusick, M.E. et al. 2009. Literature-curated protein interaction datasets. *Nature methods* 6, pp. 39–46.
Dixon, J.R. et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398), pp. 376–380.
Doherty, J.L. et al. 2012. Recent genomic advances in schizophrenia. *Clinical Genetics* 81(2), pp. 103–109.
Fleuren, W.W.M. and Alkema, W. 2015. Application of text mining in the biomedical domain. *Methods* 74, pp. 97–106.
Fraley, C. and Raftery, A.E. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* 97(458), pp. 611–631.
Frånberg, M. et al. 2015. Discovering Genetic Interactions in Large-Scale Association Studies by Stage-wise Likelihood Ratio Tests. *PLoS Genetics* 11(9).
Fromer, M. et al. 2014. De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506(7487), pp. 179–84.
Gauderman, W.J. 2002. Sample size requirements for association studies of gene-gene interaction. *American Journal of Epidemiology* 155(5), pp. 478–484.
Gibson, G. 2012. Rare and common variants: twenty arguments. *Nature Reviews Genetics* 13(2), pp. 135–145.
Gilbert-Diamond, D. and Moore, J.H. 2011. Analysis of gene-gene interactions. *Current Protocols in Human Genetics* (SUPPL. 70).
Glessner, J.T. et al. 2010. Strong synaptic transmission impact by copy number variations in schizophrenia. *Proceedings of the National Academy of Sciences* 107(23), pp. 10584–10589.
Gregersen, J.W. et al. 2006. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature* 443(7111), pp. 574–577.
Habib, M.S. and Kalita, J. 2010. Scalable biomedical Named Entity Recognition: investigation of a database-supported SVM approach. *International journal of bioinformatics research and applications* 6(2), pp. 191–208.
Hamshere, M.L. et al. 2013. Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Molecular psychiatry* 18(6), pp. 708–12.
Hasan, M.S. habbi. et al. 2015. Performance evaluation of indel calling tools using real short-read data. *Human genomics* 9, p. 20.
He, J. et al. 2011. Gene-based interaction analysis by incorporating external linkage disequilibrium information. *European journal of human genetics : EJHG* 19(2), pp. 164–72.
He, X. et al. 2010. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nature Genetics* 42(3), pp. 272–276.
Hemani, G. et al. 2014. Detection and replication of epistasis influencing transcription in humans. *Nature* 508, pp. 249–53.
Hirschhorn, J.N. and Daly, M.J. 2005. Genome-wide association studies for common

diseases and complex traits. *Nature reviews. Genetics* 6(2), pp. 95–108.
Hirschhorn, J.N. and Daly, M.J. 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6(2), pp. 95–108.
Van Houwelingen, H.C. et al. 2002. Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine* 21(4), pp. 589–624.
Hu, T. et al. 2013. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association : JAMIA* 20(4), pp. 630–6.
International HapMap Consortium 2005. A haplotype map of the human genome. *Nature* 437(7063), p. 1299–320 ST–A haplotype map of the human genome.
International Schizophrenia Consortium 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455(7210), pp. 237–41.
International, T. and Consortium, H. 2003. The International HapMap Project. *Nature* 426(6968), pp. 789–796.
Ito, T. et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98(8), pp. 4569–4574.
Jablensky, A. 2000. Epidemiology of schizophrenia: The global burden of disease and disability. *European Archives of Psychiatry and Clinical Neuroscience* 250(6), pp. 274–285.
Jensen, L.J. et al. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews. Genetics* 7(2), pp. 119–29.
Jeste, D. V. et al. 1985. Did schizophrenia exist before the eighteenth century? *Comprehensive Psychiatry* 26(6), pp. 493–503.
Kaufman, J. et al. 2004. Social supports and serotonin transporter gene moderate depression in maltreated children. *Proceedings of the National Academy of Sciences of the United States of America* 101(49), pp. 17316–17321.
Kavanagh, D.H. et al. 2015. Schizophrenia genetics: Emerging themes for a complex disorder. *Molecular Psychiatry* 20(1), pp. 72–76.
Keller, M.C. 2014. Gene × environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. *Biological Psychiatry* 75(1), pp. 18–24.
Kerrien, S. et al. 2007. IntAct--open source resource for molecular interaction data. *Nucleic acids research* 35, pp. D561–D565.
Kirov, G., Rujescu, D., et al. 2009. Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophrenia Bulletin* 35(5), pp. 851–854.
Kirov, G., Grozeva, D., et al. 2009. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Human Molecular Genetics* 18(8), pp. 1497–1503.
Kirov, G. et al. 2012. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry* 17(2), pp. 142–153.
Klapa, M.I. et al. 2013. Reconstruction of the experimentally supported human protein interactome: what can we learn? *BMC systems biology* 7(1), p. 96.
Klein, R.J. 2007. Power analysis for genome-wide association studies. *BMC genetics* 8(1), p. 58.
Knol, M.J. and VanderWeele, T.J. 2012. Recommendations for presenting analyses of effect modification and interaction. *International Journal of Epidemiology* 41(2), pp. 514–520.
Koh, G.C. et al. 2012. Analyzing protein-protein interaction networks. *J Proteome Res* 11(4), pp. 2014–2031.
Krallinger, M. et al. 2011. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC*

Bioinformatics 12, p. S3.
Lambert, J.C. et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* 45(12), pp. 1452–1458.
Lee, S.H. et al. 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* 44(7), pp. 831–831.
Lichtenstein, P. et al. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *The Lancet* 373(9659), pp. 234–239.
Lincoln, M.R. et al. 2009. Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. *Proceedings of the National Academy of Sciences* 106(18), pp. 7542–7547.
Lippert, C. et al. 2013. An exhaustive epistatic SNP association analysis on expanded wellcome trust data. *Scientific Reports* 3.
Littell, R.C. and Folks, J.L. 1971. Asymptotic optimality of fisher's method of combining independent tests. *Journal of the American Statistical Association* 66(336), pp. 802–806.
Liu, Y. et al. 2011. Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genetics* 7(3).
Lourenço, A. et al. 2009. @Note: A workbench for Biomedical Text Mining. *Journal of Biomedical Informatics* 42(4), pp. 710–720.
Mackay, T.F.C. 2013. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics* 15(1), pp. 22–33.
Mah, S. et al. 2006. Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. *Molecular psychiatry* 11(5), pp. 471–478.
Manolio, T.A. et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265), pp. 747–53.
Mathivanan, S. et al. 2006. An evaluation of human protein-protein interaction data in the public domain. *BMC bioinformatics* 7 Suppl 5, p. S19.
McCarroll, S. a et al. 2006. Common deletion polymorphisms in the human genome. *Nature genetics* 38(1), pp. 86–92.
McCarthy, M.I. et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9(5), pp. 356–369.
McGrath, J. et al. 2004. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC medicine* 2, p. 13.
Millar, J.K. et al. 2005. DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling. *Science (New York, N.Y.)* 310(5751), pp. 1187–1191.
Mishra, G.R. et al. 2006. Human protein reference database--2006 update. *Nucleic acids research* 34, pp. D411–D414.
Moore, J.H. and Williams, S.M. 2005. Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. *BioEssays* 27(6), pp. 637–646.
Morris, A.P. and Zeggini, E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* 34(2), pp. 188–193.
Moskvina, V., Craddock, N., et al. 2011. An examination of single nucleotide polymorphism selection prioritization strategies for tests of gene-gene interaction. *Biological Psychiatry* 70(2), pp. 198–203.
Moskvina, V., O'Dushlaine, C., et al. 2011. Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. *Genetic epidemiology* 35(8), pp. 861–6.
Musani, S.K. et al. 2007. Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum Hered* 63(2), pp. 67–84.

Nalls, M.A. et al. 2014. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genetics* 46(9), pp. 989–993.
Nicodemus, K.K. et al. 2010. Biological validation of increased schizophrenia risk with NRG1, ERBB4, and AKT1 epistasis via functional neuroimaging in healthy controls. *Archives of general psychiatry* 67(10), pp. 991–1001.
Nicodemus, K.K. et al. 2010. Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: Biological validation with functional neuroimaging. *Human Genetics* 127(4), pp. 441–452.
Orchard, S. et al. 2007. Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics* 7 Suppl 1, pp. 28–34.
Orchard, S. et al. 2014. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* 42(D1).
Owen, M.J. et al. 2005. Schizophrenia: genes at last? *Trends in Genetics* 21(9), pp. 518–525.
Pagel, P. et al. 2005. The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)* 21, pp. 832–834.
Pawson, T. and Nash, P. 2000. Protein-protein interactions define specificity in signal transduction. *Genes and Development* 14(9), pp. 1027–1047.
Phillips, P. 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* 9(11), pp. 855–867.
Pirinen, M. et al. 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* 44(8), pp. 848–851.
Prabhu, S. and Pe'er, I. 2012. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Research* 22(11), pp. 2230–2240.
Price, A.L. et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 38(8), pp. 904–9.
Price, A.L. et al. 2008. Long-Range LD Can Confound Genome Scans in Admixed Populations. *American Journal of Human Genetics* 83(1), pp. 132–135.
Psychiatric GWAS Consortium Bipolar Disorder Working Group 2011. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 43(10), pp. 977–983.
Purcell, S. et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81, pp. 559–575.
Purcell, S.M. et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, pp. 748–752.
Purcell, S.M. et al. 2014. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506(7487), pp. 185–90.
Raja, K. et al. 2013. PPInterFinder - A mining tool for extracting causal relations on human proteins from literature. *Database* 2013.
Rebholz-Schuhmann, D. et al. 2012. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics* 13(12), pp. 829–839.
Rees, E. et al. 2014. Analysis of copy number variations at 15 schizophrenia-associated loci. *British Journal of Psychiatry* 204(2), pp. 108–114.
Reich, D.E. and Lander, E.S. 2001. On the allelic spectrum of human disease. *Trends in Genetics* 17(9), pp. 502–510.
Rinaldi, F. et al. 2012. Using ODIN for a PharmGKB revalidation experiment. *Database* 2012.
Rinaldi, F. et al. 2014. OntoGene web services for biomedical text mining. *BMC bioinformatics* 15 Suppl 1, p. S6.
Risch, N. and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)* 273(5281), pp. 1516–1517.
Rolland, T. et al. 2014. A proteome-scale map of the human interactome network. *Cell*

159(5), pp. 1212–1226.
Rössler, W. et al. 2005. Size of burden of schizophrenia and psychotic disorders. *European Neuropsychopharmacology* 15(4), pp. 399–409.
Rual, J.-F. et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437(7062), pp. 1173–1178.
Saha, S. et al. 2005. A systematic review of the prevalence of schizophrenia. *PLoS medicine* 2(5), p. e141.
Saha, S. et al. 2006. Incidence of schizophrenia does not vary with economic status of the country: Evidence from a systematic review. *Social Psychiatry and Psychiatric Epidemiology* 41(5), pp. 338–340.
Salwinski, L. et al. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic acids research* 32, pp. D449–D451.
Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510), pp. 421–427.
Sebat, J. et al. 2007. Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)* 316(5823), pp. 445–9.
Sham, P.C. and Purcell, S.M. 2014. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* 15(5), pp. 335–346.
Sherman, B.T. et al. 2007. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC bioinformatics* 8, p. 426.
Shi, J. et al. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 460(7256), pp. 753–7.
Stark, C. et al. 2011. The BioGRID Interaction Database: 2011 update. *Nucleic acids research* 39, pp. D698–D704.
Stefansson, H. et al. 2009. Common variants conferring risk of schizophrenia. *Nature* 460(7256), pp. 744–7.
Sullivan, P.F. et al. 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry* 60, pp. 1187–1192.
Szklarczyk, D. et al. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43(D1), pp. D447–D452.
Tandon, R. et al. 2008. Schizophrenia, 'just the facts' what we know in 2008. 2. Epidemiology and etiology. *Schizophrenia research* 102(1–3), pp. 1–18.
Tandon, R. et al. 2009. Schizophrenia, 'just the facts' 4. Clinical features and conceptualization. *Schizophrenia Research* 110(1–3), pp. 1–23.
Turner, S. et al. 2011. Quality Control Procedures for Genome-Wide Association Studies. In: *Current Protocols in Human Genetics*.
Turner, S.D. et al. 2011. Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PloS one* 6(5), p. e19586.
Ueki, M. and Cordell, H.J. 2012. Improved statistics for genome-wide interaction analysis. *PLoS genetics* 8(4), p. e1002625.
Venkatesan, K. et al. 2009. An empirical framework for binary interactome mapping. *Nature methods* 6(1), pp. 83–90.
Walsh, T. et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, N.Y.)* 320(5875), pp. 539–43.
Wan, X. et al. 2010. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics* 87(3), pp. 325–340.
Wang, S. and Zhao, H. 2003. Sample Size Needed to Detect Gene-Gene Interactions using

Association Designs. *American Journal of Epidemiology* 158(9), pp. 899–914.
Wei, W.-H. et al. 2012. Genome-wide analysis of epistasis in body mass index using multiple human populations. *European Journal of Human Genetics* 20(8), pp. 857–862.
Wei, W.-H. et al. 2014. Detecting epistasis in human complex traits. *Nature Reviews Genetics* 15(11), pp. 722–733.
Wellcome, T. et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145), pp. 661–78.
Willer, C.J. et al. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* 26, pp. 2190–2191.
Willer, C.J. et al. 2010. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, pp. 2190–2191.
Williams, H.J. et al. 2011. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Molecular psychiatry* 16(4), pp. 429–41.
Williams, N.M. et al. 2010. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: A genome-wide analysis. *The Lancet* 376(9750), pp. 1401–1408.
Yang, Z. et al. 2010. BioPPISVMExtractor: A protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. *Journal of Biomedical Informatics* 43(1), pp. 88–96.
Yeger-Lotem, E. and Sharan, R. 2015. Human protein interaction networks across tissues and diseases. *Frontiers in Genetics* 6(Aug).
Yi, N. et al. 2011. Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. *Annals of human genetics* 75(1), pp. 90–104.
Yu, H. et al. 2011. Next-generation sequencing to generate interactome datasets. *Nature methods* 8(6), pp. 478–480.
Yzerbyt, V.Y. et al. 2004. Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology* 40, pp. 424–431.
Zhang, J. et al. 2004. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics* 37(6), pp. 411–422.
Zhang, L. et al. 2012. Modeling haplotype-haplotype interactions in case-control genetic association studies. *Frontiers in Genetics* 3(January), pp. 1–17.
Zhang, Y. and Liu, J.S. 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* 39(9), pp. 1167–1173.
Zuk, O. et al. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* 109, pp. 1193–1198.