Gene expression data and neuropsychiatric disease

Alexander Richards

A thesis submitted to Cardiff University for the degree of

Doctor of Philosophy

Department of Psychological Medicine

School of Medicine

Cardiff University

September 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration and statements

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .....

Date .....

Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Signed .....

Date .....

Statement 2

This thesis is the result of my own independent investigations, except where otherwise stated. References are given where other sources are acknowledged. A bibliography is appended.

Signed .....

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....

Date .....

Summary

Microarrays are a powerful technology, enabling the determination of mRNA levels for tens of thousands of genes from a single sample. However, the resulting gene expression datasets are often difficult to interpret due to their size and complexity. The overall aim of this study is to evaluate a diverse selection of methods for the analysis of large-scale gene expression data derived from human brain, and to apply them to furthering the understanding of heritable psychiatric disorders. One strand of research presented here focuses on using clustering algorithms to group genes according to their expression. Several methods for expression clustering were implemented and used upon brain expression datasets. The technique of Gene Ontology enrichment was then used to assess the concordance of the resulting clusters with current biological knowledge. The results suggest that combining different clustering methods is the most effective strategy, as it allows the discovery of the widest range of clusters. Clusters produced by these methods were then investigated for enrichment with genes associated with, or differentially expressed in, bipolar disorder or schizophrenia. Particularly enriched clusters were further studied using the functional annotation database MetaCore. The second strand of this research focused on using control adult brain expression data and expression quantitative trait analysis to divide SNPs into those with a greater and lesser effect on global gene expression. This classification was used to enhance the prediction of schizophrenia affected status from genome-wide association study SNP data using polygenic score analysis, a method which aggregates information from a large number of loci. It was found that SNPs which have a larger effect on global gene expression are significantly superior at predicting schizophrenia affected status through polygenic score analysis, a novel finding which suggests that expression data from control adult brain can have relevance to the study of schizophrenia.

Acknowledgments

My sincere thanks go to my supervisors, Dr Lesley Jones, Professor Peter Holmans and Professor Michael O'Donovan for all their help and support over the past four years.

Thanks also go to the other members of the Biostatistics and Bioinformatics Unit, all of whom have provided invaluable advice at various points during my PhD. Additional thanks go to Dr Seth Dobrin for the use of his data before it was publically available.

Finally, special thanks go to my wife Kirsty and son Alfie, for all the patience, support and love they have given me during my studies.

This work was funded by the Medical Research Council.

Table of contents

List of figures

List of tables

Abbreviations used in thesis

BP - Bipolar disorder

CDF - Chip definition file

cDNA - Copy DNA

CEU - Caucasian European

CNS - Central nervous system

CNV - Copy number variation

Con - Control

CRC - Chinese restaurant clustering

cRNA - Copy RNA

DIANA - Divisive analysis clustering

DNA - Deoxyribonucleic acid

DSM-IV - Diagnostic and statistical manual of mental disorders, 4th edition

ENO2 - Enolase 2

eQTL - Expression quantitative trait locus

FDR - False discovery rate

GABA - gamma-Aminobutyric acid

GCRMA - Guanine-cytosine robust multi-array average

GEO - Gene expression omnibus

GIFtS - GeneCard inferred functionality scores

GIMM - Gaussian infinite mixture model

GO - Gene ontology

GSEA - Gene set enrichment analysis

GTOM - Generalised topological overlap matrix

GWAS - Genome-wide association study

ISA - Iterative signature algorithm

ISC - International Schizophrenia Consortium

KEGG - Kyoto encyclopedia of genes and genomes

LD - Linkage disequilibrium

MAF - Minor allele frequency

MAPK - Map kinase

MAS5 - Microarray suite

MC66 - McLean 66

MCLUST - Model-based clustering

MEA - Modular enrichment analysis

memISA - Memory iterative signature algorithm

MeSH - Medical Subject Headings

MGS - Molecular Genetics of Schizophrenia

miRNA - MicroRNA

MM - Mismatch

mRNA - Messenger ribonucleic acid

PANTHER - Protein Analysis Through Evolutionary Relationships

PB - Perrone-Bizzozero

PCR - Polymerase chain reaction

PD - Parkinson's disease

PISA - Progressive iterative signature algorithm

PM - Perfect match

QC - Quality control

RMA - Robust multi-array average

RNA - Ribonucleic acid

rRNA - Ribosomal RNA

SAMBA - Statistical algorithmic method for bicluster analysis

SCZ - Schizophrenia

SEA - Single enrichment analysis

SNP - Single nucleotide polymorphism

snRNA - Small nuclear ribonucleic acid

SVDMAN - Singular value decomposition microarray analysis

tRNA - Transfer RNA

WTCCC - Wellcome Trust case control consortium

YRI - Yoruban, Ibidan

Chapter One

Introduction

1.1 Summary

This chapter gives an overview of the process by which large expression datasets are created, the current state of the techniques used to analyse them, and their application to the understanding of psychiatric disorders. Firstly, some background information on gene expression and the role of nucleic acids in the cell is discussed (Section 1.2), and the overall aim of the study set out (Section 1.3). A brief description of the history of microarray technology is given (Section 1.4) and the practical steps involved in their use are described (Section 1.5). The data processing and quality control steps used to transform the image of the microarray into a usable spreadsheet of expression data are described (Section 1.6). Unique features of the two most commonly used microarray platforms, Affymetrix and Illumina, are detailed (Section 1.7), and genotyping arrays and genome-wide association studies are also described (Section 1.8). In Section 1.9, the utility of expression data in understanding human disease in general, and psychiatric disease in particular, is discussed. The limitations of some types of expression data analysis, and the need to find new analytical methods, are also examined here. Finally, in Section 1.10, a brief overview of the approaches examined in the later chapters of this study is given.

1.2 Nucleic acids and gene expression

Nucleic acids are molecules that perform a variety of important functions within the cell. Deoxyribonucleic acid (DNA) is used as a storage medium, containing the information the cell machinery needs to survive and perform its functions in the organism, and is also the means by which an organism passes on characteristics to its offspring. DNA is well suited to long-term storage because it is a relatively stable molecule.

RNA is transcribed from sections of DNA called genes. Some of this, referred to as messenger RNA or mRNA, is transported from the nucleus to other locations in the cell, where it is translated into proteins, which fulfil the vast majority of structural and catalytic functions the organism requires. mRNA effectively acts as a short-term storage medium and transporter of information around the cell.

This process is supported by other types of RNA (1). Some species of small nuclear RNA (snRNA) are important in the process of splicing, where the non-protein-encoding segments of an mRNA molecule (introns) are removed. snRNAs can also play a role in the regulation of transcription. Other RNA types play a part in the translation from mRNA to protein. Ribosomal RNA (rRNA) forms the major part of the ribosome, the structure where translation occurs, while transfer RNA (tRNA) facilitates the attachment of the correct amino acid to the growing protein chain at the ribosome. MicroRNA (miRNA) molecules can bind to the 3' untranslated region of an mRNA, blocking its translation.

This whole process by which a gene can affect the behaviour of a cell, and so the whole organism, is termed 'gene expression'. It is heavily regulated at every stage of the process, especially transcription, which can require complex interactions between large numbers of DNA sequence elements and proteins referred to as 'transcription factors'. One way in which the level of expression of a gene in a tissue can be estimated is by quantifying the abundance of mRNA molecules produced by each gene. Although there are several levels of regulation before and after the production of mRNA, protein and mRNA abundance do generally correlate positively (2), and some level of mRNA production is required for the corresponding protein to be present. Hence, biological inferences can be made from mRNA abundance, although the presence of these subsequent stages of regulation must be borne in mind while doing so.

DNA microarrays are one of the most powerful ways of evaluating gene expression, as they allow estimation of mRNA abundance for a high proportion of the genes expressed in a biological sample. This study focuses on ways of using these gene expression data to improve the understanding of psychiatric disorders.

1.3 Aim of study

The overall aim of this study is to evaluate a diverse selection of methods for the analysis of large-scale gene expression data derived from human brain, and to apply them to furthering the understanding of heritable psychiatric disorders. Brain expression datasets are used both alone and in concert with genome-wide association study (GWAS) data for this purpose.

1.4 History of microarrays

The origins of the microarray begin with a 1975 paper by Southern *et al* describing the Southern blot (3, 4). In this process a target DNA molecule possibly containing a sequence of interest is cut with endonucleases and denatured into single strands with a reagent such as sodium hydroxide. The strands are separated with electrophoresis, and bound to a nylon or nitrocellulose sheet.

The sheet is then washed with a single stranded radiolabelled probe DNA, complementary to the sequence of interest. If the sequence of interest is present in the target DNA, the probe will preferentially hybridise to the target DNA fragment, so the radiolabel will remain present on the DNA even after washing. This principle was extended to use other target molecules, such as RNA (northern blotting) and protein (western blotting) (5, 6). Western blots are based upon antibodies specific to the target protein, rather than hybridisation.

A decade later, some research groups began to expand these techniques to use multiple probes bound to a nylon sheet or glass slide (3). This led to the first microarrays produced using photolithography (7). This is a fabrication technique similar to that used to etch circuitry into computer chips. The silicon is covered with a masking agent which is removed in specific areas by the application of ultraviolet light. The appropriate oligonucleotide base for that probe can then be bound to the unmasked areas. The chip is then covered in masking agent again, and the process is repeated, gradually building up the sequence of the probes.

Microarrays were rapidly turned to a number of uses, particularly determination of mRNA expression and SNP genotyping (3). Expression data derived from Affymetrix and Illumina mRNA microarrays are the primary focus of this study, but genotyping microarray data are also discussed in Chapter 3.

Early microarray studies suffered from a number of drawbacks. They were often interested purely in identifying differentially expressed genes between affected and control states. Many also used only basic metrics such as fold change (i.e. the ratio of expression in controls to expression in cases) to define 'differential expression', and did not attempt to correct for the considerable multiple testing burden created by testing thousands of gene transcripts (8).

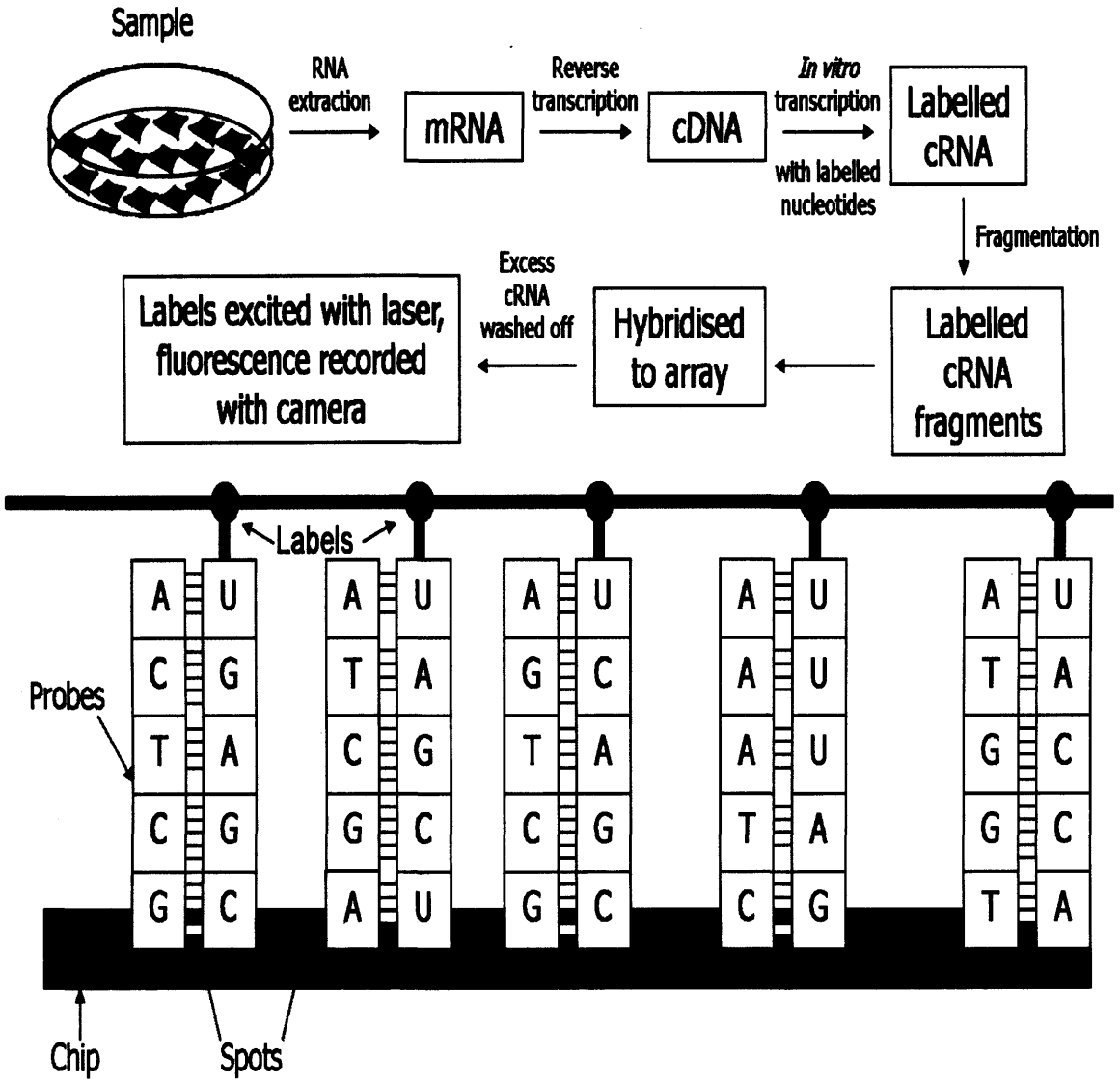
This issue was compounded by the small sample sizes in early studies, which could be as small as two samples from a single donor under different conditions (9). Although sample sizes have increased as the microarray technology has matured and chip prices have fallen, the typical sample size in brain gene expression studies targeting psychiatric disorders are not as large as for genome-wide association studies. This is because of the necessity of taking brain samples from recently deceased subjects, which limits the number of samples that are available.

Microarray chips generally have numerous probes in a grid of spots. The spots are also referred to as 'features'. All probe molecules within a single feature will have the same sequence. Many features across the microarray contain probes with the same sequence, to give multiple readings from that probe; these are said to belong to the same 'probeset'. Distinct probesets can target different parts of the same transcript. In some cases these form an extra layer of replication, while in others different probesets hybridise to alternatively spliced forms of a gene transcript.

In the following years, the number of features present on a microarray chip increased rapidly, while their cost gradually reduced (3). In addition to Affymetrix, other companies also produced microarrays, including Agilent and Illumina. These products are similar to Affymetrix microarrays, but with some architectural differences, such as Illumina applying their probes to tiny beads attached to the surface of the chip, rather than the flat surface of the chip itself.

Other types of microarray and glass slide spotted array have also been created. Antibody microarrays are used to measure protein levels in a fashion analogous to the western blot (10). As the biological importance of microRNAs (miRNA) has become apparent, microarrays designed to detect their expression levels have also been produced (11).

Figure 1.1. Practical steps needed to measure mRNA expression of a sample using a microarray



Note that the probes are represented as 5 nucleotides long for convenience – in reality they would be 25 nucleotides long on Affymetrix microarrays, and 50 nucleotides long on Illumina microarrays.

1.5 Method of use

The laboratory techniques used to quantify mRNA expression using microarrays begin with the sample of interest (Figure 1.1). Initially, RNA is extracted from the sample using standard molecular biology techniques (12). These can be divided into methods that dissolve the RNA in a solution, and column-based methods which purify the mRNA by binding the polyadenine tails of mRNA molecules to silica beads with attached polythymine nucleotide tracts.

The mRNA is reverse transcribed, producing cDNA. This is then transcribed, using nucleotides labelled with biotin, producing labelled cRNA molecules with the same sequence as the original mRNA. Generally, the biotin labels are attached to uracil bases in the cRNA, although alternative protocols do exist where other nucleotides are labelled (13). The cRNA is then fragmented, to create cRNA fragments with similar lengths, and also to prevent cRNA secondary structure interfering with the hybridisation process (14).

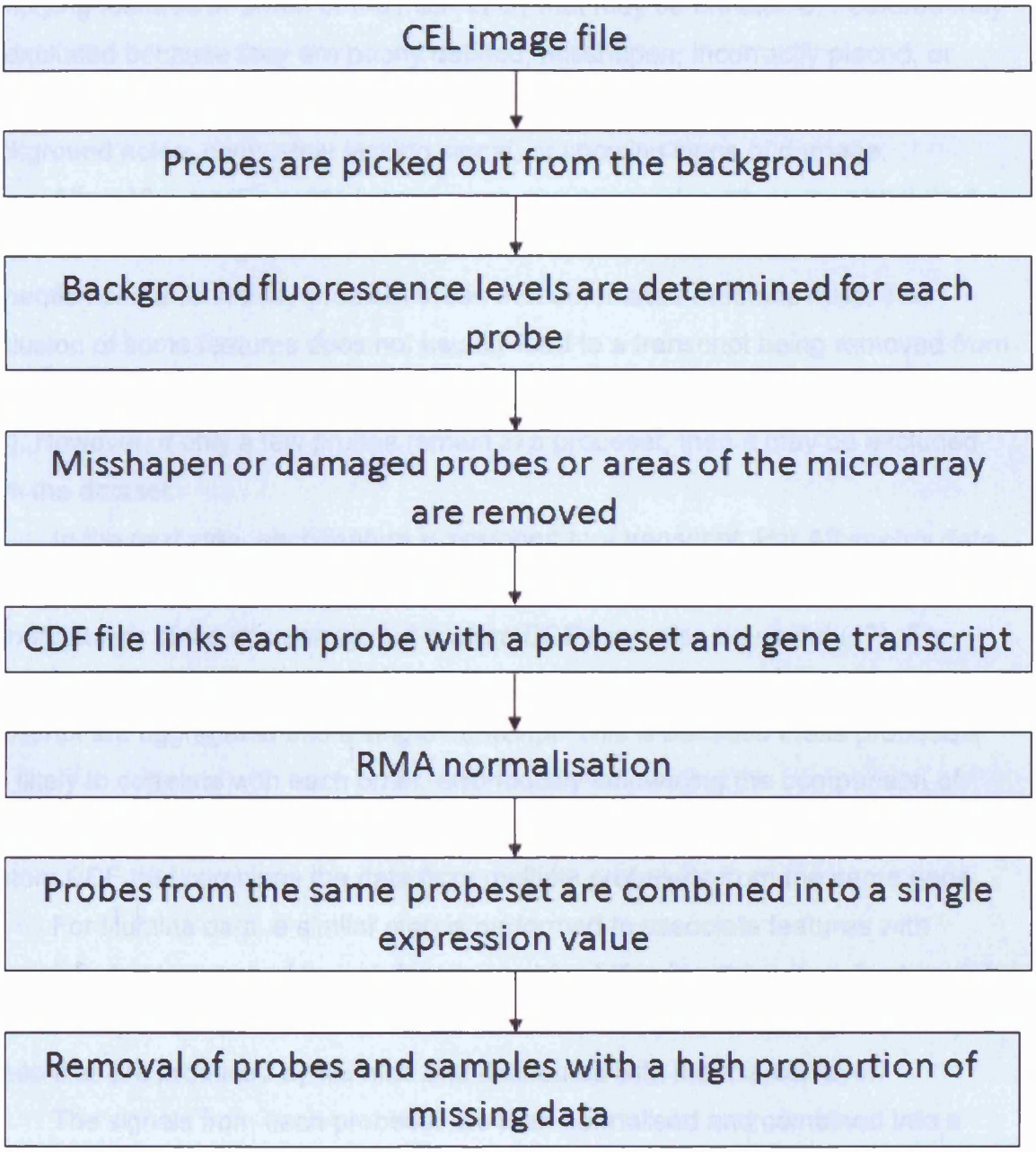
The cRNA is then applied to the surface of the microarray under tightly controlled conditions, particularly temperature. Once hybridisation has occurred between the target cRNA and the probes on the array, unhybridised cRNA is washed off using a series of salt buffers. The array is then stained with avidin, to which is attached a fluorophore such as Cy3. The avidin has an extremely high affinity for the biotin labels and binds to them, labelling the nucleotide fragments hybridised to the array with fluorophore.

The microarray is then read using a scanner, which uses a laser to excite the fluorophores attached to the cRNA and a camera to read the fluorescence produced. This produces an image file of the chip, where the fluorescence at each probe is proportional to the concentration of mRNA complementary to that probe in the original sample.

1.6 Data pre-processing

Several steps are necessary to transform this image file into a spreadsheet of data suitable for further analysis (Figure 1.2). For Affymetrix and Illumina microarrays, the first step is to process the image. The microarray features containing the probes are picked out from the darker background. A sample of the background noise level

Figure 1.2. Steps needed to transform microarray image files into a spreadsheet of data



surrounding each feature is then taken, and this value subtracted from the signal for that feature.

Quality control steps are then performed. These involve automatically identifying features or areas of the microarray that may be unreliable. Features may be excluded because they are poorly defined, misshapen, incorrectly placed, or overlapping another feature. Entire areas may be excluded because of high background noise, completely lacking signal, or showing signs of damage.

Alternative quality control metrics can also be used, such as 'harshlighting', which identifies damaged areas of the microarray (15). Non-automated visual inspection can also identify problem areas that automated methods miss. The exclusion of some features does not usually lead to a transcript being removed from the final dataset, as each transcript has multiple identical probes which map to it (16). However, if only a few probes remain in a probeset, then it may be excluded from the dataset.

In the next step, each feature is assigned to a transcript. For Affymetrix data, this requires a 'chip definition file' (CDF). CDFs can be provided by the manufacturers of the microarray, but custom CDFs can also be useful (17). For example, in Chapter 2, it is important that multiple probesets which map to the same transcript are aggregated into a single transcript. This is because these probesets are likely to correlate with each other, erroneously influencing the comparison of clustering methods (see Chapter 2, section 2.2.1). This problem is solved by using a custom CDF that combines the data from multiple probesets from the same gene.

For Illumina data, a similar step is performed to associate features with transcripts, but because of the random assembly of Illumina chips (see Section 1.7 below), a separate definition file, similar to a CDF, is needed for each microarray. These files are produced by Illumina and distributed with the microarrays.

The signals from each probeset are then normalised and combined into a single numerical value. Several different methods can be used for this. RMA (Robust Multi-array Average) is one of the most common methods (18), and is used in this study. First, the fluorescence value of each probe is log transformed, and the probes are ranked by their fluorescence value. RMA then performs quantile normalisation. The highest probe fluorescence value on each chip is replaced with the mean of the highest values across all chips. This procedure is then repeated for the next-highest

probe value on each chip, then the next-highest, and so on until all probes have been replaced. Tukey's median polish is then used to estimate the expression of each probe on each chip while excluding effects that are specific to particular chips or particular probes (19).

Other normalisation methods also exist, such as MAS5 (see Section 1.7 below) and GCRMA (16, 20). This is a variant of RMA which uses the guanine-cytosine content of probes to inform the estimation of transcript expression levels.

Once the data are in the form of a spreadsheet giving the signal intensity for each transcript in each sample, further quality control is performed. Samples and transcripts with a high proportion of missing data are excluded from further analysis, as this suggests they may have technical problems. It can also suggest that mRNA expression was at too low a level to be reliably detected by the microarray. Boxplots can also be constructed for the non-normalised expression data in each sample, and those with a considerably different mean expression level to the others can be excluded. The dataset is then ready for further analysis.

1.7 Affymetrix and Illumina microarray platforms

The majority of the information presented here applies equally to both Affymetrix and Illumina microarrays. However, there are some major differences between these array platforms. Ilumina chips have a slightly different design to the Affymetrix chips. Instead of bonding the probes directly to the chip substrate, the probes are bound to silica beads placed within wells in the surface of the chip. This allows somewhat denser packing of features, which permits Illumina chips to contain many more technical replicates of a probe for a given transcript – an average of 30 probes with the same sequence for Illumina arrays, compared to between 11 and 20 for Affymetrix U133 chips. However, smaller feature size can affect probe-level variance to some extent (21), slightly offsetting the advantage of more replicates.

Illumina oligonucleotide probes are 50 nucleotides long, compared to 25 nucleotides for Affymetrix. There are advantages and disadvantages to both choices of probe length. The longer Illumina probes are less likely to randomly match a non-target sequence, but can be more vulnerable to other kinds of non-specific binding (22). Illumina probes are more sensitive to the presence of their target sequence than Affymetrix probes, because of the increase in binding energy in the

hybridisation between a longer probe and target sequence. However, because Illumina probes are longer, they are also more likely to be affected by SNPs occurring in their target sequence, which will weaken the hybridisation between probe and target and reduce the apparent expression value. This can also potentially result in an artefactual correlation between SNP genotype and expression.

The biggest difference, however, is the way in which Illumina produces and determines the location of the probes (23). Where Affymetrix chips have a defined layout with the location of each probe defined in advance of the production of a chip, beads with different probes are applied to Illumina chips at random. The identities of the probes attached to each bead are discovered by using a secondary 'decoder' oligonucleotide sequence attached to the probes. A set of fluorescently labelled oligonucleotides, which will bind to some but not all decoder sequences (and not to the probe sequences themselves) is hybridised to the array, and the fluorescence they give off recorded. Generally, two separate sequences of oligonucleotide are bound to the array in a single hybridisation, each with a different fluorescent dye attached. The oligonucleotides are then dehybridised, and the process is repeated with two different labelled oligonucleotide sequences. The process is iterated several times, and the binding status of each decoder sequence recorded. This builds up a unique pattern of decoder hybridisation for each bead, which uniquely identifies the probe sequence. This decoding step is performed for each chip by the manufacturers, who distribute the resulting map of the chip as a data file.

Older Affymetrix microarrays, including the U133 platform which the work in chapter 2 is based upon, contain two types of probe – 'perfect match' (PM) and 'mismatch' (MM) probes. Perfect match probes are exactly complementary to their target sequence, while mismatch probes contain a single non-complementary base at a central nucleotide (20). Originally, MM probes were intended to allow correction for non-specific RNA hybridisation when using the MAS5 normalisation method. More recently it has been discovered that the relationship between PM and MM fluorescence values is not linear (24, 25). Hence, subtracting MM signal from PM signal can increase the variance of the final probe intensity, especially where PM signal is low, and so MM probe correction is not now commonly used.

Newer Affymetrix microarrays, such as the HuGene and HuExon platforms, omit MM probes, replacing them with a series of probes with varying GC (guanine or

cytosine) content, some of which match sequences in the human genome, and some of which do not (21). This is intended to allow correction for non-specific mRNA binding to the chip. GC content is an important variable for the hybridisation reaction, as link between guanine and cytosine contains three hydrogen bonds, and so is stronger than the link between adenine and thymine or uracil, which contains only two hydrogen bonds.

1.8 Genome-wide association study data

In addition to brain gene expression data, this study also uses genome-wide association study (GWAS) data, also derived using microarray-based methods. These are data which show the alleles present at a large number of common SNPs across the whole genome, often over a million SNPs in thousands or tens of thousands of samples (26). These function in a similar way to expression microarrays, but in place of the probes for specific transcripts, probes which are complementary to the sequence surrounding particular SNPs are attached to the array (27). Instead of mRNA, genomic DNA fragments are amplified using the polymerase chain reaction (PCR), labelled and hybridised to the microarray, which is scanned as described above. Software can then identify the alleles present at each SNP in the sample.

The sequence of the probes attached to the microarray differs depending on the array platform used. In earlier array-based methods, separate probes specific to the sequence containing each allele were used (27). In more recent methods, such as the Illumina GoldenGate protocol, oligonucleotide fragments containing different alleles are labelled with fluorophores producing different wavelengths of light during the PCR stage of the process (28). The probes on the microarray are designed to hybridise to sequences containing either allele present at a SNP. Hence, light from one of the two fluorophores will be produced by a feature where the sample is homozygous for the corresponding SNP, while light from both fluorophores will be produced where the sample is heterozygous for the corresponding SNP. This allows a single probe to determine the genotype of a SNP, rather than requiring a probe for each allele.

By examining the alleles present at a large number of SNPs in control populations and populations affected by a condition, it is possible to identify SNPs

where one allele is more common in affected samples and the other is more common in control samples. These SNPs are said to be associated with the condition. Such SNPs may play a role in the aetiology of the condition themselves, or they may be in linkage disequilibrium (LD) with another genomic feature that does.

Two SNPs are in LD if their genotypes at each locus are correlated. This arises because they are close together on the genome, and so there are relatively few locations where recombination events can occur during meiosis between them. As these recombination events are necessary to allow SNPs on the same chromosome to be transmitted independently during reproduction, SNPs that are close together on the genome are more likely to be transmitted together, leading to correlation of their genotypes. The presence of associated SNPs within or near to a gene suggests that the gene may be related to the condition studied.

GWAS studies have a clear advantage over brain expression studies in their ability to use genomic DNA from any body tissue. Sample availability is therefore much less limited than expression studies, allowing much larger sample sizes (e.g. 3322 cases and 3587 controls in the schizophrenia GWAS performed by the International Schizophrenia Consortium) (29).

For some conditions, such as type II diabetes and macular degeneration, GWAS studies have identified several genes as being related to the aetiology of the disease, and replicated these associations in other studies (30-32). However, in the study of psychiatric disorders, such as schizophrenia and bipolar disorder, only a small number of loci have been reproducibly associated with affected status (see Section 1.9 below) (33-36). These loci explain only a small proportion of the risk for these disorders. This relative lack of strongly implicated loci is likely to be due to the small effect sizes of common SNPs in these disorders (37). In this study, one of the aims is to use expression data and GWAS data together, to investigate whether combining them could improve understanding of the biological mechanisms underlying neuropsychiatric disorders.

1.9 Expression data and human health

Large-scale expression data produced using microarrays has helped researchers to understand many human diseases. This is particularly true in the field of cancer

research, where the possibility of using microarrays themselves as diagnostic tools has also been explored. Studies examining the power of microarray data derived from extracted tumours to predict the prognosis of cancer patients have found that using microarrays led to improvements in predicting the course of the disease and the chance of survival (38). Novel subtypes of many cancers have also been defined based upon microarray expression data (39, 40).

The primary topic of investigation in this thesis is the extent to which expression data can aid in the understanding of psychiatric disorders. Compared with other organs, the brain offers unique challenges to study. Brain tissue samples are more difficult to acquire than tissue from some other organs, as they can usually only be taken *post mortem*. This reduces the sample sizes that can be achieved, limiting the power of any analysis. Sampling after death also produces other confounders which can affect the quality of expression data, including agonal state, post mortem interval, and brain pH (41).

The brain also has particularly complex expression patterns, with at least 58% of human genes expressed to some extent (42). The variety of cell types present in the brain, broadly divisible into neurons, astrocytes and oligodendrocytes, but including many types of each, further exacerbates this complexity of expression. This can make demonstrating the existence of these expression patterns more difficult, and discoveries harder to replicate in independent datasets. Recent data suggest some psychiatric disorders are also highly polygenic (e.g. schizophrenia, bipolar disorder), although some others are not (e.g. familial early onset Alzheimer's disease, caused by mutations in PSEN1, PSEN2 and APP) (43). Highly polygenic disorders are less amenable to study through genome-wide association studies, as the effect sizes of individual truly associated variants are small and difficult to distinguish from random variation (29).

This study focuses on bipolar disorder and schizophrenia, although the methods discussed could also be applied to other disorders. The term 'bipolar disorder' is used to describe a spectrum of mood disorders that are characterised by cycles of mania and depression (44). There are several subtypes of the disorder, including bipolar disorder I, bipolar disorder II, where mania is replaced by a less pronounced state referred to as hypomania, and cyclothymia, where cycling occurs between a relatively mild mood elevation and a less severe depressive state. Bipolar

disorder can also be subdivided by the frequency of the cycles, including rapid cycling (more than four episodes annually) and ultra-rapid cycling (more than four episodes monthly) subtypes. Bipolar has a heritability of 80% according to twin studies (45).

Schizophrenia is a mental disorder characterised by psychotic symptoms such as hallucinations and delusions, irrational and disorganised speech and thought patterns, and reduced emotional responsiveness (46). The DSM-IV recognises several diagnostic categories of the disorder, including paranoid schizophrenia, which includes auditory hallucinations and paranoid delusions, disorganised schizophrenia, where irrational thinking and blunted emotional affect are prominent, and undifferentiated schizophrenia, a more general category. It is highly heritable (around 80% heritability), which implies that genetic factors play a major role in its aetiology (47). Onset generally occurs during late adolescence or early adulthood, although it can occur earlier or later (48).

In addition to the small effect sizes of at least the common risk alleles which makes gene identification challenging, many additional factors make these disorders difficult to study. Very little, if anything is known with certainty about the pathophysiological mechanisms underlying either disorder and therefore while there are large numbers of hypotheses, these are mainly quite vague, involving concepts like neurodevelopment or synaptic function (49, 50). As a result, studies that target specific genes on the basis of those hypotheses are not highly likely to be successful. In fact this lack of success of candidate gene studies is a feature of genetic studies of most common disease, even those like diabetes which at least at a superficial level, more was known in advance about biological mechanisms. As this thesis is focused on the application of gene expression studies to further investigation of disorders in a generic way and aims to apply those methods in a manner that is atheoretical in terms of disease mechanisms (other than the disorders involving genes and the brain) I do not discuss the detail of any of those hypotheses.

Both disorders are defined by highly variable symptoms. In other words the diagnostic categories span groups of individuals, pairs of which may have very little in common in terms of clinical features. The courses of the disorders are highly variable, sometimes only involving a single episode, while in other cases becoming a chronic condition that affects the sufferer for the rest of their life (36). Furthermore,

there is symptomatic overlap between bipolar disorder and schizophrenia. Bipolar patients can suffer from delusions and disorganised thinking while many schizophrenia patients have prominent symptoms of mood disorder. There is also an intermediate category, schizoaffective disorder, can be used where the symptoms of both disorders occur in the same individual. For neither disorder are their confirmatory diagnostic tests of biological validity (e.g. blood tests) and it is therefore uncertain to what extent the diagnostic groups are simply descriptive terms covering a range of unrelated disorders or are coherent syndromes with related underlying causes across cases. The effect of this is that at the level of biological validity, it is difficult to define sets of patients that one can be certain are likely to have related disorders, which in effect, adds noise to any analyses, genetic or otherwise, that attempts to find differences between patient groups and unaffected people. Adding to the difficulties, it also appears from recent epidemiology and from molecular genetic studies that there is a considerable degree of overlap in the genetic risk of the disorders (36), which strongly point to lack of validity of the diagnostic boundaries, at least with respect to genetic aetiology.

Despite these issues, current GWAS analyses and meta-analyses have found some polymorphisms that are associated with bipolar disorder or schizophrenia at a genome-wide significant level. In the case of schizophrenia, these include two independent sites in the major histocompatibility complex (51). They also include a SNP within an intron of the transcription factor TCF4 and near to the neuronal gene neurogranin. For bipolar disorder, genome-wide significant SNPs were found within the gene PBRM1 and near to the gene ANK3 (52, 53).

Other SNPs are highly significantly associated with both schizophrenia and bipolar disorder. These include SNPs near the genes ZNF804A and CACNA1C (53, 54). CACNA1C is a calcium ion channel sub-unit, but little is known about the function of ZNF804A. These significant trans-disorder associations further suggest that there may be common mechanisms in the aetiology of schizophrenia and bipolar disorder.

Another type of data that has produced some results for schizophrenia are copy number variant studies. These are studies which identify deletions and duplications of areas of the genome. Some rare deletions have been shown to dramatically increase the risk of schizophrenia, such as a microdeletion within

chromosome 22q11, which causes velocardiofacial syndrome. Generally, there are also an excess of rare CNVs in schizophrenia cases compared to controls, especially large deletions (55), suggesting rare CNVs may be aetiologically relevant to schizophrenia.

These findings demonstrate that large-scale genetic methods can produce results of relevance to the pathophysiology of neuropsychiatric disease. However, these results only explain a tiny fraction of the variation in risk. The challenges involved in understanding psychiatric disorders are considerable, and so progress may depend upon investigation and utilisation of different types of data sources (56). Expression data potentially offers such a source of data. For example, genes which both contain associated SNPs and are differentially expressed between cases and controls may be more likely to be true positives, and so further study can be focused on them (see Section 1.8 below). The analyses performed in this study are intended to illuminate some of the ways in which expression data can be used to investigate psychiatric disease.

1.10 Overview of later chapters

Chapter 2 is a comparison of four different microarray expression data clustering methods when used upon human brain expression data – k-means clustering, Chinese Restaurant Clustering (CRC), the Iterative Signature Algorithm (ISA) and the Memory Iterative Signature Algorithm (memISA). The primary metric used to compare the methods is the percentage of clusters produced that are significantly enriched for one or more Gene Ontology (GO) 'biological process' categories. This assesses the degree to which the methods agree with current biological knowledge.

Chapter 3 investigates the utility of combining expression data and GWAS data to further improve schizophrenia affected status prediction through polygenic score analysis. The purpose of this is to discover whether or not SNPs which affect gene expression are significantly enriched for SNPs associated for schizophrenia. A significant enrichment would demonstrate that expression data is relevant to schizophrenia aetiology.

Chapter 4 investigates the function of genes in selected expression clusters produced using the clustering methods described in chapter 2. The clusters are subdivided into subclusters using k-means clustering. The commercial functional

annotation database and software package MetaCore is used to determine how heavily enriched the clusters and subclusters are for GO and MetaCore functional categories. The enrichment of clusters and subclusters for genes associated with, or differentially expressed in, schizophrenia and bipolar disorder is also determined, using the program EASE.

Chapter Two

A comparison of four clustering methods for brain expression microarray data

2.1 Introduction

2.1.1 Background

Clustering genes according to their expression profiles is an important step in interpreting data from microarray studies. Clustering can help summarise datasets, reducing tens of thousands of genes to a much smaller number of clusters. It can aid understanding of systemic effects; looking for a small change in expression between disease states across many genes in a cluster could be a better strategy for finding the causes of complex, polygenic disorders than looking for large changes in single genes (57). Clustering can also help predict gene function, as coexpressed genes are more likely to have similar functions than non-coexpressed genes (58).

There are many clustering methods for microarray expression data currently available (59). However, there are few comparisons of these methods, making it hard for researchers to make a rational choice between them. The majority of papers comparing multiple clustering methods use simulated data or data from simple organisms such as bacteria and yeast (60-62), which may limit the applicability of their findings to data from more complex sources such as human tissues which express more genes. Thus, to investigate human disease, it would be useful to test the methods upon expression data derived from complex human tissues, among which brain tissue is particularly complex since it expresses a higher proportion of the genome transcribed than other tissues (63, 64). Thalamuthu *et al* (65) have previously looked at a wide range of datasets, including some human expression datasets. However, since they restricted their analysis to functionally defined subsets of genes, that analysis did not fully reflect the complexity of human expression, particularly for disorders where there is insufficient knowledge of their aetiology to focus on specific subsets of genes.

2.1.2 Clustering methods selected for comparison

Four methods were examined: k-means clustering(66), Chinese Restaurant Clustering (CRC)(67), the Iterative Signature Algorithm (ISA)(68, 69) and a new, progressive variant of ISA called memISA. These were chosen after a literature survey of the available methods (Table 2.1). All four are unsupervised methods that derive the clusters from the input data, rather than supervised methods which classify genes into user-specified clusters.

Many of the available comparative clustering studies focus exclusively on older methods (61, 70), or restrict their analysis to a single class of clustering methods (60, 62). In this study, the methods were chosen on the basis of variety. ISA and memISA are examples of biclustering methods, CRC is a mixture model based method, while k-means clustering is a simple, well-understood algorithm. They were reported as performing well by their authors and/or other studies (60, 61).

One of the weaknesses of ISA is that it does not use already-found clusters to inform further clustering. It can find a given strong cluster hundreds of times before finding a weak one. An attempt has been made to mitigate this – PISA, the progressive signature algorithm, which works by requiring the sample set of each successive cluster to be orthogonal to (i.e. as different as possible from) all the sample sets of all previous clusters(71). This allows the method to find weaker clusters obscured by stronger clusters with which they share genes. However, no implementation is yet available for this.

In order to investigate the effect of a progressive clustering strategy, memISA was created. It is a modified version of ISA, which weights against genes that are already members of a cluster (see Section 2.2.9 below for more details). The methods were also chosen partly on the basis of novelty. Apart from k-means clustering, they are too recent to have been included in many previous surveys of clustering methods, and so are particularly in need of testing.

Table 2.1 – Survey of clustering methods

Name	Abbreviation	Refs.	Ease of use	Apparent computation time	Reason for inclusion or rejection
Iterative Signature Algorithm	ISA	(60, 69)	4	High	Included, because of the five methods examined in Prelic *et al*, ISA performed well on both the simulated and real datasets.
Progressive Iterative Signature Algorithm	PISA	(71)	N/A	High	Rejected because implementation was unavailable. However, memISA was based upon PISA, to allow a progressive variant of ISA to be investigated.
CLUSTER	CLUSTER	(61, 72)	1	Low	k-means clustering in this package was included, as it was found to be effective in Riva *et al*. Self organising maps were rejected, as, unlike k-means, they restrict the clusters to a two-dimensional map.
Divisive Analysis Clustering	DIANA	(73)	1	Low	Rejected, as hierarchical methods do not offer a simple and objective way to separate the tree into multiple clusters.
Model-based Clustering	MCLUST	(74)	2	Medium	Rejected, as GIMM and CRC are infinite mixture model methods that automatically find the correct number of clusters (see Medvedovic and Sivaganesan).
Gaussian Infinite Mixture Model	GIMM	(75)	2	Very high on large datasets	Investigated, but rejected because of extremely high computation time on large datasets.

Chinese restaurant clustering	CRC	(67)	2	Medium	Included, because it allows an infinite mixture modelling approach to be applied to large datasets.
Singular Value Decomposition Microarray Analysis	SVDMAN	(68, 76)	2	Medium	Rejected, as SVD is sensitive to the noise found in microarray datasets (see Bergmann *et al*).
Gene Shaving	Gene Shaving	(68, 77)	2	Low	Rejected, as it is based upon SVD and so may suffer from the same noise sensitivity (see Bergmann *et al*).
Generalised Topological Overlap Matrix	GTOM	(78)	2	Medium	Rejected, as network based methods are too different to unsupervised clustering to be directly comparable.
Statistical Algorithmic Method For Bicluster Analysis	SAMBA	(60, 79)	2	Medium	Rejected because ISA outperformed it in terms of GO enrichment on real yeast datasets (see Prelic *et al*).
cMonkey	cMonkey	(80)	5	Very high	Investigated, but rejected because of extremely high computation time on large datasets.

Table of the methods considered for comparison. Refs. = references. 'Ease of use' is a subjective assessment of how simple the method is to use (1=most simple, 5=most complex).

2.1.3 Rejected clustering methods

Several other methods were considered for inclusion in this comparison, but subsequently rejected. A self-organising map is a clustering method similar to k-means, but it arbitrarily restricts the clusters to a two-dimensional plane (72). Hierarchical methods, such as DIANA, were rejected, because they do not offer a simple way of dividing the tree they produce into clusters (73). They also assume that the clusters are hierarchically organised, which may not be true.

Two other modelling methods were examined in addition to CRC – MCLUST and GIMM (74, 75). MCLUST was rejected because, unlike CRC and GIMM, it does not automatically choose the best number of clusters from the data. GIMM was initially investigated, but its runtime scales quadratically with number of input genes (complexity of $O(n^2)$), so its use was impractical. The runtime of CRC scales with the number of input genes times the logarithm of the number of input genes (complexity of $O(n \log n)$), so it was chosen instead.

Two methods based on singular value decomposition were also considered, SVDMAN and gene shaving (76, 77). However, it has been found that these methods perform poorly on large, noisy datasets like microarray data, so they were rejected (68). A network-based method, GTOM, was examined, but was felt to be too different to CRC, ISA and k-means for a direct comparison, as it depended on having a network of relationships between the genes already constructed (78). Another biclustering method, SAMBA, was rejected because ISA outperformed it (60, 79).

Lastly, a complex biclustering method called cMonkey was investigated. This integrated data from a number of sources, including *cis* regulatory elements, protein-protein interactions, and expression data, to group similar genes together (80). It was rejected because it was too computationally intensive to be used on a large human brain expression dataset.

2.1.4 Comparison of clustering methods

The performance of the four clustering methods (CRC, k-means, ISA and memISA) was compared by examining the results for biologically meaningful clustering by looking for gene ontology (GO) enrichments within the resulting clusters. The methods were also compared on the percentage of genes from the dataset that were assigned a cluster ('gene coverage') and computation time.

2.2 Methods

2.2.1 Datasets

Three datasets were used for testing, the Dobrin (81), McLean 66 (82) (MC66) and Perrone-Bizzozero (PB - GEO dataset GSE4036) (83) datasets (Table 2.2). They

were downloaded in CEL format from the Stanley Medical Research Online Genomics database(81), the Harvard National Brain Databank database(82) and GEO(84), respectively. They were then processed using R(85), with custom CDF files to map the probes to genes(86). Box plots were used to examine the quality of the data, and several outlier samples (defined as an average expression across all genes 10% lower or higher than the mean for all samples) were removed. Three versions of each dataset were produced. One was normalised by the RMA median polish method, for use in CRC and k-means(87). The other two were normalised to produce a gene-normalised and sample-normalised dataset for running ISA(68).

Creating the gene-normalised dataset for ISA entailed finding the square root of the sum of squares of each sample ($\sqrt{\sum x_i^2}$), and dividing every value for that sample by the result. The mean was then found for each gene, and deducted from every value for that gene. Lastly, the square root of the sum of squares of each gene ($\sqrt{\sum y_j^2}$) was found, and every value for that gene divided by the result. To create the sample-normalised dataset, the procedure was followed in reverse – finding and dividing by $\sqrt{\sum y_j^2}$ for each gene, deducting the mean from every sample, and finding and dividing by $\sqrt{\sum x_i^2}$ for each sample.

2.2.2 Gene coverage

Gene coverage, the percentage of genes on the chip that are put into at least one cluster, was assessed for the cluster set produced by each method.

Table 2.2 – Characteristics of datasets used to test clustering methods

	Pre quality control number of samples			Post quality control number of samples			Tissue	Chip	Number of genes
	Con	SCZ	BP	Con	SCZ	BP			
Dobrin	25	26	27	20	22	22	Brodmann Area 46	Affymetrix 133 plus 2.0	20292
McLean 66	27	18	19	27	15	19	Dorsolateral Prefrontal Cortex	Affymetrix 133A	12757
Perrone-Bizzozero cerebellum	14	14	0	14	14	0	Cerebellum	Affymetrix 133 plus 2.0	20292

Con=control, SCZ=schizophrenia, BP=bipolar disorder.

2.2.3 Speed

The methods were also assessed by speed. As ISA and memISA are dependent on parallelisation to run at a reasonable speed, this is taken as real-world time taken to run, rather than computer run-time used. For k-means and penalised k-means, this includes the time taken to estimate k.

2.2.4 GO enrichment

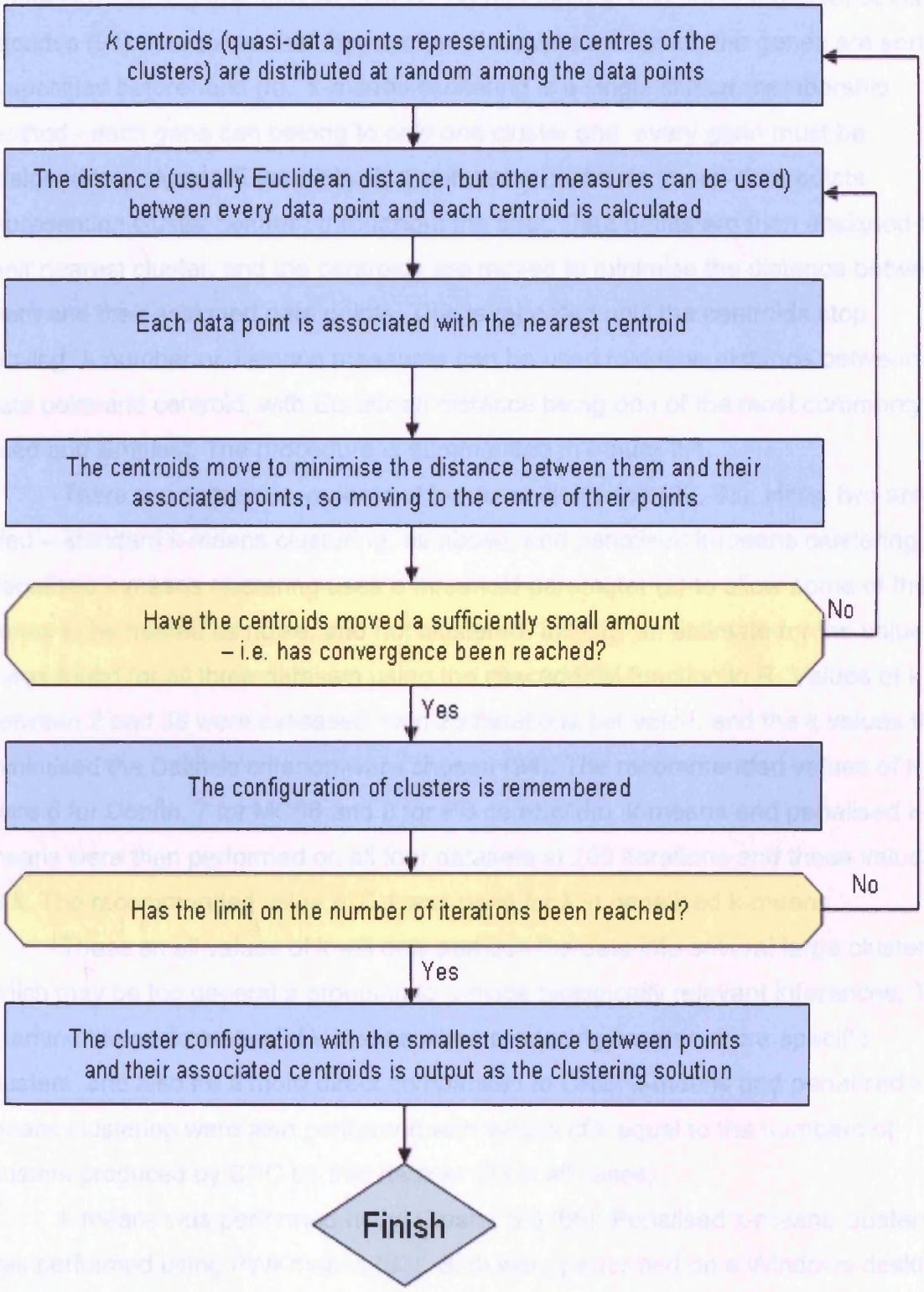
GO enrichment is an overrepresentation analysis method that assesses the percentage of clusters that are significantly enriched (compared to all annotated genes on the microarray) with genes from one or more Gene Ontology categories (using the goa_human dataset provided with GOstat) at different significance levels, using Fisher's exact test and the Benjamini false discovery rate multiple testing correction(88). Clusters were tested for enrichment (using Fisher's exact test) for all GO biological process terms 3 or more levels deep into the hierarchical tree of GO terms, at several different levels of significance. At least 3 genes from the input cluster had to match a GO category for the cluster to be counted as enriched for that category, to ensure that chance appearance of 1 or 2 genes from a GO category with few members could not affect the results. The percentage of clusters matching this criterion gives a measure of the biological, functional relevance of the clusters.

GO enrichment was determined with the web-based service, GOstat (89). This accepts multiple kinds of gene name or ID as input, allowing approximately 85% of genes within the input clusters to be included. This was automated using WWW-Mechanize, a Perl module(90).

2.2.5 Random cluster set construction

To compare the results of GO enrichments for the various clustering algorithms, several random cluster sets were also examined using GO enrichment. Four sets of clusters with the same distribution of cluster sizes as those made by k-means (at the value of k recommended by cascadeKM), CRC, ISA and memISA (both after removal of overlapping clusters) were produced. The cluster sets made from the Dobrin, MC66 and PB datasets were combined when determining the distribution of sizes. The new cluster sets had genes chosen at random from all those available on the Affymetrix U133 Plus 2.0 chip.

Figure 2.1 - Flowchart summarising the method used by k-means clustering



k is a user-defined input parameter which sets the number of clusters k-means clustering will find.

2.2.6 k-means

k-means clustering is a standard clustering method that has been in use for several decades (91). It requires that the number of clusters into which the genes are sorted is specified beforehand (*k*). k-means clustering is a single cluster membership method - each gene can belong to only one cluster and every gene must be assigned to a cluster. Essentially, it distributes *k* centroids (quasi-data points representing cluster centres) throughout the data. Data points are then assigned to their nearest cluster, and the centroids are moved to minimise the distance between them and their assigned data points. This is repeated until the centroids stop moving. A number of distance measures can be used to define distance between data point and centroid, with Euclidean distance being one of the most commonly used and simplest. The procedure is summarised in Figure 2.1.

There are numerous variants of k-means clustering (92, 93). Here, two are tried – standard k-means clustering, as above, and penalised k-means clustering. Penalised k-means clustering uses a threshold parameter (λ) to allow some of the genes to be treated as noise, and not clustered. Initially, an estimate for the value of k was found for all three datasets using the cascadeKM function in R. Values of k between 2 and 35 were assessed, with 25 iterations per value, and the k values that minimised the Calinski criterion were chosen (94). The recommended values of k were 6 for Dobrin, 7 for MC66 and 8 for PB cerebellum. k-means and penalised k-means were then performed on all four datasets at 200 iterations and these values of k. The recommended value of 0.1 was used for λ in penalised k-means.

These small values of k will only partition the data into several large clusters, which may be too general a grouping to provide biologically relevant inferences. To examine the performance of k-means when producing smaller, more specific clusters, and also for a more direct comparison to CRC, k-means and penalised k-means clustering were also performed with values of k equal to the numbers of clusters produced by CRC on that dataset (23 in all cases).

k-means was performed using Cluster 3.0 (66). Penalised k-means clustering was performed using PWKmeans (93). Both were performed on a Windows desktop PC with 2 GB RAM, using a 2.66 Ghz processor.

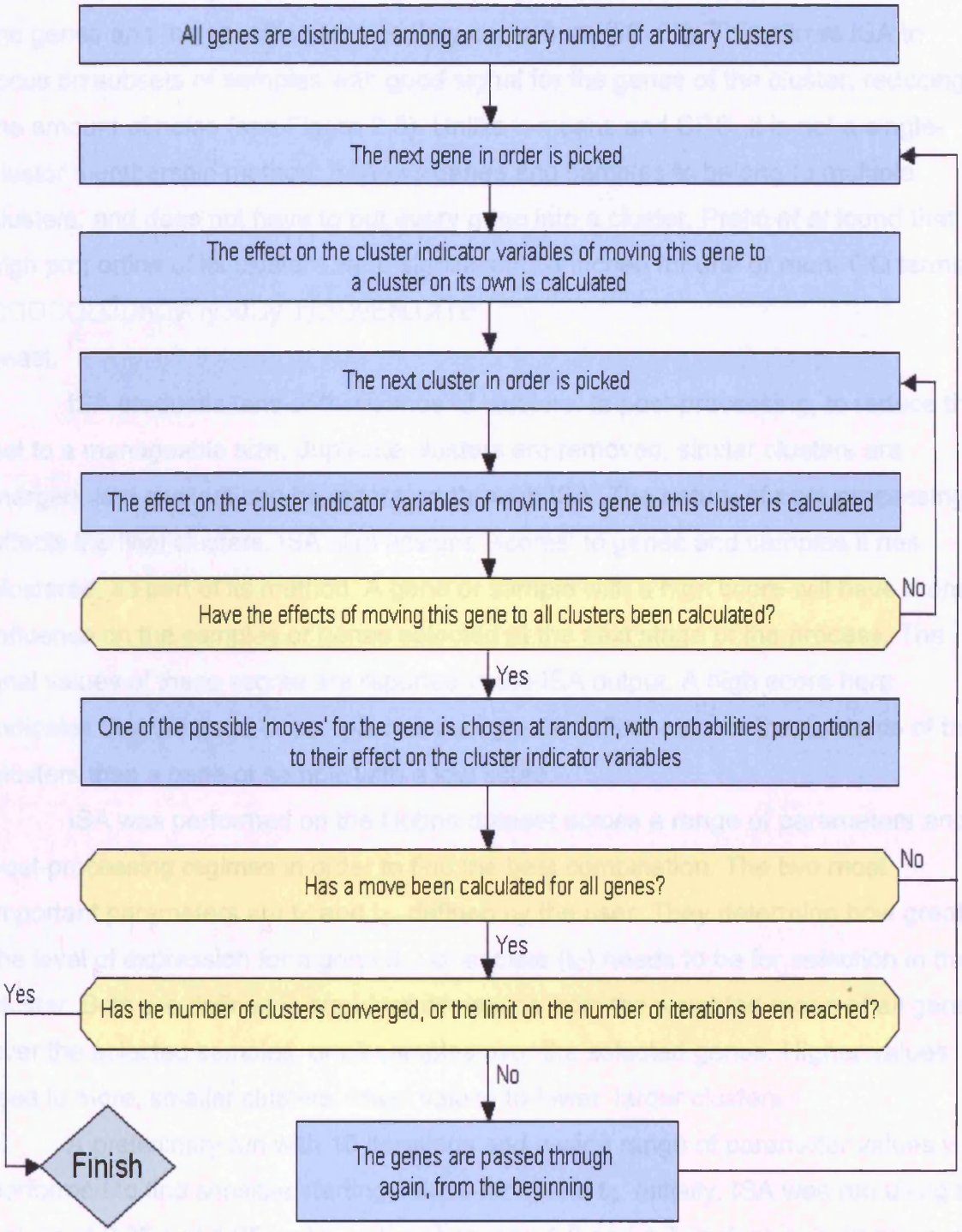
2.2.7 CRC

CRC(67) is a model-based clustering method.. The name arises from a metaphor where genes are regarded as customers in a Chinese restaurant with unlimited tables of unlimited size, each representing a cluster, and their food orders represent the expression profile of each gene. The customers are then seated at tables according to the similarities of their food orders. CRC has several advantages over other methods. It can handle missing data and cluster genes based on negative correlation and time-shifted correlation. Like k-means it is a single cluster membership method. Its methodology is complex, and is based upon treating the expression profiles of the genes as the sum of multiple normal distributions.

The procedure is outlined in Figure 2.2. Each iteration of the flowchart in Figure 2.2 can be considered a Markov chain process. CRC runs a number of these chains in parallel (set by the user - 10 is the recommended amount), and reports the highest likelihood cluster set as the final output. The chains are also limited to a certain number of iterations through the flowchart before reporting their clusters. This is another parameter decided by the user, and is recommended to be set at 20. Finally, a probability cut-off can be input, which determines how high the likelihood of a gene being a member of a cluster needs to be in order for it to be included in the final output. In practice, most genes are members of their cluster with probability 1, so this removes few genes.

CRC was performed on all three datasets. It was performed at two parameter sets for each dataset – 10 chains/20 cycles per chain/probability cut-off of 0.7, and 20 chains/40 cycles per chain/probability cut-off of 0.9. CRC was performed using a standalone program (67). It was performed on a Unix server running Redhat OS with 32 GB RAM, using one 2.2 Ghz processor.

Figure 2.2 - Flowchart summarising the method used by CRC



One run through this flowchart equates to a single chain in CRC, with several chains being run in parallel. The number of parallel chains and the maximum number of iterations are user-defined parameters.

2.2.8 ISA

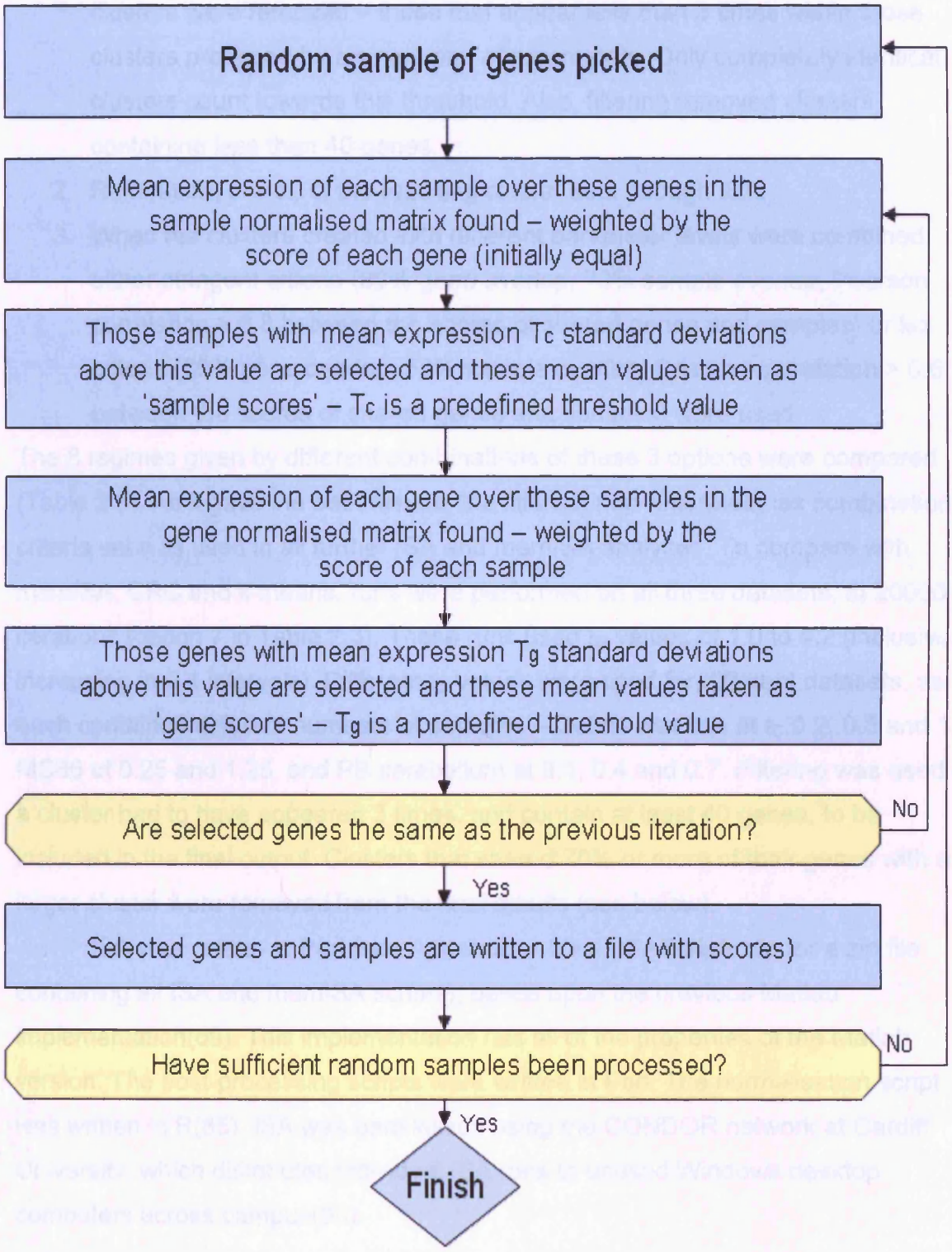
ISA is a biclustering method – it clusters both rows and columns of the dataset, here the genes and the specific samples they come from (68, 69). This allows ISA to focus on subsets of samples with good signal for the genes of the cluster, reducing the amount of noise (see Figure 2.3). Unlike k-means and CRC, it is not a single-cluster membership method: it allows genes and samples to belong to multiple clusters, and does not have to put every gene into a cluster. Prelić *et al* found that a high proportion of its clusters were significantly enriched for one or more GO terms □□□□□□Ĵ□n□ÿ□ÿ30□ÿ□ÿ□□ÿEN.CITE
yeast.

ISA produces tens of thousands of clusters. In post-processing, to reduce this set to a manageable size, duplicate clusters are removed, similar clusters are merged, and clusters can be reiterated through ISA. The nature of post-processing affects the final clusters. ISA also assigns 'scores' to genes and samples it has clustered, as part of its method. A gene or sample with a high score will have more influence on the samples or genes selected at the next stage of the process. The final values of these scores are reported in the ISA output. A high score here indicates that the gene or sample has had greater influence over the contents of the clusters than a gene or sample with a low score.

ISA was performed on the Dobrin dataset across a range of parameters and post-processing regimes in order to find the best combination. The two most important parameters are t_G and t_C, defined by the user. They determine how great the level of expression for a gene (t_G) or sample (t_C) needs to be for selection in the cluster. Both are defined in standard deviations from the weighted mean of all genes over the selected samples, or all samples over the selected genes. Higher values lead to more, smaller clusters, lower values to fewer, larger clusters.

A preliminary run with 10 iterations and a wide range of parameter values was performed to find sensible starting values for t_C and t_G. Initially, ISA was run using t_C values of 0.25 and 1.25 and t_G values between 1.0 and 4.7 (inclusive, increasing in 0.1 intervals), with 10000 iterations. Runs were also performed at 20000 and 30000 iterations. The effects of 3 post-processing options, each with 2 choices, on the clusters produced from the Dobrin dataset were investigated:

Figure 2.3 - Flowchart summarising the method used by ISA



1. The presence or absence of filtering. With filtering used, low-occurrence clusters were removed – those that appear less than 3 times within those clusters produced by a single pair of parameters. Only completely identical clusters count towards this threshold. Also, filtering removed clusters containing less than 40 genes.
2. Reiteration, or not, of the resulting cluster sets through ISA.
3. When the clusters created with different parameter levels were combined, either stringent criteria (80% gene overlap, 70% sample overlap, Pearson correlation > 0.8 between the scores of shared genes and samples) or lax criteria (60% gene overlap, 50% sample overlap, Pearson correlation > 0.6 between the scores of shared genes and samples) were used.

The 8 regimes given by different combinations of these 3 options were compared (Table 2.3). As it gave the best results, the filtered, non-reiterated, lax combination criteria set was used in all further ISA and memISA analyses. To compare with memISA, CRC and k-means, runs were performed on all three datasets, at 20000 iterations (option 7 in Table 2.3). These runs used t_G values of 1.0 to 4.2 (inclusive, increasing in 0.1 intervals). Different t_C values were used for different datasets, as each contained different numbers of samples – Dobrin was run at t_C 0.2, 0.5 and 1.0, MC66 at 0.25 and 1.25, and PB cerebellum at 0.1, 0.4 and 0.7. Filtering was used – a cluster had to have appeared 3 times, and contain at least 40 genes, to be included in the final output. Clusters that shared 70% or more of their genes with a larger cluster were removed from the final results (see below).

ISA was written in Perl (see Appendix A file 1, ISAScripts.zip for a zip file containing all ISA and memISA scripts), based upon the previous Matlab implementation(69). This implementation has all of the properties of the Matlab version. The post-processing scripts were written in Perl. The normalisation script was written in R(85). ISA was parallelised using the CONDOR network at Cardiff University, which distributes individual ISA runs to unused Windows desktop computers across campus(95).

Table 2.3 – Effects of different post-processing techniques on GO enrichment of clusters derived from ISA on Dobrin dataset

	1	2	3	4	5	6	7	8
Filtered for size and occurrence	No	No	No	No	Yes	Yes	Yes	Yes
Reiterated	Yes	Yes	No	No	Yes	Yes	No	No
Stringent combining criteria	No	Yes	No	Yes	No	Yes	No	Yes
% enriched at p-val < 0.3	78.2	80.0	80.2	82.3	86.2	84.2	75.7	87.0
% enriched at p-val < 0.1	48.7	47.2	51.6	49.3	52.3	56.3	48.6	57.5
% enriched at p-val < 0.05	34.6	36.4	38.5	39.6	40.0	46.2	45.9	48.7
% enriched at p-val < 0.01	28.2	25.4	28.6	31.2	36.9	33.0	35.1	36.0
% enriched at p-val < 0.001	23.1	17.1	25.3	22.6	30.8	23.7	29.7	24.9
% enriched at p-val < 0.0001	20.5	12.5	18.7	17.9	27.7	19.2	27.0	19.9

Clusters containing 40 or less genes, or that appeared less than 3 times in the output, were filtered out in the 'filtered' sets (sets 5-8). Reiterated sets were passed through ISA again (sets 1-2 and 5-6). Stringent combination criteria (80% gene overlap, 70% sample overlap, r > 0.8) were used in sets 2, 4, 6 and 8, and lax combination criteria (60% gene overlap, 50% sample overlap, r>0.6) were used in sets 1, 3, 5 and 7.

2.2.9 memISA

The underlying method of memISA is closely based on ISA and similar to PISA(71) (Figure 2.3). It biases against both genes and samples that have already been put into a cluster, according to two user input parameters, f and n. The bias is calculated relative to the highest scoring gene and sample in a cluster – this has its gene/sample score multiplied by the factor $(1 - f)$ in future iterations. All other genes/samples found in a cluster have their future scores reduced by a smaller amount. This is determined by the proportion of their score and the highest gene/sample score – a gene with a quarter of the score of the highest gene will have its future scores multiplied by $1 - (f * 0.25)$. The intent of this is to bias against the highest scoring genes of a cluster while allowing lower scoring genes to be relatively unaffected and still be included in subsequent clusters (the highest scoring genes typically have scores 10 times greater than the majority of genes in a cluster). If a gene/sample is included in a subsequent cluster, the biases are multiplied together – a gene which is the strongest gene in two successive clusters would have its score multiplied by $(1 - f)^2$ in following iterations. These biases are only remembered for a

certain number of iterations (*n*). Every *n* iterations, the slate is wiped clean. This is to ensure that memISA does not begin returning noise once it has found all the available clusters in the data, and to limit the effect that an early misclustering can have on the results.

memISA was run on the Dobrin dataset at 20000 iterations with a number of different values for *f* and *n* (Table 2.4). It was found the results were generally robust to the values of *f* and *n*, and that *f*=0.75 and *n*=5 produced clusters with the highest GO enrichment, so these values were used in all further analysis. A filtering step was also attempted on one dataset to see if it would improve GO enrichment. For this, those genes whose gene scores were in the bottom 10% for their cluster were removed from the cluster. This step reduced both gene coverage and GO enrichment and so was not used further.

memISA was run on the Dobrin, MC66 and PB cerebellum datasets at t_G 1.0 to 4.2 (inclusive, increasing in 0.1 intervals) and t_C 0.2, 0.5 and 1.0. Filtering was carried out as with ISA, using an occurrence threshold of 3 and a size threshold of 40. memISA was implemented in Perl, and was based upon the new Perl implementation of ISA. Like ISA, it was parallelised using CONDOR.

Table 2.4 – Comparison of GO enrichments for different memISA parameters in Dobrin (overlaps not removed)

% enriched at varying p-vals	$f = 0.5$, $n = 10$	$f = 0.75$, $n = 5$	$f = 0.75$, $n = 5$, 10% of genes with lowest gene scores removed from clusters	$f = 0.5$, $n = 3$
p-val < 0.3	62.5	92.3	88.5	85.7
p-val < 0.1	50.0	65.4	57.7	60.7
p-val < 0.05	50.0	57.7	53.8	50.0
p-val < 0.01	43.8	42.3	42.3	46.4
p-val < 0.001	37.5	38.5	38.5	42.9
p-val < 0.0001	31.3	34.6	26.9	28.6
Gene coverage	61.1	78.8	74.7	74.7
Number of clusters found	16	26	26	28

The parameter f controls how heavily the method biases against already-found genes. The parameter n controls how many iterations of memISA the biases are stored for.

2.2.10 Assessing overlap between clusters

We examined inter-method overlap in gene membership of clusters for the four methods and intra-method overlap of ISA and memISA. CRC and k-means, as single-cluster membership methods, had no intra-method overlap between their clusters. ISA and memISA cluster sets, however, both contained a large amount of intra-method overlap, making them impossible to compare fairly with clusters produced by k-means or CRC. To try to facilitate fair comparison, clusters with gene overlap above a certain level (values of 60, 70 and 80% gene overlap were tried) were merged but since this resulted in datasets with fewer than 3 clusters, this approach was abandoned. As an alternative, where over 70% of the genes in the smaller of a pair of clusters was shared with a larger cluster, the smaller cluster was removed. This process was performed on a subset of ISA and memISA output – those raw clusters produced at $t_G = 2.1$ or greater were used, and the rest discarded. This was in order to prevent a few very large clusters causing the removal of nearly all smaller clusters. This overlap removal step was applied after all other post-processing.

2.2.11 Combining methods

As there was not a large amount of overlap in clusters obtained between the ISA methods and either CRC or k-means, the possibility of combining their cluster sets to improve GO enrichment was investigated. The cluster sets were simply combined and clusters that had over 70% gene overlap with a larger cluster were removed as above. One set contained k-means, memISA and ISA clusters, one set contained CRC, memISA and ISA clusters. The memISA and ISA clusters had already had overlaps removed before combining. The CRC set used was the 10 chains/20 cycles per chain /0.7 cut-off. The k-means sets used were the k=23 sets.

2.3 Results

All four methods performed better than the random cluster sets when examined using GO enrichment to represent known biological relationships (Figures 2.4-2.6). This implies that all the clustering methods result in groupings of biological significance. Of the three random cluster sets, those with the same size distribution as ISA had slightly lower GO enrichment than those with the same size distribution as memISA or CRC. This may suggest that GO enrichment has a small bias against ISA due to the sizes of clusters it produces. However, at $p<0.05$ the difference dropped to under 1% GO enrichment, suggesting that any such bias is slight and may well be due to chance.

2.3.1 k-means and penalised k-means

k-means and penalised k-means produced clusters with high GO enrichments, especially at the lower k values recommended by cascadeKM. In these low-k cluster sets, k-means obtained higher GO enrichments than penalised k-means. In the k=23 cluster sets, they produced cluster sets with similar GO enrichment (Figures 2.4-2.6). As k-means gave similar GO enrichment to penalised k-means and by definition clustered more genes it was used in comparisons with the other methods.

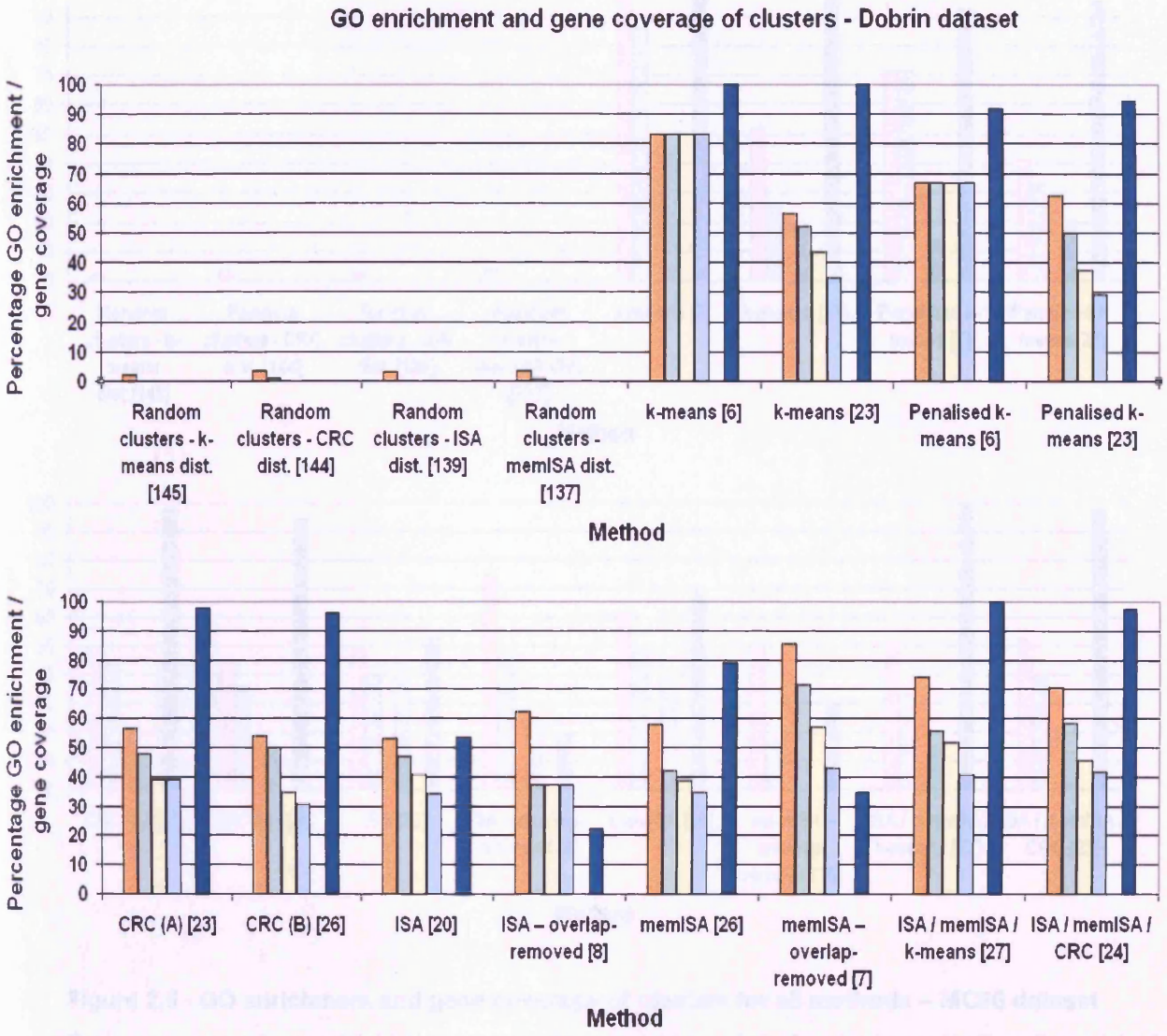


Figure 2.4 - GO enrichment and gene coverage of clusters for all methods – Dobrin dataset
Orange, green, yellow and light blue bars are the percentage of clusters that are significantly enriched for one or more GO categories at $p<0.05$, 0.01, 0.001 and 0.0001 respectively. The dark blue bar is gene coverage, the percentage of genes available on the chip that are assigned to at least one cluster. Numbers in square brackets are the number of clusters produced by that method. 'Dist.' = distribution of sizes. Parameter set A for CRC is 10 chains and 20 iterations per chain. Parameter set B for CRC is 20 chains and 40 iterations per chain.

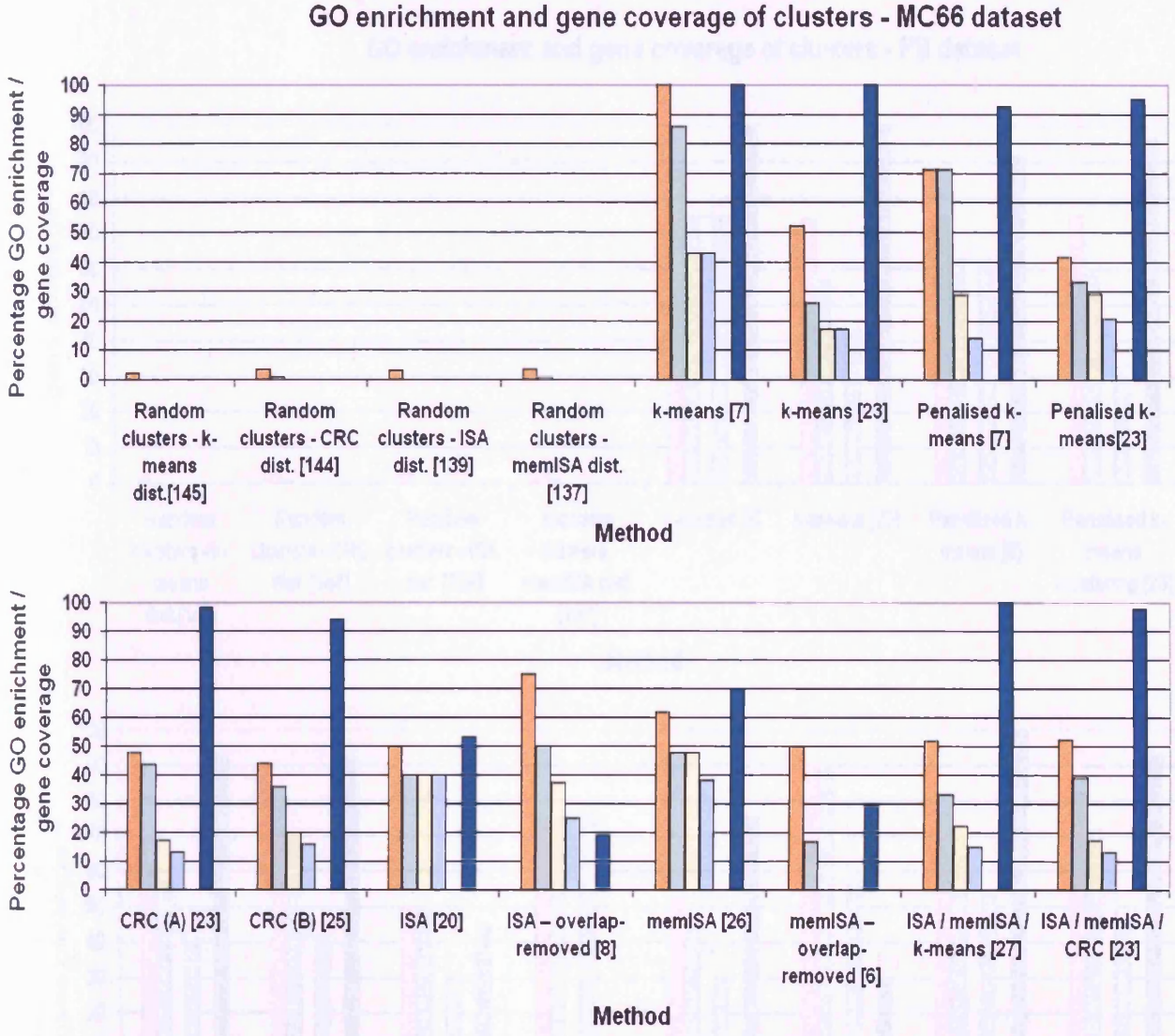


Figure 2.5 - GO enrichment and gene coverage of clusters for all methods – MC66 dataset
Orange, green, yellow and light blue bars are the percentage of clusters that are significantly enriched for one or more GO categories at $p<0.05$, 0.01, 0.001 and 0.0001 respectively. The dark blue bar is gene coverage, the percentage of genes available on the chip that are assigned to at least one cluster. Numbers in square brackets are the number of clusters produced by that method. 'Dist.' = distribution of sizes. Parameter set A for CRC is 10 chains and 20 iterations per chain. Parameter set B for CRC is 20 chains and 40 iterations per chain.

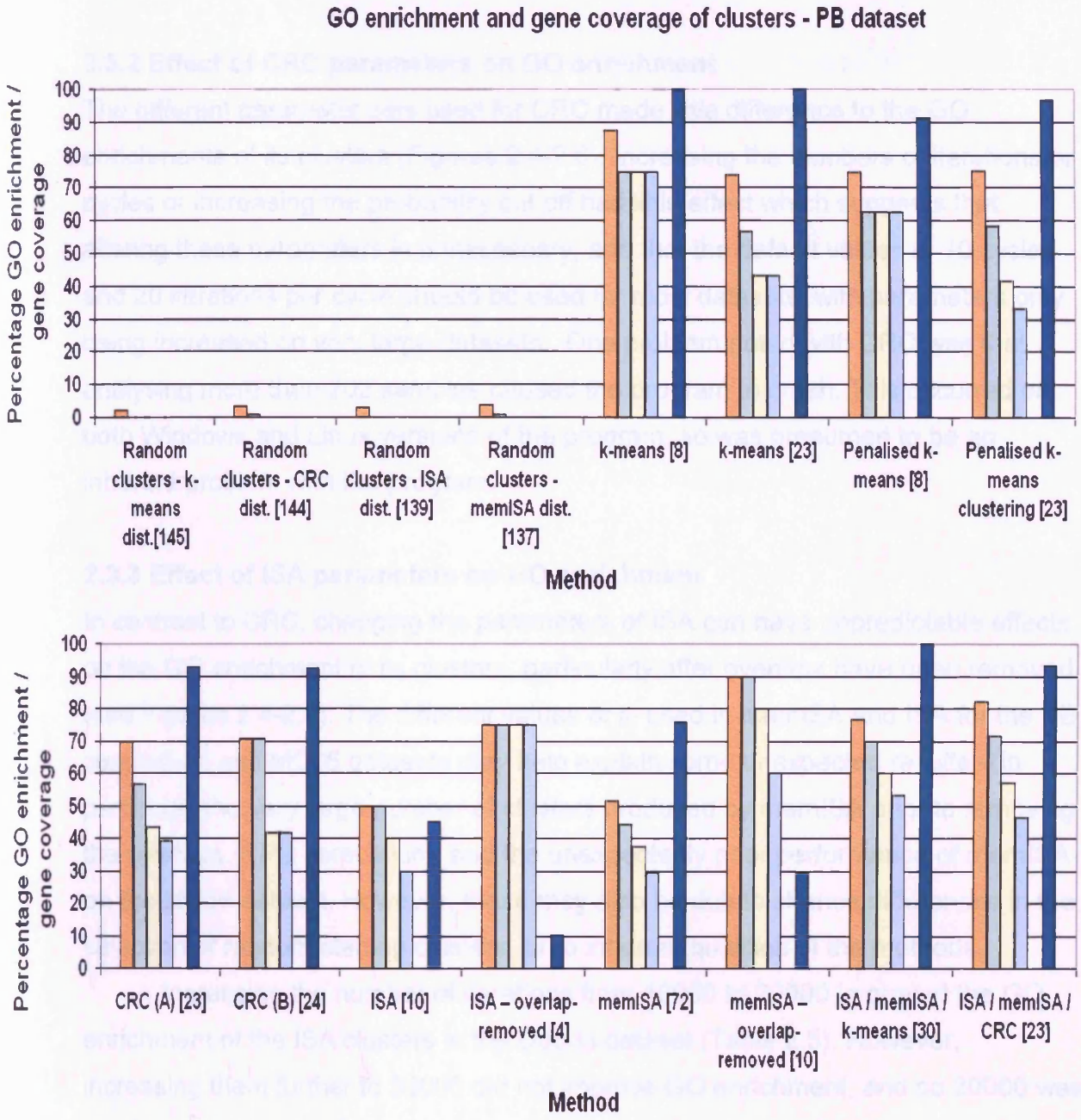


Figure 2.6 - GO enrichment and gene coverage of clusters for all methods – PB dataset
Orange, green, yellow and light blue bars are the percentage of clusters that are significantly enriched for one or more GO categories at $p<0.05$, 0.01, 0.001 and 0.0001 respectively. The dark blue bar is gene coverage, the percentage of genes available on the chip that are assigned to at least one cluster. Numbers in square brackets are the number of clusters produced by that method. 'Dist.' = distribution of sizes. Parameter set A for CRC is 10 chains and 20 iterations per chain. Parameter set B for CRC is 20 chains and 40 iterations per chain.

2.3.2 Effect of CRC parameters on GO enrichment

The different parameter sets used for CRC made little difference to the GO enrichments of its clusters (Figures 2.4-2.6). Increasing the numbers of iterations or cycles or increasing the probability cut off had little effect which suggests that altering these parameters is unnecessary, and that the default values of 10 cycles and 20 iterations per cycle should be used for most datasets, with parameters only being increased on very large datasets. One problem noted with CRC was that analysing more than 202 samples caused the program to crash. This occurred on both Windows and Linux versions of the program, so was presumed to be an inherent problem with the program.

2.3.3 Effect of ISA parameters on GO enrichment

In contrast to CRC, changing the parameters of ISA can have unpredictable effects on the GO enrichment of its clusters, particularly after overlaps have been removed (see Figures 2.4-2.6). The different values of t_C used in memISA and ISA for the PB cerebellum and MC66 datasets may help explain some unexpected results – in particular, the very large number of clusters produced by memISA prior to removing the overlaps in PB cerebellum, and the unexpectedly poor performance of memISA on the MC66 dataset. However, these may also be due to chance differences in the selection of random starting clusters, or to inherent qualities of the methods.

Increasing the number of iterations from 10000 to 20000 improved the GO enrichment of the ISA clusters in the Dobrin dataset (Table 2.5). However, increasing them further to 30000 did not improve GO enrichment, and so 20000 was used as the number of iterations for all other runs.

Table 2.5 – Comparison of GO enrichment and gene coverage of ISA clusters at different numbers of reiterations

% enriched at varying p-values	10,000 iterations	20,000 iterations	30,000 iterations
p-val < 0.3	75.7	75.0	77.4
p-val < 0.1	48.6	56.3	51.6
p-val < 0.05	45.9	53.1	48.4
p-val < 0.01	35.1	46.9	38.7
p-val < 0.001	29.7	40.6	35.5
p-val < 0.0001	27.0	34.4	32.3
Gene coverage	53.3	53.2	53.3

Table showing the GO enrichment and gene coverage of ISA clusters at 10000, 20000 and 30000 iterations.

2.3.4 Effect of memISA parameters on GO enrichment

memISA is robust to the choice of f and n, as all of the combinations tried produced reasonable GO enrichments (see Table 2.4). f=0.7 and n=5 were chosen because they produced clusters with slightly better GO enrichments than other parameter sets.

2.3.5 Comparison of clusters detected

There was a large amount of overlap between the clusters produced using penalised k-means and k-means at k= 23, with the majority of clusters (from all three datasets) having over 70% overlap with a cluster from the other method, and all others having over 40% overlap (see Table 2.6 and Appendix A file 2, AllOverlaps.xls for more detail).

Since these methods found similar clusters, further analysis was focused on standard k-means clustering, as it had 100% gene coverage.

Table 2.6 - Percentage overlap between clusters produced by different methods

	k-means	Penalised k-means	CRC	ISA	memISA
k-means	100	62.3	52.2	8.7	8.7
Penalised k-means	63.8	100	57.5	4.3	4.3
CRC	52.2	54.5	100	7.2	27.5
ISA	25	25	23.1	100	95.2
memISA	26.1	26.1	26.1	56.5	100

Values in table indicate the percentage of clusters produced by the method in the first column that have over 70% gene overlap with one or more clusters produced by the method in the top row.

There was considerable overlap in the results obtained between k-means and CRC across all three datasets. This suggests that k-means and CRC find similar patterns within the datasets. Conversely, there was little overlap between either k-means or CRC and either memISA or ISA clusters. In the case of ISA, there were a few overlaps at 70% or above for each dataset. In the case of memISA, there was a large cluster that overlapped with several of the smaller clusters produced by CRC at 70% or more, plus one other 70% plus overlap between more similarly sized clusters, in all three datasets.

Removing clusters with over 70% intra-method gene overlap from the ISA and memISA cluster sets reduced the number of clusters considerably. These sets contained only 4-10 clusters and were much smaller than the original ones. However, their GO enrichments were generally considerably higher (see Figures 2.4-2.6) but at the cost of a considerable drop in gene coverage.

2.3.6 Combining methods

The cluster sets produced by combining the methods had similar gene coverage to those produced by CRC/k-means alone (see Figures 2.4-2.6). They generally had a higher number of clusters. For the CRC/ISA/memISA combined set, the GO enrichment of these clusters was higher in the Dobrin and PB cerebellum datasets. In the k-means/ISA/memISA combined sets, the gains in GO enrichment relative to k-means alone were generally smaller: under 5% at most levels of p. There were a

few small losses in GO enrichment in some datasets and at some levels of p, but generally the impact on GO enrichment was still positive.

2.3.7 Gene coverage

Before highly overlapping clusters were removed from the clusters produced by ISA, k-means had the highest gene coverage (100% by definition), followed by CRC, and then by memISA and lastly ISA. However, these cluster sets are not directly comparable on number of clusters or on GO enrichment, as the cluster sets produced by ISA and memISA contain a large amount of redundancy.

As memISA and ISA had much lower gene coverage than k-means or CRC, the relationship between mean gene expression levels and cluster membership was examined for these methods in the Dobrin dataset. For both ISA and memISA, no significant correlation was found (r=-0.132 for ISA, r=-0.081 for memISA).

2.3.8 Cluster size

The number of genes per cluster for each method and dataset was also examined, and the mean cluster size and standard deviation computed (see Appendix A file 3, SizeDistribution.xls). Generally, CRC, k-means and penalised k-means were consistent in their cluster sizes, which appear to vary only with the number of genes in the dataset. The average cluster size was between 800 and 900 for these three methods in both 133P datasets (Dobrin and PB), and between 500 and 600 in the MC66 dataset. ISA generally produces clusters that are smaller than this, between 400 and 600 on average (with no obvious relationship to number of genes or samples in the dataset). memISA, conversely, is particularly prone to producing datasets with one or two particularly large clusters, giving it a higher average cluster size and standard deviation. This is because the larger number of unique clusters it produces makes it more likely for clusters to overlap and be merged, leading to these extremely large clusters.

To examine whether cluster size affected enrichment, cluster size was checked for correlation with $\log_{10}$ of the p-values of the best GO hit for each cluster (unenriched clusters were treated as having a p-value of 1). No significant correlation was found for any of the methods.

2.3.9 Speed

The three datasets were used to evaluate approximate runtimes for the four methods (Table 2.7). CRC and k-means are very fast methods, with a runtime of a few hours on current computer technology with typical parameters on human brain datasets. ISA and memISA, meanwhile, are much slower, taking up to a month without parallelisation. Even with parallelisation using CONDOR, ISA and memISA can take over 24 hours for a full parameter set when post-processing is included. Restricting the parameters to t_G 2.1 and above, as in the non-overlapping cluster set(see section 2.2.10 above), reduces these times by up to half.

Table 2.7 - Comparison of method runtimes

Runtime on different datasets	ISA (using CONDOR)	memISA (using CONDOR)	CRC – 10/20	CRC – 20/40
Dobrin	23h 6min	37h 22min	2h 12min	7h 53min
MC66	17h 23min	28h 55min	1h 15min	4h 33min
PB cerebellum	15h 11min	24h 13min	1h 7min	3h 53min

Table showing the real-world time taken for the methods to run on each dataset.

2.4 Discussion

2.4.1 Inter- and intra-method gene overlap

Nearly all ISA clusters had over 70% overlap with a memISA cluster across all three datasets. However, less than half of the memISA clusters had over 70% overlap with a cluster from ISA, as many of the ISA clusters overlap with the same memISA cluster. This level of overlap is surprisingly high, considering that their post-processing regimes already include a step to merge similar clusters. However, this step requires high sample overlap and correlation of shared gene/sample scores in addition to simple gene overlap. It also uses the size of the larger cluster to calculate overlap – i.e. 50% overlap in this step indicates that 50% of the genes in the larger cluster are found in the smaller cluster. As a result, it tends to only combine clusters of a similar size. The ability of memISA to bias against already-found clusters may

help it find clusters that would previously have been hidden by a stronger cluster, a useful feature when looking for novel clusters.

The tendency of the cluster merging step in ISA and memISA to only combine clusters of a similar size may help to explain the improvement in GO enrichment the removal of overlapping clusters produces. Requiring a similar size and similar samples and gene/sample scores may help to ensure that only those clusters which come from the same signal are actually merged, excluding noise clusters with a coincidentally high gene overlap. The overlap removal process would then remove these clusters from the dataset altogether, improving GO enrichment.

The reasons for the poorer performance of memISA on the MC66 dataset are not known. It is possible that the difference in the t_C and t_G parameters between memISA and ISA for this dataset was critical. The smaller number of genes in this dataset might also be important, and so reducing the values of t_G used may help. Alternatively, it might be that chance played a role. memISA may be inherently more prone to chance variation than ISA or CRC.

2.4.2 Comparisons with other clustering method surveys

The findings here broadly agree with several other surveys of clustering methods (Figures 2.4-2.6). ISA is an effective method that produces clusters with high GO enrichment, as suggested previously by Prelić *et al* (60), but the cluster sets presented here generally do not have as high a proportion of GO enriched clusters as theirs. This is likely to be a consequence of the greater complexity of the input data. Using synthetic datasets, Prelić *et al* found that ISA coped well with high levels of noise, but was affected by overlapping clusters. It is likely that the complex human brain datasets used here have more overlapping clusters than the *S. cerevisiae* datasets Prelić *et al* used. This may explain why memISA generally had superior GO enrichments to ISA in my analysis, as the capacity to use previously found clusters to inform further clustering should help it uncover clusters that are overlapped by a more obvious cluster.

Garge *et al* found k-means clustering effective (70) on a wide range of input datasets. They compared the cluster sets produced by subsets of the datasets, using similarity as a measure of cluster stability. Although k-means performed well, none of the methods they tested had high stability scores (over 0.55), even on

datasets with sample sizes of 50 or more. The results do not examine stability directly, but suggest that clusters can be biologically meaningful at similar sample sizes, despite this instability.

The results of Garge *et al* are echoed by the k-means cluster sets reported here, which have high GO enrichment and gene coverage scores. These scores were generally higher than CRC, the mixture modelling method examined here. This contrasts with the findings of Thalamuthu *et al*, who found that modelling methods were superior to k-means clustering (65). This difference is again likely to be due to the datasets used; in particular the datasets used here were much larger in size. When using synthetic datasets, Thalamuthu *et al* found that MCLUST (a model-based clustering method) and tightClust were better at ignoring scattered background genes in favour of genuine clusters (74, 96). In this study, the probability cut-off parameter of CRC or the penalised version of k-means are similarly intended to allow the method to exclude peripheral cluster genes that may only be cluster members by chance. However, these methods did not prove superior to standard k-means in terms of GO enrichment scores, so this ability may not be critical to successfully clustering complex human datasets.

k-means clustering, CRC, ISA and memISA are all potentially useful methods. Considered alone, k-means clustering is probably the most useful of the four, as it is fast, does not require parallelisation, and produces clusters with slightly higher levels of GO enrichment than CRC when producing similar numbers of clusters. When used to find smaller numbers of clusters more in line with the estimation of k, the GO enrichments are higher still, reaching 100% at some levels of p. It also assigns a cluster to every gene (100% gene coverage), unlike overlap-removed ISA and memISA (under 30% gene coverage). Although this must lead to some false positives, this does not seem to have affected the GO enrichment scores unduly, and is an advantage in exploratory studies where as wide a view as possible is desired.

Furthermore, k-means is a relatively simple and very well understood method. This simplicity may be the reason for its good performance here, as it may allow it to cope with a wide variety of input data. CRC, conversely, has many more parameters and so may have had scope to become optimised for the smaller yeast and bacterial datasets it was built for and tested upon.

However, for the fullest picture of clusters available in a dataset, combining memISA, ISA and k-means is the best option, as it offers higher GO enrichment than k-means alone in two out of the three test datasets while retaining 100% gene coverage (see Figures 2.4-2.6). Even in the MC66 dataset, it added additional clusters not found by k-means without reducing GO enrichment. One of these memISA clusters (found in both dorsolateral prefrontal cortex datasets) was found to be significantly enriched for schizophrenia-associated genes and genes differentially expressed in schizophrenia (see Chapter 4 below), further emphasising the utility of combining methods. If time allows, this combined method should be the method of choice for clustering microarray brain expression data.

Chapter Three

Expression quantitative trait loci and polygenic score analysis in schizophrenia

3.1 Introduction

3.1.1 Summary

It is widely thought that alleles that influence susceptibility to common diseases, including schizophrenia, will frequently do so through effects on gene expression. Since only a small proportion of the genetic variance for schizophrenia has been attributed to specific loci, this remains an unproven hypothesis. The International Schizophrenia Consortium (ISC) recently reported analyses that would support a substantial polygenic contribution to that disorder, and showed that schizophrenia risk alleles are enriched among SNPs selected for marginal evidence for association ($p<0.5$) from genome wide association analyses (29). It follows that if schizophrenia susceptibility alleles commonly influence gene expression, those marginally associated SNPs which are also eQTLs should be enriched for true association signals compared with SNPs which are not eQTLs. To test this, I identified marginally associated ($p<0.5$) SNPs from two of the largest available schizophrenia GWAS datasets. eQTL status was assigned to those SNPs based upon eQTL datasets derived from adult human brain. Using the polygenic score method of analysis reported by the ISC, I observed that higher probability *cis*-eQTLs predicted schizophrenia status better in independent datasets than those with a lower probability for being a *cis*-eQTL. My data support the hypothesis that a proportion of common alleles confer risk of schizophrenia through impact on gene expression. Moreover, of considerable practical importance, my data show that notwithstanding the likely developmental origin of schizophrenia, studies of adult brain tissue can in principle allow relevant susceptibility eQTLs to be identified.

3.1.2 Background

A high proportion of mutations for simple (Mendelian) genetic disorders exert their pathogenic functional effects by altering the structure of the protein encoded by the mutant gene. However, this does not appear to be the case for the majority of susceptibility alleles for common phenotypes, at least not those that are common enough to be identified through genome-wide association studies (GWAS) (97). This is compatible with the hypothesis that inherited variation that impacts upon mRNA expression plays an important part in susceptibility to complex traits, including many human diseases (98-100). Since only a small proportion of the genetic variance for risk to common diseases has been attributed to specific loci (101, 102), this is an unproven hypothesis. This issue is a particular problem for schizophrenia and other psychiatric disorders, where only a small number of strongly supported associations to common alleles have been reported (29, 51, 103-105) and none of these has yet been functionally characterised.

3.1.3 Expression data and the validity of genetic association with disease

From the perspective of identifying risk alleles of genes, the hypothesis that susceptibility variants for schizophrenia will be enriched for variants that influence mRNA expression is not merely of academic interest. For example some authors (106-108) have reported associations between gene expression changes and particular genetic variants or haplotypes whose associations with schizophrenia are controversial, the idea being that association with expression lends additional support to the association with disease status. Others have also used this principle in non-psychiatric disorders to localize the likely susceptibility genes or functional variants within regions of association (109). Given that the effect sizes of most common alleles are small (29) and unlikely to be reliably separated from chance findings in the full genome context in the near future with available sample sizes (103), the ability to assign an enhanced prior probability to variants associated with gene expression may be of additional value in identifying novel disease associations.

Although the convergent use of expression and genetic data for informing pathophysiological theory seems intuitively reasonable (110), the validity of this approach for informing genetic studies is crucially dependent on the assumption that true associations are in fact enriched among variants that impact upon gene

expression. Moreover, in the case of schizophrenia, attempts to relate disorder-associated variants to effects on gene expression are generally based upon mRNA studies of adult brain, adult peripheral tissues, or cell lines derived from those peripheral tissues. Whether such studies are justified for disorders like schizophrenia, whose origins are thought to be developmental, is unclear, although it seems plausible they may be relevant since some characteristics of schizophrenia, such as grey matter loss, tend to appear in adolescence (111), and it seems reasonable to postulate that the effects of many regulatory variants relevant to aetiology may persist into adulthood, even those which exert their pathogenic effects in development.

3.1.4 Expression quantitative trait loci (eQTLs)

Here, I sought to test the hypothesis that polymorphisms that are associated with gene expression in adult brain samples are enriched among those that show evidence for association to schizophrenia. Loci that exert an effect on gene expression are often called expression quantitative trait loci (eQTLs) (112). At a genome-wide level, putative eQTLs can be identified by combining GWAS SNP data with global transcriptomic data obtained from the same subjects, the aim being to identify eQTLs by correlating genotypes at SNP loci with gene expression levels. In the present study, to identify putative eQTLs, I used a dataset reported by Myers and colleagues (42, 113), currently one of the largest expression datasets derived from human brain available that also contains genotype data for each sample. A few months prior to submission of this thesis, and genotype and expression data based upon human frontal cortex became available from another study, that of Gibbs *et al* (114), and this allowed me to attempt to replicate my primary observations from the dataset of Myers. I also used a dataset based upon human lymphoblastoid cell lines to examine whether eQTLs derived from non-brain tissue might have relevance to schizophrenia (115).

The Myers *et al* and Gibbs *et al* studies, which allow the determination of eQTLs by including both expression and genotype data, do not include data from brains of individuals who, during life, had suffered from schizophrenia (42, 114).

3.1.5 Polygenic scores

To identify sets of variants enriched for schizophrenia susceptibility alleles, I exploited the approach of the International Schizophrenia Consortium (29) who recently demonstrated the existence of thousands of risk alleles for schizophrenia. They also showed that these risk alleles are enriched among large sets of SNPs surpassing very liberal thresholds of association (e.g. $P<0.5$). In essence, the ISC defined sets of putative schizophrenia risk alleles in a training GWAS dataset as being those alleles that were more common in cases than controls at loci meeting very relaxed thresholds of significance for association. Individuals in independent test GWAS datasets were assigned what can be referred to as a 'polygenic score' based upon the number of putative risk alleles carried by that individual, and then the scores for cases and controls in those datasets were compared. The main finding was that in independent datasets, these 'polygenic scores' were significantly higher in cases than in controls. The ISC explored several thresholds for association in the training GWAS, including $p<0.1$, $p<0.2$, $p<0.3$, $p<0.4$ and $p<0.5$. The most significant distinction between diagnostic groups in the test samples occurred when the threshold for association in the training GWAS was set at $p<0.5$. Although increasing numbers of false positive SNPs must be included at more lax thresholds, it appears this was outweighed by the inclusion of more SNPs that captured susceptibility alleles.

The ISC also performed polygenic score analysis using both maximum and minimum p-value thresholds (e.g. SNPs within a range $0.2<p<0.5$). Polygenic scores based upon these SNP sets also regressed significantly upon affected status, further demonstrating that SNPs at these association thresholds are informative in relation to schizophrenia status. Modelling studies suggested that the most plausible explanation for the effectiveness of lax association thresholds was that there is a substantial polygenic component to schizophrenia comprising of thousands of risk alleles and that this contributes at least 30% of the overall variance in risk of the disorder at the population level.

Here, I used this general approach to test whether eQTLs are enriched among schizophrenia associated alleles. Schizophrenia 'risk' alleles were defined according to the method reported by the ISC (29) in a subset of the ISC data and also in the European American subset of the Molecular Genetics of Schizophrenia

study (104). These SNPs were then classified as 'top eQTL' and 'bottom eQTL' sets based upon their p-value for association with expression levels of transcripts in the various eQTL datasets, and these sets were then tested for differences in their polygenic scores in cases and controls independent of the training sets.

3.1.6 eQTL p-value collation

Three methods for collating the eQTL p-values into a single measure were considered. The first was to take the most significant eligible eQTL p-value for each SNP. The second method was simply to count the number of associations to all eligible transcripts that a given SNP has that surpass a nominal level of significance. The third method was 'cumulative minus log p', where the negative log of all eQTL p-values a SNP has is added together. This method offered a balance between the first method, which ranks SNPs with a single strong eQTL highest, and the second, which gives the highest scores to SNPs with a larger number of smaller eQTLs.

The first method was chosen for further analysis as it is simple and easy to implement. Also, when used in *cis* context, which requires SNPs and transcripts to be within a certain distance of each other before their eQTLs are considered (see Section 3.2.2 below), it is the least biased by the number of transcripts close to a SNP. When using the second method, a SNP within range of 10 genes will have, on average, 5 times as great a score as one in range of 2.

3.1.7 Subgroups of genes for calculating eQTLs

To test whether SNPs that affect the expression of specific subgroups of genes can offer superior predictive power compared to SNPs that affect any gene in the expression dataset, top and bottom eQTL SNP sets were determined for a number of gene subgroups. These included sets of genes which are differentially expressed in the post mortem brains of people who had suffered schizophrenia or bipolar disorder, according to the Stanley Medical Research Institute Online Genomics database (116). The bipolar disorder genes were included on the basis that schizophrenia and bipolar disorder have shown considerable genetic overlap in several studies (55).

Gene sets based upon coexpressed clusters of genes which are also enriched for schizophrenia-related genes were also examined. See Chapter 2 for

more details of the methods used to find these clusters, and Chapter 4 for functional analysis of these clusters. A gene set consisting of genes present in the Dobrin 3093 cluster (defined in Chapter 4, Section 4.2.1) was included; subsequently this is referred to as the 'Dobrin 3093' gene set.

3.2 Methods

3.2.1 Dataset acquisition and preparation

In the primary analysis, eQTL p values were calculated from the dataset of Myers and colleagues (42, 113). This contains genotypes (Affymetrix GeneChip Human Mapping 500K Array Set) for 380157 SNPs which met their quality control criteria (defined below) and expression (Illumina v1 Human RefSeq-8 BeadChip) data for 176 Alzheimer's disease cases and 188 controls. The analysis was restricted to the control samples to exclude the impact of neurodegeneration on gene expression measures. I selected this option rather than allowing for affected status in the analysis as a crude categorical adjustment will not allow for a number of variables within the affected group that will be expected to have major effects on gene expression, including possible aetiological heterogeneity, duration of illness prior to death, and rate of disease progression.

Beginning with the rank-invariant normalised expression data (42), samples with over 10% missing data were removed, as were probes with over 25% missing data. These values were arbitrarily chosen on the basis to be somewhat more stringent than the original publication without removing a large proportion of the data. Only the probe with the lowest proportion of missing data was retained for each gene (arbitrarily retaining the first to appear in the dataset file in the case of a tie). This was to prevent the results being biased in favour of transcripts with multiple probes, since multiple probes provide each SNP multiple opportunities to be designated an eQTL. However, by adopting this procedure, I effectively ignore the impact of eQTLs on transcript splicing isoforms. To minimize the impact of different brain regions in the dataset, I included only samples from the two most common regions represented in the study (frontal cortex and temporal cortex). Overall, 163 samples and 8361 probes were retained for analysis.

As in the primary publication, the expression data were log transformed to minimise the effect of departures from normality in the analysis (using the statistical package R (117)) and the data were adjusted for a number of non-genetic covariates using linear regression in R. These were gender, post mortem interval, age at death, institute performing the chip analysis, and hybridisation date, as well as for brain area (frontal or temporal cortex). I additionally covaried for the expression value for *Enolase 2* (ENO2), a neuronal marker. The intention in making this correction was to reduce expression variance arising from differing proportions of neurons between the samples (118, 119).

To examine whether the results from the primary analysis replicated, I used the frontal cortex dataset of Gibbs to derive eQTLs (114). The expression data were normalised and log transformed as described in the original publication. Samples and transcripts where over 50% of the data were missing were removed, leaving 133 samples and 14467 transcripts. The data were adjusted for covariates using linear regression in R (117) – these were gender, age at death, post mortem interval, institute performing the chip analysis, hybridisation batch, and ENO2 expression.

As a secondary analysis, the lymphoblastoid cell line GeneVar expression dataset was also used to derive eQTLs (115). This was to investigate the effect of using a non-brain tissue for eQTL determination. To prevent population differences affecting the results, this analysis was restricted to the CEU (North Americans of European descent) section of the GeneVar dataset. No additional quality control was performed; the genotype and expression datasets were used to derive eQTLs exactly as provided on the GeneVar website (http://www.sanger.ac.uk/humgen/genevar/).

For the Myers *et al* genotype data, the same quality control metrics as the original publication were used (42). All SNPs were required to have minor allele frequency of at least 1%, a call rate of at least 90%, and an exact Hardy-Weinberg equilibrium p-value > 0.05. The same quality control metrics as the original publication were also used for the Gibbs *et al* genotype data. SNPs in the Gibbs *et al* dataset were required to have at least three samples that were homozygous for the minor allele, a call rate of at least 95%, and an exact Hardy-Weinberg equilibrium p-value > 0.001.

The International Schizophrenia Consortium (ISC) (29) and Molecular Genetics of Schizophrenia (MGS) (104) GWAS datasets were used for the study as these are currently the largest GWAS datasets available to me. In exploiting these datasets, I essentially followed the study design of the ISC. For the initial analysis, the ISC dataset was divided to create training and target subsets. The dataset was split by assigning alternate cases and alternate controls to the training and target datasets, so that each contained half of the cases and half of the controls. These training and target datasets derived from within the ISC sample are termed the 'Split ISC datasets'. To derive a set of putative risk alleles fully independent of the ISC, I used the MGS European American dataset (104). P values were provided by the authors of that study as per the analysis reported in the primary publication.

3.2.2 eQTL determination

Linear regression of the expression values for each gene (correcting for covariates as described above) on SNP genotypes (coded on the number of minor alleles: 0, 1 or 2) was performed using PLINK (120, 121). This gave p-values for association between each SNP and the mRNA expression as measured by each probe-set. To test my hypothesis, I based the analysis upon *cis*-eQTL p-values. *Cis*-eQTLs are variants that are in chromosomal proximity to the transcripts they putatively regulate, and previous studies suggest *cis*-eQTLs have a higher prior probability for being true eQTLs than *trans* eQTLs (112), the latter being defined on the basis of association with transcripts with which they are not physically co-located. Moreover, *trans* eQTL analysis involves a much greater degree of multiple testing (all SNPs against all probesets) than *cis*-eQTL analysis.

These considerations suggest that sets of 'top *cis* eQTLs' will be more greatly enriched for true eQTLs than sets of top *trans* eQTLs, so restriction to *cis* eQTLs should enhance the power of the analysis. *cis* eQTLs were ranked by p value with respect to any transcript within a certain distance of the SNP locus – the '*cis* window'. A window of 100kb was used for the primary analysis. Exploratory windows of 50kb and 150kb were used for secondary analyses. The choice of distance is arbitrary, but 100kb was used in the primary analysis based upon a previous study suggesting that *cis* eQTLs are enriched within this boundary (122).

If a SNP was within range of multiple transcripts, the lowest p-value for any transcript was taken as the eQTL p-value. SNPs within range of multiple transcripts have multiple chances to attain a significant *cis* eQTL p-value; this is a potential source of bias but it may also be that SNPs that are close to a large number of genes are more likely to tag an eQTL than those near only one gene.

Given the presumed lower probability for any *trans* eQTL representing a true association, I expected that even if the primary hypothesis was correct, SNPs selected on this basis of *trans* eQTL status would be less effective at distinguishing between cases and controls. In the ISC study, when polygenic scores were calculated separately for SNPs within genes and SNPs over 500kb away from any gene, the former correlated with affected status more significantly than the latter (29). This suggests that SNPs which are distant from genes are less enriched for susceptibility alleles, although that analysis did not distinguish between SNPs that are or are not eQTLs. As a secondary analysis, I also explored the relative ability of top and bottom eQTLs after ranking those loci by the most significant p-value for association to any transcript in the dataset, that is, SNPs were ranked on the basis of *cis* and *trans* effects.

3.2.3 Risk allele counts

The SNPs available in the training datasets were placed into the following categories according to eQTL p-value: top 5% eQTLs, top 50% eQTLs, bottom 50% eQTLs, and bottom 5% eQTLs. As in the ISC study, the SNPs in all sets were linkage disequilibrium (LD) pruned according to the estimate of r^2 in the particular training dataset being used. A window size of 200 consecutive SNPs was compared for LD, which was moved along the chromosomes in steps of 5 SNPs. The maximum r^2 permitted between two compared SNPs was 0.25 – an r^2 greater than this resulted in the arbitrary removal of one of the SNPs.

In the randomly split ISC training datasets, as in the primary ISC paper (29), allelic P values and odds ratios for association were calculated by a Cochrane-Mantel-Haenszel test conditioned by country of origin using the QC-cleaned datasets provided by that group. Training on the MGS European American Sample was based upon the association results that formed the basis of the primary publication (104) and did not require access to individual genotypes. SNPs that had association

$p<0.5$ in training sets were carried through for polygenic score analysis. The alleles of these SNPs that were more common in cases were defined as risk alleles. PLINK (using the --score option) was then used to perform a count of the number of risk alleles for each sample in the target dataset, weighted by the odds ratio at each SNP. PLINK gives the mean risk allele score for each individual, that is, the risk allele score is divided by the number of SNPs for which there are data in that individual.

3.2.4 Controlling for minor allele frequency and population stratification

For each pruned *cis*-eQTL SNP list in the primary analysis, and for significant results from the secondary analyses, I calculated the mean and standard deviation of MAF for pruned SNPs in the target association datasets (full ISC dataset for MGS/ISC analyses, the target split ISC dataset for the split ISC analyses). These MAF values were then compared using t-tests. This was necessary in case the process of ranking the SNPs by their most significant eQTL p-value introduced systematic differences in allele frequency between high and low eQTL SNP sets. These differences in allele frequency could affect the results of the analyses, as alleles which affect complex traits can have different distribution of MAF to the majority of SNPs. For example, SNPs which affect human height tend to have a lower MAF than typical SNPs (37). In the polygenic score analysis performed by the ISC, a disproportionately large proportion of the signal was carried by alleles with relatively high MAF (29), and therefore if selection of SNPs by eQTL status resulted in groups with very different MAFs, this could in principle generate spurious results.

Although the regression analyses described below (see Section 3.2.6) include population of origin as a covariate, it is possible that population stratification may have complex effects on our data that cannot be corrected for with a linear covariate. Population stratification can inflate or deflate the significance of the link between phenotype and polygenic score, if members of a population are over-represented in the same phenotypic category in both the training and target datasets (29). Although both the top and bottom eQTL SNP lists use the same training and target datasets, it is possible that if a SNP list contains particularly heavily stratified SNPs it will be more susceptible to this effect. This could create false positive differences between top and bottom eQTL SNP lists.

Hence, to examine whether the results might be influenced by population stratification, I obtained F_{ST} values for each SNP based upon the ISC sample that are part of a previous study (123, 124). F_{ST} is a measure of population stratification and is based upon the sequence similarity of members of a subpopulation, compared to their similarity with the population as a whole. In a heavily stratified population, members of the subpopulations will be much more similar to each other than to the whole population, leading to a high F_{ST} score. Mean F_{ST} scores were determined for each SNP list, and compared using t-tests.

SNPs with as close a F_{ST} value as possible to each SNP in the smaller of the two SNP lists (top or bottom eQTL) were extracted without replacement from the larger SNP list to create eQTL sets matched for F_{ST}. A small number of SNPs could not be matched (arbitrarily defined as those where the closest match had a F_{ST} value over 0.0005 different) and were removed from the analysis. This definition of unmatchable SNPs was chosen because it created pairs of SNP lists with the same number of SNPs and the same mean and standard deviation of F_{ST}, while only excluding a small proportion of the SNP lists. As above, I calculated risk allele scores for each F_{ST} matched SNP list, and compared the scores for top and bottom eQTL SNP lists using logistic regression.

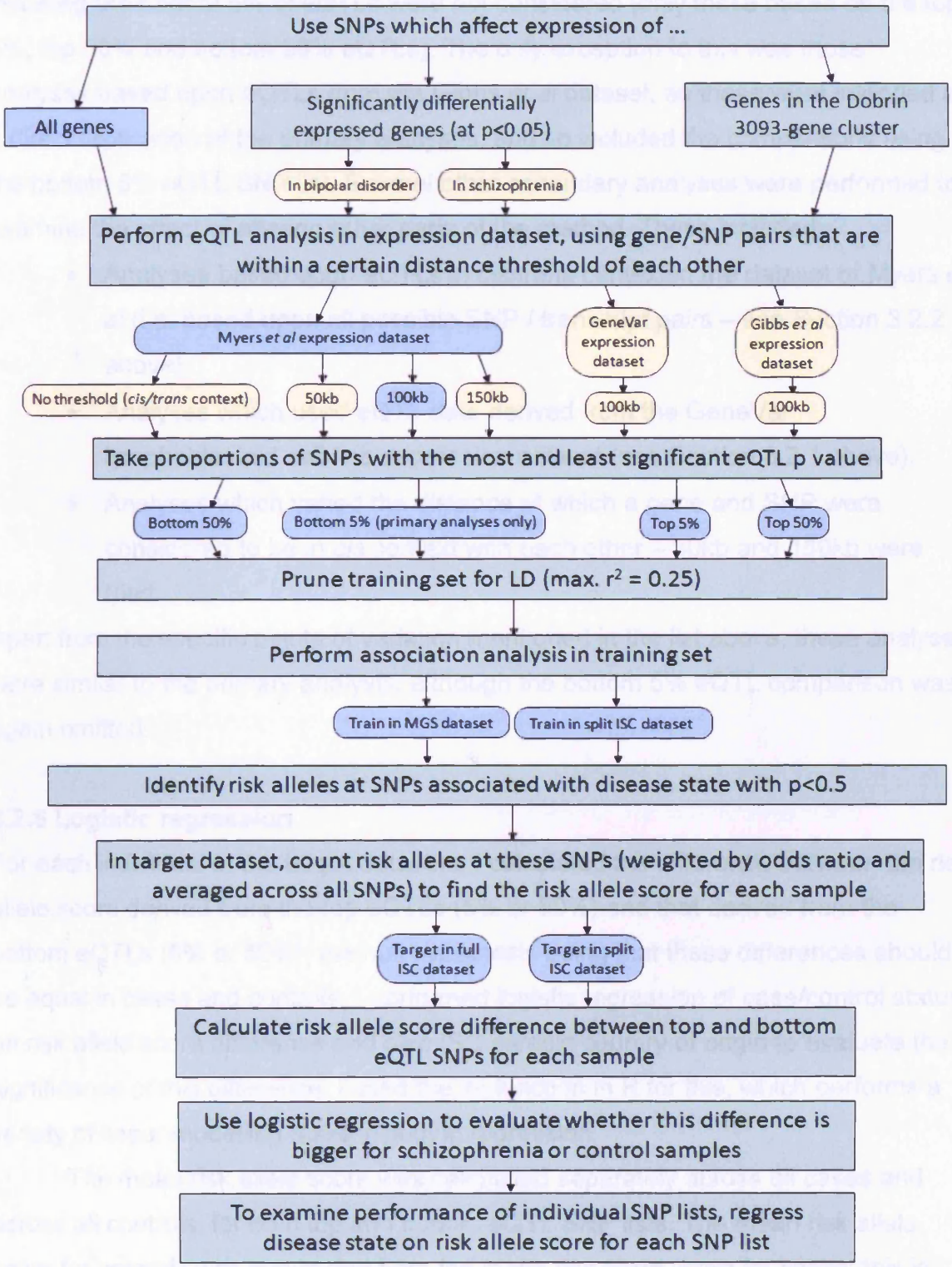
3.2.5 Secondary analyses

Top and bottom eQTL SNP lists were also constructed based upon eQTLs calculated using a subset of gene transcripts (see Figure 3.1 for an overview of the primary and secondary analyses). These were:

- Genes differentially expressed in schizophrenia or bipolar disorder according to the Stanley Medical Research Institute Online Genomics database, at $p<0.05$.
- Genes from the Dobrin 3093-gene cluster, found by using the clustering method memISA on the Dobrin expression dataset (see Chapters 2 and 4 for more details of this cluster).

In each case, only SNP / transcript comparisons involving transcripts from the subgroup were considered. The subgroup analyses were otherwise similar to the primary analysis. They were performed in *cis* context (100kb *cis* window only), using the dataset of Myers *et al* to define eQTLs and the MGS, ISC and split ISC datasets

Figure 3.1 Overview of the process of eQTL production and polygenic score analysis



Options coloured in blue form part of the primary analysis – all possible combinations of these options are used. Options coloured in yellow are secondary analyses.

for training and targeting. To reduce the multiplicity of similar analyses, comparisons involving the bottom 5% of eQTLs were not considered (only those based on the top 5%, top 50% and bottom 50% eQTLs). The only exception to this was those analyses based upon eQTLs from the Gibbs *et al* dataset, as these were intended as a direct replication of the primary analyses, and so included the comparisons using the bottom 5% eQTL SNP list. Several other secondary analyses were performed to examine the effect of altering other parts of the method. These included:

- Analyses based upon eQTLs in *cis/trans* context in the dataset of Myers *et al* (i.e. based upon all possible SNP / transcript pairs – see Section 3.2.2 above).
- Analyses which used eQTL data derived from the GeneVar lymphoblastoid cell line expression dataset (see Section 3.2.1 above).
- Analyses which varied the distance at which a gene and SNP were considered to be in *cis* context with each other – 50kb and 150kb were tried.

Apart from the specific points of variation mentioned in the list above, these analyses were similar to the primary analysis, although the bottom 5% eQTL comparison was again omitted.

3.2.6 Logistic regression

For each individual in the target datasets, I calculated the difference between the risk allele score derived from the top eQTLs (5% or 50%) and that derived from the bottom eQTLs (5% or 50%), the null hypothesis being that these differences should be equal in cases and controls. I performed logistic regression of case/control status on risk allele score difference and also ISC sample country of origin to evaluate the significance of this difference. I used the lm function in R for this, which performs a variety of linear modelling tasks including regression.

The mean risk allele score was calculated separately across all cases and across all controls, for both top and bottom eQTL SNP lists. The mean risk allele score for controls was subtracted from the mean risk allele score for cases; this is subsequently referred to as risk allele score case/control difference. As a measure of effect size in each comparison between top and bottom eQTL SNP lists, the risk allele score case/control difference for the bottom eQTL SNP list was subtracted

from the risk allele score case/control difference for the top eQTL SNP list. This is subsequently referred to as the difference in risk allele score disparity – a positive value indicates that the top eQTL SNP list predicts schizophrenia affected status better than the bottom eQTL SNP list, while a negative value indicates the reverse.

Logistic regression of disease status on risk allele score was also calculated to determine how significantly each individual SNP list predicted disease status. For each SNP list, I also calculated the Nagelkerke pseudo-R^2 (125), which is a measure of how well the risk allele score predicts schizophrenia disease state, by subtracting the R^2 of the regression without the risk allele score term included from the R^2 of the regression with the risk allele score term included.

3.3 Results

3.3.1 Primary analysis – *cis* eQTLs derived from the Myers *et al* brain expression dataset

When I defined risk alleles using half of the ISC sample as the training set (Table 3.1, rows 1-3), the difference in the risk allele scores between the top and bottom *cis*-eQTLs was greater in the cases than in the controls for all tested pairs (reflected in positive values in the 'difference in risk allele score disparity' column). Moreover, for all tests, this difference was significant (reflected in the columns labelled 'regression p-value'). This is consistent with the hypothesis that schizophrenia susceptibility alleles are enriched among *cis*-eQTLs. Qualitatively similar findings were observed when the risk alleles were defined from the MGS European dataset (which is entirely independent of the ISC dataset) in that the differences in the scores between the top and bottom *cis*-eQTLs were greater in the cases than in the controls (Table 3.1, rows 4-6). Although the comparisons are not independent (top 5% versus bottom 50% clearly overlaps with top 50% v bottom 50%), two *cis* tests are significant if the Bonferroni correction for three independent replication tests is used.

For all but one SNP list, the regression of schizophrenia affected status upon risk allele score was significant (Table 3.2, 'regression p-value' column)). This shows that the significant results in Table 3.1 are not due to the bottom eQTL SNP lists failing to significantly predict schizophrenia affected status, rather that the top eQTLs

perform better. However, the bottom 5% eQTL analyses are less significant. Their SNP counts are 20 to 25% of the size of the bottom 50% eQTL analyses, but their significance is several orders of magnitude less (Table 3.2, rows 4 and 8, columns 'SNP count' and 'Regression p-value'). The Nagelkerke pseudo-R^2 values of the bottom 5% SNP lists, which indicate the percentage of variation in disease state explained by the polygenic score value, are also low. In both the MGS/ISC and split ISC results, the bottom 5% eQTL SNP list is considerably bigger than the top 5% eQTL list, but has a much smaller Nagelkerke pseudo-R^2 (Table 3.2, rows 2, 4, 6 and 8, columns 'SNP count' and 'Nagelkerke pseudo-R^2').

Table 3.1. Difference in risk allele score case-control disparity between top and bottom *cis* Myers *et al* brain eQTL SNP lists (100kb *cis* window)

Row	Expression dataset	Gene subgroup	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Regression p-value	Difference in risk allele score disparity
1	Myers *et al*	All genes	Split ISC	Split ISC	50	50	0.014	2.56E-05
2	Myers *et al*	All genes	Split ISC	Split ISC	5	50	0.014	8.15E-05
3	Myers *et al*	All genes	Split ISC	Split ISC	5	5	0.012	9.63E-05
4	Myers *et al*	All genes	MGS	ISC	50	50	0.298	1.63E-05
5	Myers *et al*	All genes	MGS	ISC	5	50	0.002	9.27E-05
6	Myers *et al*	All genes	MGS	ISC	5	5	0.003	8.57E-05

Table 3.2. Regression of affected status on risk allele score, primary analyses (*cis* context, 100kb *cis* window)

Row	Expression dataset	Training dataset	Target dataset	Top / bottom eQTL percentage	Nagelkerke pseudo-R^2	Regression p-value	Case risk allele score	Control risk allele score	Case/ control risk allele score difference	SNP count
1	Myers *et al*	Split ISC	Split ISC	Top 50%	1.59	1.43E-14	0.04748	0.04743	5.36E-05	10805
2	Myers *et al*	Split ISC	Split ISC	Top 5%	0.47	2.09E-05	0.04646	0.04635	1.10E-04	1285
3	Myers *et al*	Split ISC	Split ISC	Bottom 50%	0.63	9.22E-07	0.04849	0.04846	2.86E-05	10967
4	Myers *et al*	Split ISC	Split ISC	Bottom 5%	0.04	1.12E-01	0.04579	0.04578	1.38E-05	2033
5	Myers *et al*	MGS	ISC	Top 50%	0.50	8.78E-10	0.04709	0.04706	3.08E-05	3903
6	Myers *et al*	MGS	ISC	Top 5%	0.30	1.47E-06	0.04565	0.04554	1.07E-04	435
7	Myers *et al*	MGS	ISC	Bottom 50%	0.30	1.63E-06	0.04709	0.04707	1.45E-05	4037
8	Myers *et al*	MGS	ISC	Bottom 5%	0.11	2.70E-03	0.04469	0.04467	2.15E-05	1154

3.3.2 Replication analysis – *cis* eQTLs derived from Gibbs *et al* brain dataset

In the dataset of Gibbs *et al*, there were no significant differences between top and bottom eQTL SNP lists (Table 3.3, rows 1-3, column 'Regression p-value') in the split ISC analyses, although for all comparisons, the top eQTL SNP list produced a higher risk allele score case/control difference than the bottom eQTL SNP list (Table 3.3, rows 1-3, column 'Difference in risk allele score disparity'). However, for the MGS/ISC analysis, the top eQTL SNPs were highly significantly better predictors of affected status than the bottom SNP lists (Table 3.3, rows 4-6, columns 'Regression p-value' and 'Difference in risk allele score disparity'). These would remain highly significant even after Bonferroni correction for three tests in two test samples.

Table 3.3. Difference in risk allele score case-control disparity between top and bottom *cis* Gibbs *et al* eQTL SNP lists (100kb *cis* window)

Row	Expression dataset	Gene subgroup	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Regression p-value	Difference in risk allele score disparity
1	Gibbs *et al*	All genes	Split ISC	Split ISC	50	50	3.99E-01	7.82E-06
2	Gibbs *et al*	All genes	Split ISC	Split ISC	5	50	1.70E-01	3.10E-05
3	Gibbs *et al*	All genes	Split ISC	Split ISC	5	5	4.40E-01	2.16E-05
4	Gibbs *et al*	All genes	MGS	ISC	50	50	1.76E-04	2.71E-05
5	Gibbs *et al*	All genes	MGS	ISC	5	50	5.02E-05	6.71E-05
6	Gibbs *et al*	All genes	MGS	ISC	5	5	9.51E-04	6.68E-05

3.3.3 Secondary analysis – *cis/trans* Myers *et al* brain eQTLs

There were no consistent patterns or significant differences in the risk allele score case/control disparity between the top and bottom *cis/trans* eQTL SNP list (Table 3.4, rows 1-6).

Table 3.4. Difference in risk allele score case-control disparity between top and bottom *cis/trans* Myers *et al* brain eQTL SNP lists

Row	Expression dataset	Gene subgroup	Context	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Regression p-value	Difference in risk allele score disparity
1	Myers *et al*	All genes	*cis/trans*	Split ISC	Split ISC	50	50	0.252	-7.33E-06
2	Myers *et al*	All genes	*cis/trans*	Split ISC	Split ISC	5	50	0.518	8.91E-06
3	Myers *et al*	All genes	*cis/trans*	Split ISC	Split ISC	5	5	0.687	1.54E-05
4	Myers *et al*	All genes	*cis/trans*	MGS	ISC	50	50	0.715	-2.72E-06
5	Myers *et al*	All genes	*cis/trans*	MGS	ISC	5	50	0.747	-2.02E-06
6	Myers *et al*	All genes	*cis/trans*	MGS	ISC	5	5	0.323	-5.59E-06

3.3.4 Secondary analysis – *cis* lymphoblast cell line eQTLs

No *cis* eQTL analysis based on SNPs derived from the GeneVar lymphoblastoid cell line expression dataset was significant (Table 3.5, rows 1-4).

Table 3.5. Difference in risk allele score case-control disparity between top and bottom *cis* lymphoblastoid cell line eQTL SNP lists

Row	Expression dataset	Gene subgroup	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Regression p-value	Difference in risk allele score disparity
1	GeneVar	All genes	Split ISC	Split ISC	50	50	0.535	-2.55E-06
2	GeneVar	All genes	Split ISC	Split ISC	5	50	0.735	1.05E-05
3	GeneVar	All genes	MGS	ISC	50	50	0.868	-1.96E-06
4	GeneVar	All genes	MGS	ISC	5	50	0.940	3.11E-07

3.3.5 Secondary analysis – *cis* brain eQTLs based upon genes differentially expressed in schizophrenia or bipolar disorder

There were no significant results in the analyses using genes differentially expressed in schizophrenia (Table 3.6, rows 1-4, 'regression p-value' column) or bipolar disorder (Table 3.6, rows 5-8). One possible reason for this is the small size of some of the SNP lists. For example, the top 5% eQTL SNP list for genes differentially expressed in schizophrenia contains only 46 SNPs (see row 2 of Table S6 in Appendix B, column 'SNP count').

Table 3.6. Difference in risk allele score case-control disparity between *cis* brain eQTL SNP lists based on genes differentially expressed in schizophrenia or bipolar disorder

Row	Expression dataset	Gene subgroup	Context	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Regression p-value	Difference in risk allele score disparity
1	Myers *et al*	Schizophrenia differential expression	*cis*	Split ISC	Split ISC	50	50	0.141	8.72E-05
2	Myers *et al*	Schizophrenia differential expression	*cis*	Split ISC	Split ISC	5	50	0.653	0.000105
3	Myers *et al*	Schizophrenia differential expression	*cis*	MGS	ISC	50	50	0.532	-1.03E-05
4	Myers *et al*	Schizophrenia differential expression	*cis*	MGS	ISC	5	50	0.569	2.85E-06
5	Myers *et al*	Bipolar disorder differential expression	*cis*	Split ISC	Split ISC	50	50	0.891	-1.44E-05
6	Myers *et al*	Bipolar disorder differential expression	*cis*	Split ISC	Split ISC	5	50	0.256	0.000121
7	Myers *et al*	Bipolar disorder differential expression	*cis*	MGS	ISC	50	50	0.693	1.63E-05
8	Myers *et al*	Bipolar disorder differential expression	*cis*	MGS	ISC	5	50	0.152	-4.86E-05

3.3.6 Secondary analyses based upon alternate *cis* windows

Using varying *cis* ranges had little effect when the split ISC dataset was used for training and targeting – the results for a range of 50kb and 150kb were very similar to the results for 100kb from the primary analysis (Table 3.7, rows 1-6). However, in the analysis trained in the MGS and targeted in the ISC, the 50kb and 150kb results no longer had a nominally significant positive difference in risk allele score disparity (Table 3.7, rows 9-12).

Table 3.7. Difference in risk allele score case-control disparity between top and bottom eQTL SNP lists, secondary analyses with eQTLs based upon all genes (*cis* results with variant *cis* windows, 100kb results included for comparison)

Row	Expression dataset	Gene subgroup	Training dataset	Target dataset	*cis* window	Top eQTL (%)	Bottom eQTL (%)	Regression p-value	Difference in risk allele score disparity
1	Myers *et al*	All genes	Split ISC	Split ISC	100kb	50	50	0.014	2.56E-05
2	Myers *et al*	All genes	Split ISC	Split ISC	100kb	5	50	0.014	8.15E-05
3	Myers *et al*	All genes	Split ISC	Split ISC	150kb	50	50	0.0171	2.67E-05
4	Myers *et al*	All genes	Split ISC	Split ISC	150kb	5	50	0.0131	7.23E-05
5	Myers *et al*	All genes	Split ISC	Split ISC	50kb	50	50	0.0556	1.85E-05
6	Myers *et al*	All genes	Split ISC	Split ISC	50kb	5	50	0.0140	7.63E-05
7	Myers *et al*	All genes	MGS	ISC	100kb	50	50	0.298	1.63E-05
8	Myers *et al*	All genes	MGS	ISC	100kb	5	50	0.002	9.27E-05
9	Myers *et al*	All genes	MGS	ISC	150kb	50	50	0.117	-1.44E-05
10	Myers *et al*	All genes	MGS	ISC	150kb	5	50	0.202	-1.15E-05
11	Myers *et al*	All genes	MGS	ISC	50kb	50	50	0.562	1.23E-05
12	Myers *et al*	All genes	MGS	ISC	50kb	5	50	0.836	-2.04E-05

3.3.7 Secondary analyses using eQTLs based upon expression cluster genes (*cis* context)

In the analyses based upon eQTLs from genes in the Dobrin 3093 coexpression cluster, there was one nominally significant result. This was not in the direction predicted by the hypothesis, since the bottom eQTL SNP list outperformed the top eQTL SNP when the top 5% and bottom 50% lists were compared (p=0.02, Table 3.8, row 4). However, in the split ISC dataset, no such effect was observed, with the (non-significant) trend being for top eQTLs to outperform bottom.

Table 3.8. Difference in risk allele score case-control disparity between top and bottom eQTL SNP lists, secondary analyses with eQTLs based upon expression cluster genes (*cis* context with a *cis* window of 100kb)

Row	Expression dataset	Gene subgroup	Training dataset	Target dataset	Top eQTL SNP list (%)	Bottom eQTL SNP list (%)	Regression p-value	Difference in risk allele score disparity
1	Myers *et al*	Dobrin 3093 cluster	Split ISC	Split ISC	50	50	0.567	1.83E-05
2	Myers *et al*	Dobrin 3093 cluster	Split ISC	Split ISC	5	50	0.869	3.52E-05
3	Myers *et al*	Dobrin 3093 cluster	MGS	ISC	50	50	0.989	3.92E-06
4	Myers *et al*	Dobrin 3093 cluster	MGS	ISC	5	50	0.027	-6.70E-05

3.3.8 Minor allele frequency and population stratification, primary analysis

In the primary analyses, of the 5 tests in which the top *cis*-eQTLs were significantly better at discriminating case-control status, the mean MAF was slightly but significantly higher in 2 of the top *cis*-eQTLs (Table 3.9, rows 2 and 5, column 'T-test significance of MAF difference'), whereas for the other three tests, any trends were for a lower MAF in the top *cis*-eQTLs set (Table 3.9, rows 1,2 and 6). (This suggests that the findings are unlikely to be due to differences in MAF between the sets.

In each analysis, the top *cis*-eQTL set had significantly higher mean F_{ST} than the bottom eQTL SNP lists (Table 3.9, column 'T-test significance of F_{ST} difference'), indicating that my analysis might be confounded by enhanced stratification in the top *cis*-eQTL set. Also, the most significant F_{ST} difference occurred in the MGS/ISC top 50% versus bottom 50% result (Table 3.9, row 4). In that analysis, there was no significant difference in risk allele score case/control disparity between top and bottom eQTL SNP lists, suggesting F_{ST} difference alone of the magnitudes being observed, sufficient to cause a significant result. This is as expected since in order for this bias to affect my analysis, the MGS sample and the ISC samples would have to be ascertained in such a manner that the same alleles (not just the same loci) are similarly biased towards overrepresentation in cases in each dataset.

Table 3.9. Regression of affected status on difference in risk allele score case-control disparity between top and bottom eQTL SNP lists, plus t-tests of difference in MAF and F_{ST} – primary analysis SNP lists not matched for F_{ST}

Row	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Difference in risk allele score case-control disparity	Reg. p-value	Mean top eQTL MAF	Mean bottom eQTL MAF	T-test sig. of MAF diff.	Mean top eQTL F_{ST}	Mean bottom eQTL F_{ST}	T-test sig. of F_{ST} diff.
1	Split ISC	Split ISC	50	50	2.56E-05	0.014	0.227	0.228	0.948	0.0027	0.0026	0.006
2	Split ISC	Split ISC	5	50	8.15E-05	0.014	0.241	0.228	0.001	0.0028	0.0026	0.019
3	Split ISC	Split ISC	5	5	9.63E-05	0.012	0.241	0.246	0.268	0.0028	0.0026	0.020
4	MGS	ISC	50	50	1.63E-05	0.298	0.227	0.228	0.594	0.0027	0.0026	0.006
5	MGS	ISC	5	50	9.27E-05	0.002	0.238	0.228	0.047	0.0028	0.0026	0.019
6	MGS	ISC	5	5	8.57E-05	0.003	0.238	0.249	0.054	0.0028	0.0026	0.020

Although I do not consider F_{ST} difference a likely explanation (29) for the observations of better performance of top eQTLs in the primary analysis, to evaluate this further, all primary analyses were repeated using F_{ST} matched SNP sets. After matching, there were no significant differences in mean F_{ST} between pairs of comparator groups (Table 3.10, column 'T-test significance of F_{ST} difference'). Nevertheless, for two of the three analyses in the split ISC datasets, the top *cis*-eQTLs significantly discriminated better between cases and controls than the bottom *cis*-eQTLs (Table 3.10, rows 2-3), and I obtained significant replication for both findings when the MGS sample was used as the training set (Table 3.10, rows 5-6). Moreover, for two of the F_{ST} matched analyses that were significant, the top *cis*-eQTL sets had lower MAF than the bottom set (Table 3.10, rows 3 and 6), and for two of the results, the top group had higher MAF (Table 3.10, rows 2 and 5). I therefore conclude that the primary analysis findings are not driven by systematic biases in these variables.

Table 3.10. Regression of affected status on difference in risk allele score case-control disparity between top and bottom eQTL SNP lists, plus t-tests of difference in MAF and F_{ST} – primary analysis SNP lists matched for F_{ST}

Row	Trained in	Targeted in	Top eQTL (%)	Bottom eQTL (%)	Difference in risk allele score case-control disparity	Reg. p-value	Mean top eQTL MAF	Mean bottom eQTL MAF	T-test sig. of MAF diff.	Mean top eQTL F_{ST}	Mean bottom eQTL F_{ST}	T-test sig. of F_{ST} diff.
1	Split ISC	Split ISC	50	50	2.33E-05	0.076	0.227	0.227	0.999	0.0026	0.0026	0.565
2	Split ISC	Split ISC	5	50	1.18E-04	0.006	0.241	0.227	0.007	0.0028	0.0028	0.809
3	Split ISC	Split ISC	5	5	1.55E-04	0.005	0.242	0.244	0.603	0.0026	0.0026	0.573
4	MGS	ISC	50	50	5.18E-06	0.544	0.227	0.227	0.999	0.0027	0.0027	0.814
5	MGS	ISC	5	50	5.61E-05	0.022	0.241	0.227	0.007	0.0027	0.0027	0.968
6	MGS	ISC	5	5	5.06E-05	0.019	0.242	0.244	0.626	0.0027	0.0027	0.833

3.3.9 Minor allele frequency and population stratification, secondary analysis

I also examined the mean F_{ST} and MAF for the significant run based upon the Dobrin 3093 gene subgroup (Table 3.8, row 4). Both MAF and F_{ST} were significantly different between top and bottom eQTL SNP lists (Table 3.11, row 1). A SNP list matched for F_{ST} was constructed, which no longer had a significant regression p-value (Table 3.11, row 2, 'Regression p-value' column). However, the magnitude of the difference in risk allele score case/control disparity is similar to the non-matched result, suggesting this lack of significance may simply be due to the removal of SNPs in the matching process rather than indicative that MAF and/or F_{ST} were responsible for the nominally significant observed effect.

Table 3.11. Regression of affected status on difference in risk allele score case-control disparity between top and bottom eQTL SNP lists, plus t-tests of difference in MAF and F_{ST} – secondary analysis SNP lists based upon Dobrin 3093 coexpression cluster genes

Row	Trained in	Targeted in	Matched on	Top eQTL (%)	Bottom eQTL (%)	Difference in risk allele score case-control disparity	Reg. p-value	Mean top eQTL MAF	Mean bottom eQTL MAF	T-test sig. of MAF diff.	Mean top eQTL F_{ST}	Mean bottom eQTL F_{ST}	T-test sig. of F_{ST} diff.
1	MGS	ISC	Not matched	5	50	-6.70E-05	0.027	0.235	0.220	9.0E-03	0.0025	0.0027	0.033
2	MGS	ISC	F_{ST}	5	50	-6.99E-05	0.167	0.235	0.205	8.63E-05	0.0025	0.0026	0.515

3.4 Discussion

3.4.1 Relevance of genetic regulation of expression to schizophrenia aetiology

To date, only a small proportion of genetic susceptibility to schizophrenia, or indeed any psychiatric disorder, has been explained by robustly associated DNA variants. Moreover, in no case has the functional effect of a DNA variant responsible for a robust schizophrenia association been determined. It follows that the basic mechanisms by which genetic variation contribute to this disorder are unknown. One leading hypothesis is that a substantial amount of genetic risk is conferred by

common alleles that influence gene expression, that is, common eQTLs. However, while the existence of many common schizophrenia risk alleles has been demonstrated (29), there is no evidence to support the hypothesis that any of these influence gene expression. In the light of a recent rekindling of interest in the hypothesis that genetic risk for the disorder is likely to be attributable to rare variants of major effect, which by analogy with Mendelian disorders are likely to be dominated by mutations that change the protein coding sequences of genes, the demonstration or refutation of a contribution from eQTLs is of practical importance for several reasons.

The search for functional variants underpinning disease associations observed in GWAS studies in general is proving to be far from a trivial endeavour. Although it is relatively simple to scan the exonic sequences of individual genes for common (and even fairly rare) non-synonymous variants, the process of scanning the full genomic context of a gene for potential *cis*-eQTLs, and then demonstrating that those variants impact on expression in a disease relevant manner remains difficult. Comprehensive variant discovery is increasingly being facilitated by high capacity sequencing technology, but the demonstration of relevant functionality is not. In order to justify those endeavours, it is therefore important to demonstrate that effects on gene expression are in fact relevant mechanisms underpinning the influence of common susceptibility variants. As discussed above, the use of gene expression data to support less than fully convincing genetic associations, or in other words, to assign higher prior probability to particular variants, requires evidence that *cis*-eQTLs do in fact have a higher probability of being truly associated with disease than random sets of alleles.

Even if risk variants are enriched for common *cis*-eQTLs, it cannot be taken for granted that control adult brain tissues, far less other sources of mRNA, are suitable substrates for generating eQTLs for disorders like schizophrenia whose presumed origins are developmental (49).

3.4.2 *cis* eQTL SNPs predict disease state better than non-eQTL SNPs

To undertake the first large scale test of the involvement of eQTLs in schizophrenia, I exploited a recent finding of the ISC that sets of marginally associated alleles derived from large GWAS datasets contain large numbers of true schizophrenia-

associated alleles. Using two independent GWAS datasets I demonstrated that among the variants selected for marginal association to schizophrenia, those that additionally show evidence for being *cis*-eQTLs predict affection status better than those variants showing no evidence for being *cis*-eQTLs. In other words, I show for the first time that schizophrenia risk alleles are indeed enriched for eQTLs. As expected from the ISC study, no set of SNPs explained more than a small fraction of the variance in disease risk (Table 3.2, column 'Nagelkerke pseudo-R^{2}'), although more comprehensive genome coverage may explain a much higher proportion of this variance (29).

This finding was further reinforced by the results based upon the dataset of Gibbs *et al*. Although the results trained and targeted in the split ISC dataset did not reach significance, they did show a trend toward the top eQTL SNP lists predicting affected status better than the bottom eQTL SNP lists. However, in the analyses trained in the MGS and targeted in the ISC dataset, the top eQTL SNP lists predicted disease state better than the bottom eQTL SNP lists. This finding was highly significant, would survive correction for multiple testing of the primary hypothesis, and therefore provides a replication of my results using the Myers *et al* eQTLs.

In contrast to the findings with *cis*-eQTLs, SNPs, classified on the basis of potential *trans* effects were not superior at predicting schizophrenia disease status. This may be because the much greater multiple testing burden inherent to *trans* eQTL analysis means a smaller proportion of the top rated *trans*-eQTLs are true positives. The weaker performance of *cis/trans* eQTLs may also reflect a lesser importance of *trans* eQTLs (112). The work presented here does not distinguish between these possibilities.

While top sets of *cis*-eQTLs perform better than bottom sets, it is evident (Table 3.2, column 'Regression p-value') that even the latter significantly, often highly significantly, predict affected status. This might be because a substantial part of the true association signal is not related to variants that alter gene expression. Alternatively, though *a priori*, I do not consider it particularly likely it may be that virtually all true common associations derive from eQTLs, but that many of these are incorrectly classified as such in this study. The samples from which I derived eQTL status are relatively small in GWAS terms, and therefore have limited power to

identify weak eQTLs. Moreover, the already limited power will be further constrained by variance introduced by the many well known confounders that plague the use of *post mortem* expression datasets, such as post mortem interval, brain pH and agonal factors (126). Both factors are likely to result in eQTL classification errors.

Potentially pointing to an important impact of eQTL misclassification, comparisons of the most extreme *cis*-eQTL categories (top and bottom 5% sets) revealed considerable differences in the ability of those groups to discriminate case and control status in the primary analysis (Tables 3.1 and 3.2). Thus, the risk allele score differences between cases and controls were about 10 times greater for the top 5% of *cis*-eQTLs (Table 3.2, rows 2 and 6), and were 3-4 orders of magnitude more significant, than for the bottom 5% of *cis*-eQTLs (Table 3.2, rows 4 and 8). The former also had better predictive power as indicated by a larger Nagelkerke R^2, despite greater numbers of SNPs in the bottom 5% group. Indeed the bottom 5% of *cis*-eQTLs were either not significant predictors at all (trained in the split ISC dataset, Table 3.2, row 4) or the statistical significance of prediction was relatively modest (trained in the MGS dataset, Table 3.2, row 8). Assuming the extreme top and bottom *cis*-eQTL groups contain SNPs that are least likely to be misclassified, I postulate that the proportion of the polygenic signal captured by eQTLs might be greatly enhanced by more precise delineation of eQTL status.

3.4.3 Secondary analyses

In contrast to the primary analyses, the majority of the secondary analyses did not show any significant results. The analyses based upon Myers *et al cis/trans* eQTLs (Table 3.4), the analyses using GeneVar *cis/trans* eQTLs (Table 3.5) and the analyses based upon Myers *et al* eQTLs derived from transcripts with evidence for differential expression in bipolar disorder or schizophrenia (Table 3.6) all showed no significant differences between top and bottom eQTL SNP lists. The analyses using eQTLs based upon transcripts present in the Dobrin 3093 expression cluster showed one significant result, where the bottom 50% eQTL SNP list had a higher risk allele score case/control difference than the top 5% eQTL SNP list when the MGS and ISC datasets were used for training and targeting (Table 3.8, row 4). However, this result

did not replicate when the split ISC dataset was used for training and targeting (Table 3.8, row 2).

The results based upon Myers *cis* eQTLs using *cis* windows of 50kb and 150kb agreed with the primary analysis when the split ISC dataset was used for training and targeting, finding the top eQTL SNP lists produced significantly higher risk allele score case/control differences than the bottom eQTL SNP lists (Table 3.7, rows 3 to 6). However, in the MGS/ISC analyses, the 50kb and 150kb results are all non-significant, and the bottom eQTL SNP lists produced higher risk allele score case/control differences than the top eQTL SNP lists (Table 3.7, rows 9 to 12).

This is a surprising shift, considering that the MGS/ISC top 5% versus bottom 50% result in the primary analysis showed a significantly higher risk allele score case/control difference for the top 5% eQTL SNP list, and that changing the *cis* window is a relatively minor change to the method. The shift also occurs regardless of whether the *cis* window is decreased (50kb) or increased (150kb) relative to the primary analysis. Part of the explanation may be the nature of the pruning step, which when given a pair of SNPs in high linkage disequilibrium arbitrarily selects one to be excluded. This means that a small difference between SNP lists before pruning can be amplified into a larger difference in SNP complement after pruning.

Another reason for the shift could be the relatively small effect sizes of the analyses in general, compared to the values of the risk allele scores. The former are less than, or close to, $1e^{-4}$ (Tables 3.1 and 3.3-3.8, column 'Difference in risk allele score disparity'), while the risk allele scores for cases or controls are typically in the range 0.04 to 0.05 (Table 3.2, columns 'Case risk allele score' and 'Control risk allele score'). This means that a relatively small percentage change in any one of four values (top eQTL case risk allele score, top eQTL control risk allele score, bottom eQTL case risk allele score, bottom eQTL control risk allele score) can produce a large percentage change in the difference in risk allele score disparity.

3.4.4 Secondary analyses – GeneVar expression dataset

The results based upon putative eQTLs from the GeneVar lymphocyte dataset showed no significant differences between top and bottom eQTL SNP lists (Table

3.5, rows 1-4). Furthermore, there was no trend toward either top or bottom eQTL SNP lists having higher risk allele score case/control differences – 2 of the results favoured the top eQTL SNP list (Table 3.5, rows 2 and 4), 2 favoured the bottom eQTL SNP list (Table 3.5, rows 1 and 3). Thus, my analysis provides no support for the use of eQTLs derived from lymphoblastoid cell lines in molecular genetic studies of schizophrenia. This loosely agrees with the findings of Rollins *et al*, who found that only 22.9% of transcripts are expressed at the same level in cerebellar cortex and peripheral blood mononuclear cells (127).

However, there are several important caveats. The GeneVar dataset only contains expression data for 55 samples of European origin, so it will have less power to calculate eQTLs and sort SNPs into top and bottom categories than the datasets of Myers *et al* (163 samples) or Gibbs *et al* (133 samples). Also, additional quality control or analysis could improve the relevance of the GeneVar dataset to brain function and schizophrenia aetiology. For example, setting a minimum mean expression level for transcripts could exclude those transcripts that are not expressed in lymphoblastoid cell lines. Also, restricting the transcripts to those known to correlate well between blood and brain may be a useful step.

3.4.5 Comparisons with other studies

Schadt *et al* used eQTL data from human liver to provide evidence to prioritise candidate type I diabetes and coronary artery disease susceptibility genes found through GWAS studies (122). Unlike the study here, they did not use a method which combines information from multiple SNPs into a single score. They found that eQTL and expression evidence suggested that RPS26, not ERBB3 as previously thought, was responsible for a novel type I diabetes association signal on chromosome 12q3. They demonstrated that RPS26 expression was significantly associated with the SNPs most strongly associated with affected status, and that RPS26 had higher expression than ERBB3 in pancreas. Using network methods, they also showed that SNPs near RPS26 affected the expression of another gene associated with diabetes, H2-Eb1. In coronary artery disease, similar methods showed that SORT1 and CELSR2 were responsible for an association signal on chromosome 1q13. Although their study was focused on individual genes and association signals, rather than aggregating the effects of multiple SNPs as I do

here, that study suggests that expression data can be relevant to complex traits and useful in explaining association data.

Nicolae *et al* (128) examined the hypothesis that eQTL SNPs might be more likely to be associated with disease traits compared with sets of randomly drawn SNPs with the same distribution of minor allele frequency. The phenotypes they investigated were Crohn's disease, rheumatoid arthritis, type 1 and 2 diabetes, hypertension, coronary artery disease and bipolar disorder. They used genotype and lymphoblastoid cell line expression data from the CEU (European descent) and YRI (Yoruban descent) populations of the GeneVar study (115, 129) to define eQTLs, and the Wellcome Trust Case-Control Consortium genotype data to determine whether eQTL SNPs are enriched for SNPs associated with disease traits. They found enrichment of eQTL SNPs for SNPs associated with Crohn's disease, type 1 diabetes and rheumatoid arthritis, but not the other disorders (type 2 diabetes, hypertension, coronary artery disease and bipolar disorder). One reason why their analysis revealing enrichment of eQTL SNPs for SNPs associated with autoimmune system related disorders might relate to the use of lymphoblastoid cell lines to define eQTL status as that cell line is more likely to be relevant to diseases where immune system behaviour plays a major role. As noted above, one study reported only a modest degree of convergence between expression in cerebellar cortex and peripheral blood mononuclear cells (127), suggesting as many might predict that such peripheral tissues may not be well suited for producing expression data with relevance to conditions which affect the brain. This is borne out by my study, where using the GeneVar and HapMap data to define eQTLs did not help to predict schizophrenia affected status through polygenic score analysis (Table 3.5, rows 1-4, see Section 3.4.4 above), although other explanations for the poor performance of that dataset are discussed above.

The converse analysis, examining whether disease-associated SNPs are likely to be enriched for eQTL SNPs, was also performed. Nicolae *et al* found that SNPs most associated with affected status in bipolar disorder were enriched for eQTLs. This was also true of SNPs associated with Crohn's disease, hypertension, rheumatoid arthritis and type I diabetes.

However, no p-value was reported for these findings, and it was based upon only the top 10,000 most associated SNPs for each disorder. The study also

examined 34 other groups of 10,000 SNPs each (ordered on strength of association), but did not account for the multiple testing burden this places upon their findings. In most cases, these 34 groups were not more significantly enriched for eQTLs than would be expected by chance.

The difference between this analysis and the analysis examining whether eQTL SNPs were enriched for disorder associated SNPs may be the use of alternative eQTL score thresholds. In this analysis, the associated SNPs were examined for enrichment with any SNPs with an eQTL score that exceeded a threshold of 3, while in the previous analysis only the 10,000 SNPs with the highest eQTL scores were used.

Nonetheless, the finding that SNPs highly significantly associated with bipolar disorder are enriched for lymphoblastoid cell line eQTL SNPs is potentially relevant to the study of psychiatric disease if it can be further verified, as it suggests that even expression data from non-brain tissues can have some relevance to psychiatric research.

3.4.6 Future work

The work presented in this chapter could be continued in a number of directions. Better eQTL classification could, in principle, be relatively simply achieved by 1) using larger human brain expression and SNP datasets 2) increasing the transcriptome coverage; the primary analysis here only incorporates 8361 probes representing only 25-30% of the protein encoding genes in the human genome (130) 3) using expression datasets derived from different brain regions and from different stages of human development.

More work could be done to accurately account for the multiple testing burden the large number of secondary analyses places upon significant results using eQTLs based on the Myers *et al* dataset. The results are heavily interdependent – top 5% eQTL SNP lists necessarily overlap with top 50% eQTL SNP lists, and many of the gene subgroups overlap with one another (especially the schizophrenia and bipolar disorder differential expression SNP lists). As a result, it will probably be necessary to use a permutation-based method to assess significance. One possibility would be

a permutation analysis based upon random partitions of SNPs in place of partitions based upon eQTL status.

Another issue worth considering is the possibility that some SNPs attained erroneously significant eQTLs because they lay within a sequence to which the expression probe binds. Changes in the genetic code in such a location will affect probe binding and so appear to alter expression levels. However, misclassifying eQTLs due to this will only add noise to the analysis, reducing power rather than increasing the risk of a false positive.

Lastly, the power of the analysis could be increased by using a larger association dataset for polygenic score analysis. The large meta-analysis forthcoming from the Psychiatric Genome-Wide Association Consortium would be ideal for this purpose.

3.4.7 Conclusions

In summary, I have undertaken the first large scale analysis of the hypothesis that schizophrenia risk is mediated in part by common DNA variants that influence gene expression. My results broadly support this hypothesis, although given failure to replicate the split ISC findings in the Gibbs data, additional replication in better powered samples will be required before this can be fully confidently accepted. Nevertheless, my data provide the first demonstration that gene expression studies in human adult brain can be informative for genetic investigations of schizophrenia. Larger eQTL datasets, representing different brain regions and developmental stages, will be required to maximally exploit the enhanced prior probability for cis-eQTLs as genetic susceptibility loci.

Chapter Four

Functional analysis using MetaCore of gene expression clusters derived from human brain

4.1 Introduction

4.1.1 Background

In Chapter 2, four different clustering methods were used on the Dobrin and MC66 brain gene expression datasets, and compared by how enriched the clusters they produced are for Gene Ontology (GO) terms. The most effective method of clustering was to combine the output from memISA and k-means clustering. Twenty-six gene clusters were found by using these two clustering methods on the Dobrin brain expression dataset, and another 25 from using these methods on the MC66 dataset. In this chapter, two clusters from these two sets enriched for genes associated with or differentially expressed in schizophrenia or bipolar disorder are identified and further analysed using enrichment analysis in MetaCore and network analysis.

4.1.2 Aetiology of schizophrenia and bipolar disorder

Schizophrenia and bipolar disorder are debilitating neuropsychiatric conditions. Schizophrenia is characterised by psychosis, disorganised thought, and blunted affect (though symptoms can vary considerably from case to case), while bipolar disorder is characterised by dramatic shifts in mood between depression and mania. Both disorders are serious public health problems, affecting around 0.5% (schizophrenia) and 1% to 1.5% (bipolar disorder) of the UK population (131, 132). They are also highly heritable, with estimates as high as 80% for both schizophrenia and bipolar disorder (55, 133).

Numerous hypotheses have been proposed for the genetic causes of schizophrenia and bipolar disorder. One of the first for schizophrenia was the dopamine hypothesis, suggesting that dysregulation of the neurotransmitter

dopamine was involved in schizophrenia. This was originally based upon the observation that dopamine levels in rat brains were affected by antipsychotic medication (134). Other schizophrenia hypotheses include the neurodevelopmenal hypothesis, which suggests that insults early in brain development can lead to schizophrenia in adolescence, and the myelin hypothesis, which is based upon the reduced presence of white matter in the brains of schizophrenia cases (49, 135). Other molecules among the many which have been suggested as playing a role in schizophrenia include glutamate, GABA, and oestrogen (136-138).

Fewer hypotheses have been suggested for bipolar disorder aetiology. The two primary theories are the serotonin hypothesis and the noradrenaline hypothesis (139). These hypotheses are thought to be complementary, rather than opposing.

Many of the schizophrenia hypotheses are also not mutually exclusive (e.g. neurodevelopmental insults can potentially lead to adult abnormalities in dopamine regulation). Furthermore, it is uncertain the extent to which these conditions are single syndromes with relatively unified causes, or whether subtypes with distinct causes and characteristics might be defined in the future.

A reason for focusing on both bipolar disorder and schizophrenia is that there is believed to be considerable overlap between them (140). In terms of symptoms, schizophrenia sufferers can display mood swings, while psychosis can be a feature of mania in some cases of bipolar disorder. Cases which display features of both bipolar disorder and schizophrenia without either predominating occur, a condition referred to as schizoaffective disorder. There is also some evidence that similar genetic causes underlie both schizophrenia and bipolar disorder (36).

The multiplicity and non-mutually exclusive nature of schizophrenia and bipolar disorder hypotheses can make them difficult to compare objectively. Large scale expression and GWAS datasets offer a data-driven, rather than hypothesis-driven, mode of investigation into these neuropsychiatric conditions. Clustering of genes according to mRNA expression has particular potential in moving beyond the individual differentially expressed or associated genes. The identification of gene clusters enriched for disorder associated or differentially expressed genes and which contain genes related to particular biological functions may implicate those functions in the aetiology of the disorder.

4.1.3 Enrichment analysis

Clustering genes according to their expression profiles is only the first step in understanding how they work together and their function. One common method which can give further insight into the function of an expression cluster is enrichment analysis. This compares the number of genes present in a cluster that belong to a variety of functional categories, compared with the number of genes present in the 'background' gene list that the clusters are drawn from. A statistical distribution (most frequently the hypergeometric distribution) is used to assess significance for any enrichment of genes belonging to a particular category in a cluster.

As with clustering methods (see Chapter 2), there are an enormous variety of tools available to perform enrichment analysis. Huang *et al* identify three broad classes of enrichment analysis – singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA) (141, 142).

SEA enrichment methods are the simplest class, where genes from each functional category are checked for frequency in the background and target lists individually. GSEA methods are more complex. They involve calculating a score of interest for each gene (e.g. the fold change in a differential expression microarray experiment) and then ranking the genes according to this score. The GSEA algorithms then calculate whether the genes in each functional category have higher ranks than would be expected by chance, using a number of parametric statistical methods (141).

One difficulty with GSEA methods is the necessity of distilling the behaviour of each gene in an experiment down to a single value. This is a particular problem for expression cluster gene lists, where the methods are rarely designed to ascribe a continuous value for cluster membership to every gene in a dataset.

Also, some measures are difficult to reduce to a single value. For example, in the case of GWAS data, each gene will contain a different number of SNPs, each with its own association value. Furthermore, these SNPs may not be independent, as SNPs close to each other on the chromosome may be in linkage disequilibrium (i.e. they are more likely to be transmitted together during reproduction). There is no standard, universally accepted method of combining association values across multiple SNPs while taking into account linkage disequilibrium and the differing numbers of SNPs.

MEA methods are similar to SEA, but use the hierarchical, interconnected nature of GO to enhance their performance. Functional categories are at an advantage if genes from related functional categories are also present in the target list. However, they do have the disadvantage that they bias against 'orphan' categories that have few related functional categories. As the brain is one of the most complex and poorly understood human organs (143, 144), it is likely that the relationships between functional categories related to it are also less well understood. Hence, data from the brain may be particularly badly affected by the tendency of MEA to bias toward already well-annotated groups of GO terms, thus SEA methods were preferred over MEA methods in this analysis.

A number of different functional ontologies can be used in enrichment analysis. One of the most common is the Gene Ontology (GO), which is a publically available, hierarchical system of terms divided into three groups – the molecular function of the product of a gene, the biological process it participates in, and the cellular locations it appears in (145). Alternate versions of GO are available, such as GOslim. This removes many of the more detailed terms, making it useful for giving a broad functional overview of a group of genes, but also risking missing enrichments for detailed gene categories.

Other public ontologies also exist, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), which organising genes into interacting pathways (146). Protein Analysis Through Evolutionary Relationships (PANTHER) is also pathway-based, but is focused upon proteins rather than genes (147, 148). Commercial ontologies are also available, such as MetaCore, a commercial ontology database produced by GeneGO Inc. (see Section 4.1.4 below).

As the number of functional categories present in the ontologies used in enrichment analysis is large, it is necessary to correct the significance values of each category for multiple testing. Since traditional statistical correction methods such as the Bonferroni correction or the family-wise error rate are extremely conservative when dealing with a large number of non-independent categories, false discovery rate (FDR) multiple testing correction is commonly used (149).

In this analysis, the clusters were analysed with the web service GOstat (150) (see Chapter 2, section 2.2.4), the standalone program EASE (151), and the commercial package MetaCore (152). GOstat was chosen because it can use gene

lists containing multiple types of gene ID, allowing a high proportion of the cluster to be included in the analysis.

EASE was used because, unlike most tools, which only work using a particular functional ontology, it can be used to examine gene sets for enrichment with user-specified gene lists. It was used to examine the enrichments of the clusters with genes found to be associated with schizophrenia and bipolar disorder according to a WTCCC GWAS study (153, 154), and genes differentially expressed in bipolar disorder or schizophrenia according to the Stanley Medical Research Institute Online Genomics Database (116).

4.1.4 MetaCore

MetaCore, from GeneGO Inc. (URL: http://www.genego.com), is a commercial database and software package designed for functional analysis (152). It allows the user to construct networks of genes linked by interactions drawn from the curation of PubMed abstracts. This curation is performed entirely by individual scientists, rather than using automated searches through the literature, which avoids many of the pitfalls associated with automatic text mining (155).

MetaCore has several levels of functional category and annotation. The MetaCore maps are the most tightly controlled level, consisting of well supported pathways whose members and interactions are selected by curation by individual researchers (rather than being automatically derived from a database of interactions) (Figure 4.1). These maps display relationships between proteins and other biologically relevant entities. The maps can include activating effects (green arrows), deactivating effects (red arrows), less well-characterised or more complex relationships (grey arrows), and can also show when an interaction between proteins is broken or formed in a particular disease state. Explanatory notes help to explain the details of the process or disease, and also show where the map is believed to connect to other MetaCore maps.

Although user-defined interactions cannot be added to MetaCore maps, it is possible to annotate them with the contents of one or more gene lists (Figure 4.1, red bars adjacent to some proteins). If a gene list has numerical data linked to it, these bars can display this, for instance mean expression of each gene in case

samples and control samples. Figure 4.2 shows the MetaCore map and network legend and illustrates the other kinds of information a MetaCore map can display.

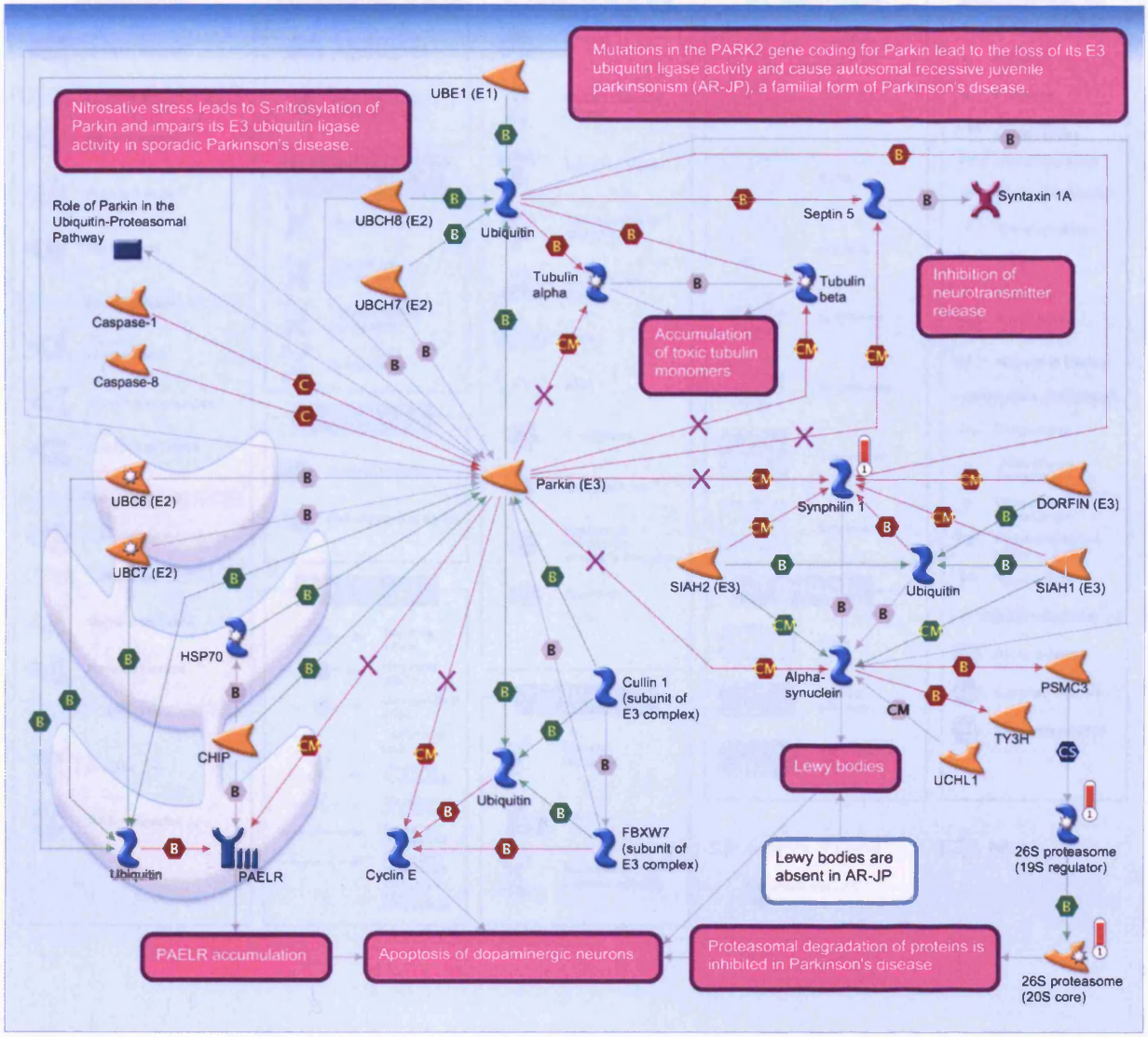
The next level of curation down contains the MetaCore networks. These can be dynamically created by MetaCore based on user input lists, or they can correspond to a particular functional category (Figure 4.3). Again, they display interactions between genes and other biologically relevant entities as arrows. As with MetaCore maps, they can mark genes that are present in gene lists of interest (red circles), and can also mark genes that are members of a particular functional category (blue circles).

Below the maps and networks, MetaCore also supports Gene Ontology (GO) biological process, molecular function and cellular localisation annotations and also MeSH terms, which link genes to diseases (http://www.nlm.nih.gov/mesh/) (145, 156). These categories are identical to the publicly available GO and MeSH ontologies.

MetaCore has several advantages over the publicly available ontologies. The hand-curated database it uses contains a considerable amount of information that is not contained in GO. This is especially important in the field of psychiatric genetics, where the biological systems concerned are incompletely understood, and likely to be complex. Hence, any additional information on the relationships between genes is invaluable. The MetaCore maps are another useful resource, defining which gene links are sufficiently well supported to be generally considered reliable.

The MetaCore networks, conversely, are useful because of their flexibility – any gene can be included, and interactions can be added from outside MetaCore. A variety of algorithms to build a network from a gene list and genes closely related to those in the list is provided. For large gene lists, only using direct interactions between list members can produce an easy-to-interpret network, although it will always omit interactions that function through another biological entity unless these entities are included in your input gene list.

Figure 4.1. Example of a MetaCore map – 'Parkin disorder in Parkinson's Disease'



See Figure 4.2 for legend. Red bars indicate genes present in activated datasets within MetaCore (genes containing SNPs associated with schizophrenia at $p<0.005$ in the WTCCC dataset here).

Figure 4.2. Legend of MetaCore maps and networks

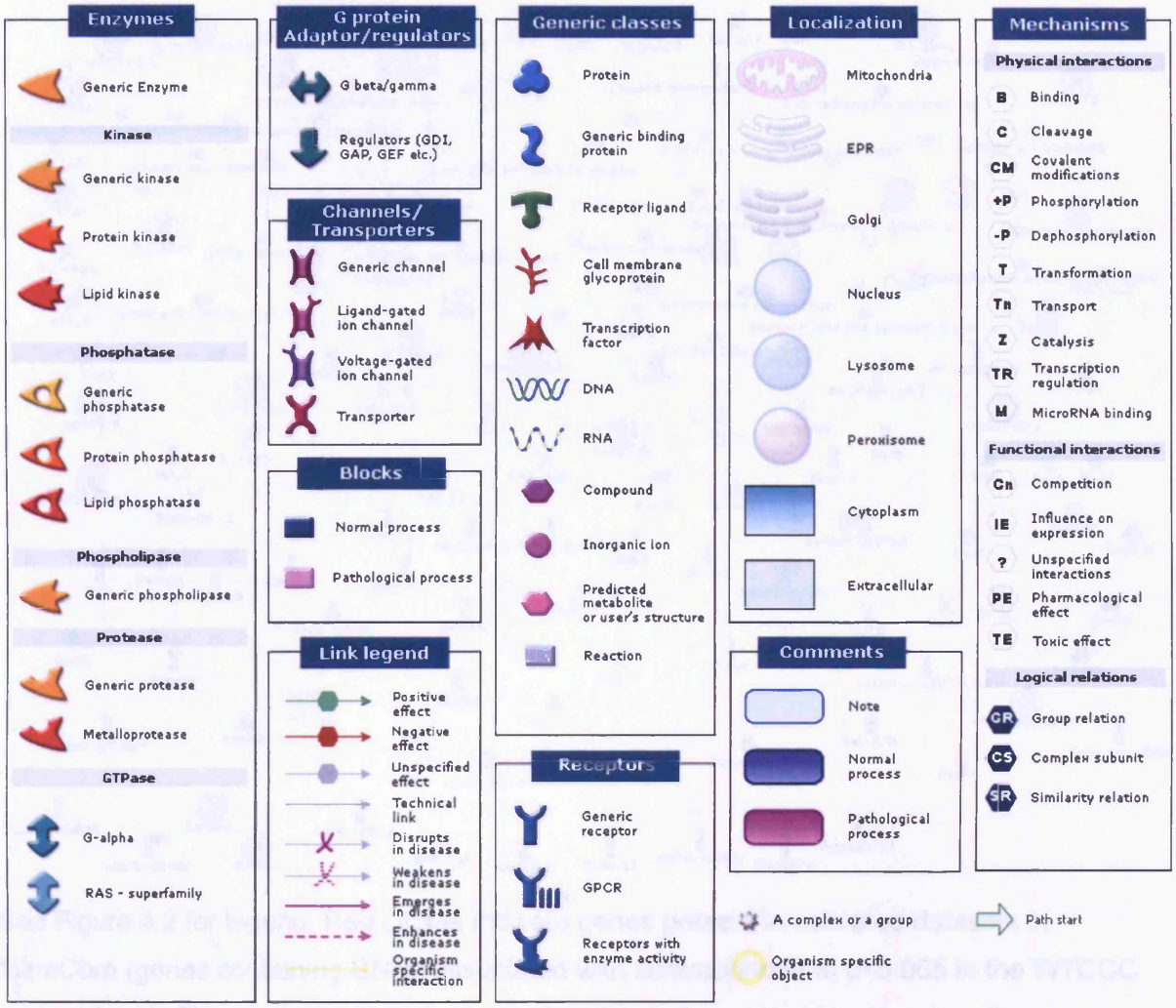
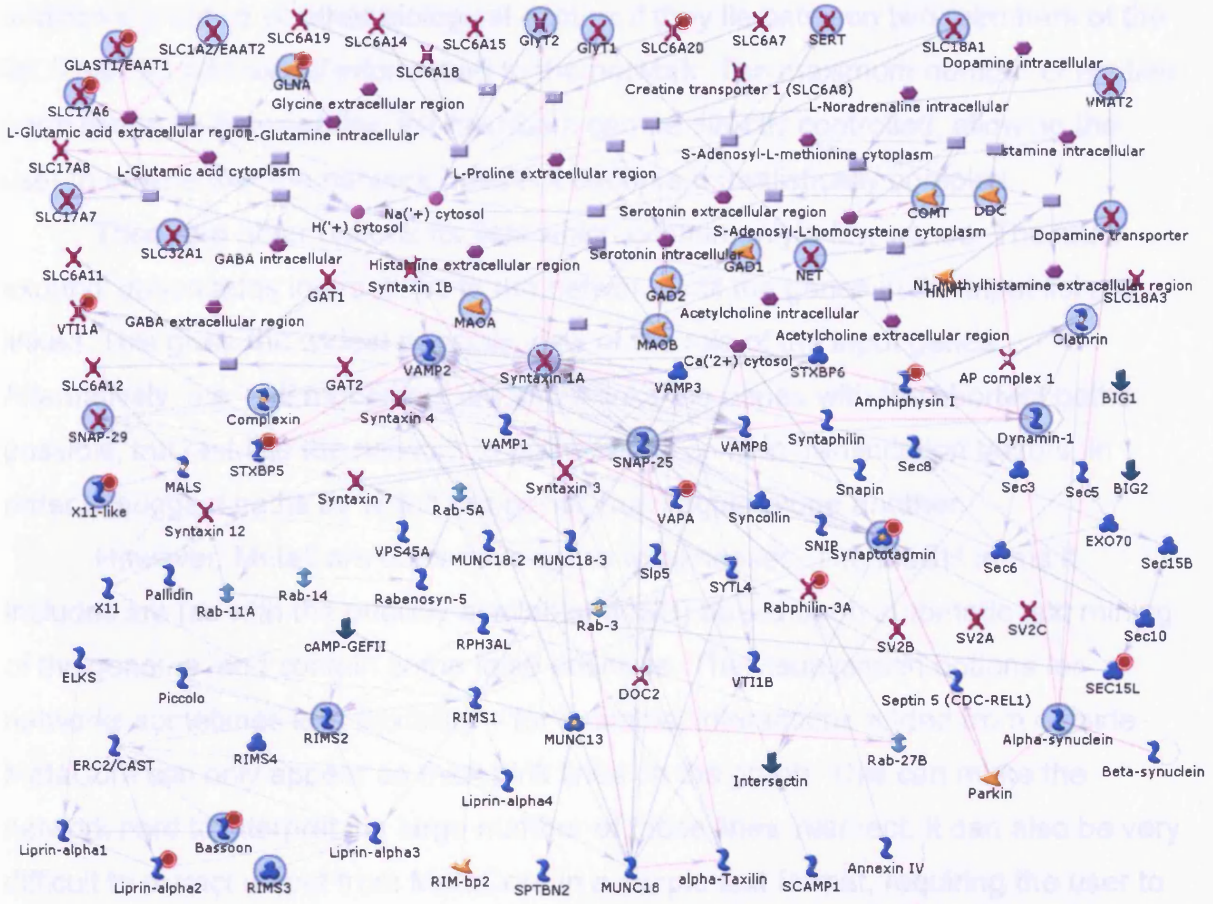


Figure 4.3. Example of a MetaCore network – 'synaptic vesicle exocytosis'



See Figure 4.2 for legend. Red circles indicate genes present in activated datasets in MetaCore (genes containing SNPs associated with schizophrenia at p<0.005 in the WTCCC dataset here). Blue circles indicate genes belonging to the current functional or disease category of interest ('schizophrenia and disorders with psychotic symptoms' here). Edges have been faded for maximum clarity of gene names.

For smaller gene lists, the 'shortest path' algorithm allows MetaCore to add additional proteins or other biological entities if they lie between two members of the list. This can add useful information to the network. The maximum number of entities permitted to lie between two list members can be directly controlled, allowing the user to ensure that the network does not become unrealistically complex.

There are other options for lists which contain only a few genes. The 'auto expand' option adds interactions to the network until the genes in the input list are linked. This gives the widest possible view of the role of the input genes. Alternatively, the 'self regulation' algorithm links the genes with the shortest paths possible, but restricts the network to paths which contain transcription factors, in order to suggest paths by which the genes may regulate one another.

However, MetaCore does have some weaknesses. The MeSH terms it includes are (as with the publicly available MeSH) based upon automatic text mining of the genome, and contain some false positives. The visualisation options for networks sometimes lack flexibility – for example, interactions added from outside MetaCore can only appear as thick pink lines on the graph. This can make the network hard to interpret if a large number of these lines intersect. It can also be very difficult to extract output from MetaCore in a simple text format, requiring the user to transcribe genes of interest by hand. Furthermore, it is impossible to have MetaCore annotate a network with functional category membership information unless that category is in the top 12 most represented functional categories for that network. This can create difficulties when using the networks to address previously formed hypotheses.

4.1.5 Expression correlation network analysis

Expression correlation networks are created by linking genes whose expression patterns have high positive correlation. They exist as hard-threshold and soft-threshold types (157). In the former, edges are binary, existing where the correlation between two nodes exceeds a threshold. In the latter, edges have weights derived from the extent of the correlation. However, when visualising a soft-threshold network, it is usually still necessary to exclude edges below a certain level to keep the graph readable.

These correlation networks are useful for directly visualising the relationships between genes, and showing groups of potentially related genes within a cluster. However, they can become difficult to interpret as the number of genes increases, as the number of possible edges increases quadratically with gene number. This means they are best used upon gene clusters containing at most a few hundred genes.

Correlation networks can be particularly vulnerable to false positives, as a single false positive correlation can pull two functionally disparate areas of the network together, distorting the graph and hampering any biologically relevant interpretation. This effect becomes more pronounced the more nodes there are in a network.

Here, an alternative method of hard-threshold expression correlation network production is used (referred to here as 'two-step network production'). Initially, the network is seeded with the edges with the highest correlation. Then, nodes are iteratively added to the network if they are correlated above a lower threshold with two nodes that are already linked by an edge. The robustness of the two-step method to random noise is compared to the standard method, by using a large number of perturbed datasets based upon a cluster from the MC66 dataset.

Note that this network construction method (like most similar methods) only uses positive correlations. This is because genes that negatively correlate may possess opposing functions (e.g. being related to opposite sides of the cell cycle).

Two-step networks are also constructed for any subclusters heavily enriched for schizophrenia or bipolar disorder associated or differentially expressed genes. The edges of the correlation networks are then added to the MetaCore networks.

4.1.6 GeneCard Inferred Functionality Scores (GIFtS)

To detect clusters containing particularly poorly studied genes, GeneCard Inferred Functionality Scores (GIFtS) were used (158). These scores are a count of the number of GeneCard data types a gene has information for (out of a total of 77). The mean GIFtS was calculated for each cluster and subcluster, and mean GIFtS of subclusters compared to parent clusters to identify particularly heavily or lightly annotated subclusters.

The intent of this was to find particularly lightly annotated subclusters that were also enriched for genes associated with or differentially expressed in schizophrenia or bipolar disorder. Such subclusters may be particularly worthwhile to focus further research on, as relatively little is known about their biology.

4.2 Methods

4.2.1 Enrichment of clusters for schizophrenia related genes

The 26 clusters produced from the combined k-means/ISA/memISA method on the Dobrin dataset (see Chapter 2, Section 2.3.6) were tested for enrichment with 607 genes which contained at least one SNP associated with schizophrenia at nominal $p<0.005$ according to a recent genome-wide association study (the UK schizophrenia study – see Chapter 3, Section 3.2.1). This enrichment test was performed using the program EASE (153, 159), which implements a version of Fisher's Exact Test, and used the full complement of genes in the Dobrin dataset as a background. The UK schizophrenia dataset consists of 2938 control samples from the WTCCC genome-wide association study (160) and 479 cases from a UK schizophrenia study (153).

Clusters enriched for schizophrenia-associated genes were also tested for enrichment with 352 genes found to be differentially expressed between schizophrenics and controls in the analysis of the Stanley Medical Research Institute Online Genomics Database (81) at an uncorrected p-value of 0.02 or lower. This choice of p-value was somewhat arbitrary, but was found to produce a differentially expressed gene list of a reasonable size (355 genes). Again, this used EASE, and used the full set of genes present in the Dobrin dataset as a background.

It should be pointed out that, as the Dobrin dataset is a large part of the Stanley database, both this analysis and the equivalent bipolar analysis (see Section 4.2.3 below) are somewhat circular when applied to clusters based on Dobrin expression data. Hence, it is particularly important to show replication for any enrichment of differentially expressed genes in these clusters.

These clusters were also examined for enrichment in KEGG and BioCarta pathways, using the Composite Regulatory Signature Database (161) (http://140.120.213.10:8080/crsd/main/home.jsp), and for enrichment in GO biological process categories using GOstat.

EASE was also used to test these clusters for enrichment with genes found to be ten-fold or more upregulated in specific cell types within brain tissue according to Cahoy *et al* (162)– specifically, neurons, oligodendrocytes and astrocytes. Clusters enriched for disease associated or differentially expressed genes from the Dobrin dataset were examined for overlap with every cluster from the MC66 dataset. Clusters that shared a high proportion of their genes with any enriched cluster from the Dobrin dataset were then identified. Their enrichment for schizophrenia-associated genes and genes differentially expressed in schizophrenia was then determined with EASE.

One cluster from the Dobrin dataset was particularly enriched for genes associated with schizophrenia (see Results, Section 4.3.1 below). This cluster contained 3093 genes, so is subsequently referred to as the 'Dobrin 3093' cluster. Two clusters from the MC66 dataset overlapped with it, one 2546-gene cluster and one 436-gene cluster. 52.3% of genes in the 2546 gene cluster were also present in the Dobrin 3093 gene cluster and 48.8% of genes in the 436 gene cluster were present in the Dobrin 3093 gene cluster. These are subsequently referred to as the 'MC66 2546' and 'MC66 436' clusters, and they were examined for enrichment with schizophrenia-associated or differentially expressed genes with EASE in the same way as the Dobrin 3093 (except using the full set of MC66 genes as a background list). Similarly, they were also examined for enrichment with KEGG or BioCarta pathways using the Composite Regulatory Signature Database, and for genes upregulated in brain cell types using EASE. Further analyses primarily focused on the Dobrin 3093 and MC66 2546 clusters.

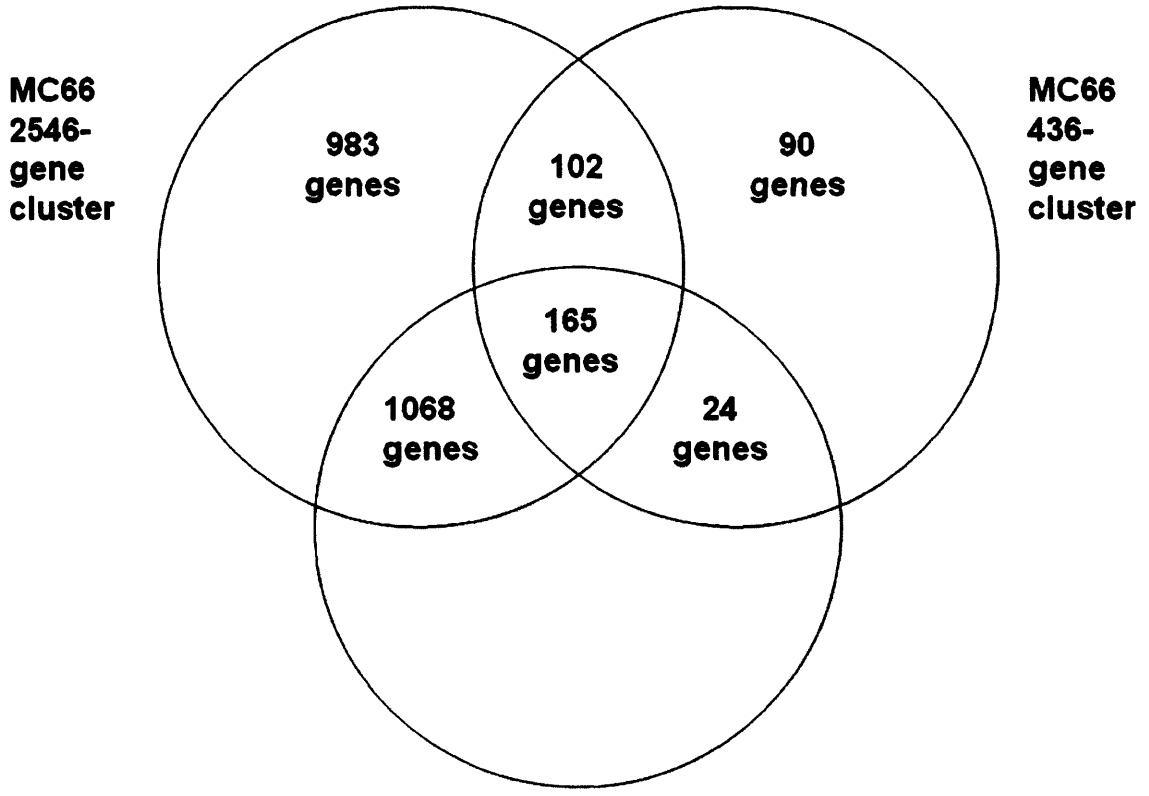
4.2.2 Permutation-based enrichment significance

In order to prevent the overlap between the Dobrin 3093 cluster and the MC66 2546 and 436 clusters producing an apparent enrichment of schizophrenia-associated genes in the latter by chance, an additional permutation-based method was used to examine whether the MC66 2546 and 436 clusters were enriched for schizophrenia-associated or differentially expressed genes. If a cluster remained enriched when using this permutation-based method, it could be considered an independent replication of the enrichment of the Dobrin 3093 for schizophrenia-associated genes.

4000 pairs of clusters were constructed at random from the genes present on the Affymetrix 133A chip, as follows. Firstly, the number of genes shared between the three clusters was calculated (see Figure 4.4). These figures were then used to create randomised MC66 clusters with the same level of overlap with the Dobrin cluster and each other.

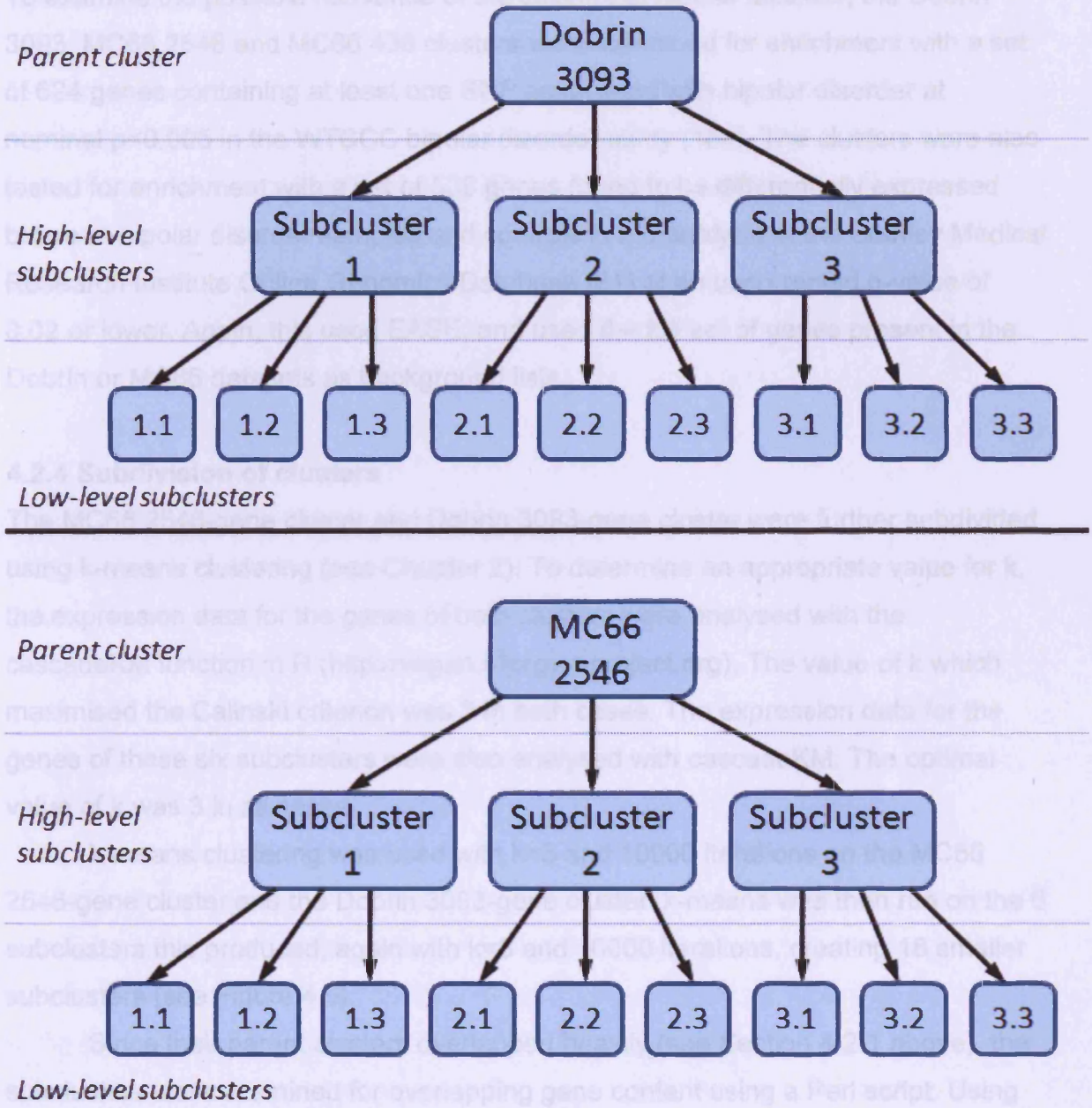
165 genes from the Dobrin 3093-gene cluster were selected at random, and placed in both the 2546-gene and 436-gene MC66 randomised clusters. From the remaining Dobrin cluster genes, 1068 and 24 genes were selected at random, the former placed in the 2546-gene randomised cluster, the latter placed in the 436-gene randomised cluster. Then, 102 genes from the genes on the chip not present in the Dobrin 3093-gene cluster were selected at random, and placed in both the 2546-gene and 436-gene randomised clusters. From the remaining genes on the chip not present in the Dobrin 3093-gene cluster, 983 and 90 genes were selected at random, the former placed in the 2546-gene randomised cluster, the latter in the 436-gene randomised cluster. This was repeated 4000 times to produce a population of 8000 random clusters. These permuted clusters were then processed with EASE in the same way as the original cluster (looking for enrichment for genes associated with schizophrenia and genes differentially expressed in schizophrenia), allowing the original results to be compared to them. The permutation p-value for enrichment was defined as the number of permuted clusters which were more significantly enriched than the original cluster, divided by the total number of permuted clusters.

Figure 4.4. Overlap between putative schizophrenia-related clusters produced from Dobrin and MC66 datasets



Venn diagram showing the amount of overlap between the clusters enriched for schizophrenia-related genes, in order to construct randomised clusters for permutation

Figure 4.5. Subdivision of Dobrin 3093-gene and MC66 2546-gene clusters using k-means clustering



At each level, clusters are subdivided using k-means clustering (k=3).

4.2.3 Enrichment for genes associated with or differentially expressed in bipolar disorder

To examine the possible relevance of the clusters to bipolar disorder, the Dobrin 3093, MC66 2546 and MC66 436 clusters were examined for enrichment with a set of 624 genes containing at least one SNP associated with bipolar disorder at nominal $p<0.005$ in the WTCCC bipolar disorder study (160). The clusters were also tested for enrichment with a set of 538 genes found to be differentially expressed between bipolar disorder samples and controls in the analysis of the Stanley Medical Research Institute Online Genomics Database (81) at an uncorrected p-value of 0.02 or lower. Again, this used EASE, and used the full set of genes present in the Dobrin or MC66 datasets as background lists.

4.2.4 Subdivision of clusters

The MC66 2546-gene cluster and Dobrin 3093-gene cluster were further subdivided using k-means clustering (see Chapter 2). To determine an appropriate value for k, the expression data for the genes of both clusters were analysed with the cascadeKM function in R (http://vegan.r-forge.r-project.org). The value of k which maximised the Calinski criterion was 3 in both cases. The expression data for the genes of these six subclusters were also analysed with cascadeKM. The optimal value of k was 3 in all cases.

k-means clustering was used with k=3 and 10000 iterations on the MC66 2546-gene cluster and the Dobrin 3093-gene cluster. k-means was then run on the 6 subclusters this produced, again with k=3 and 10000 iterations, creating 18 smaller subclusters (see Figure 4.5).

Since their parent clusters overlapped heavily (see Section 4.2.1 above), the subclusters were examined for overlapping gene content using a Perl script. Using EASE, these subclusters were also examined for enrichment with genes differentially expressed in schizophrenia or bipolar disorder, genes associated with schizophrenia or bipolar disorder, and genes linked to brain cell types in the same way as the original clusters.

4.2.5 GIFtS

The mean GIFtS values were calculated for the genes of each cluster, subcluster and background gene list using data from the GeneCards website (http://www.genecards.org). To ensure independence from the tests for enrichment for functional categories, the presence or absence of GO category data was excluded from the total GIFtS value of each gene. These mean GIFtS values were compared to each other and to the parent clusters to identify clusters and subclusters with particularly high GIFtS values, which could indicate that enrichments for functional or disease-related categories may be due to annotation bias. Conversely, clusters and subclusters with lower GIFtS values were also identified, as their enrichments are less likely to be due to annotation bias.

4.2.6 MetaCore

The 24 subclusters and their two parent clusters were uploaded to the functional analysis tool MetaCore. They were each tested for enrichment with genes from a number of types of MetaCore functional category, including:

1) MetaCore maps
2) MetaCore networks
3) GO biological process categories
4) GO molecular function categories
5) GO localisation categories

These tests used the proprietary enrichment algorithm included in MetaCore. The genes from clusters that were heavily enriched for genes associated with or differentially expressed in schizophrenia according to EASE were further examined using the network-building capabilities of MetaCore. Only direct, trusted links between genes were used to populate the network, and analysis focused on any large groups of linked genes that emerged. These groups were expanded by adding links found by performing two-step coexpression network construction upon the expression data of the cluster (see below). The clusters were also examined for enrichment with the 'disease' functional category in MetaCore. This links genes with diseases based upon literature mining.

4.2.7 Two-step coexpression network construction and testing

To examine the usefulness of coexpression network construction while limiting the effect of random false positive links, a two-step network construction method was used. Initially, correlations between the expression profiles of the genes were calculated. The network is then seeded with edges. Nodes are linked whenever their Pearson's correlation coefficient is greater than a stringent threshold *s*. In the second step, nodes that have a Pearson's correlation coefficient above a more lax threshold *l* with two already linked genes are added to the network. This step is repeated until no further nodes can be added to the network. In order to allow maximum comparability between networks generated from different data, *s* and *l* are expressed in terms of standard deviations above the mean correlation coefficient for all gene pairs. Note that the two-step method only considers positive correlations between genes – no edges are placed based upon negative correlations.

The intent of this two-step process was to limit the effect of single, random false positive links between nodes. These can bring biologically disparate sections of the graph close together when visualised, hindering interpretation of the network.

To provide a comparison, a single-step coexpression network construction method was written. This simply assigned links between nodes whenever the Pearson correlation between their expression profiles exceeded a threshold, *t*. Again, *t* was expressed in terms of standard deviations above the mean correlation coefficient for all gene pairs, and only positive expression correlations between genes were considered.

To examine the sensitivity of the one-step and two-step method to noise, three sets of perturbed datasets were produced. Each perturbed dataset was based upon the expression profiles of a 401-gene subcluster (cluster 2.1) of the MC66 2546-gene expression cluster (see above). In each set, 1000 perturbed versions of the expression data were created using a Perl script. In the first set, a random 20% of the expression values were increased or decreased by a factor between 0 and 20% (also chosen at random for each cell). In the second set, 40% of the data was perturbed at random, by up to 40% of their expression value. In the third set, 60% of the data was perturbed at random, by up to 60% of their expression value.

In all two-step analyses presented here, the parameters *s* and *l* are set to 1.8 and 1.5 respectively, as, when used on MC66 subcluster 2.1, these values were

found to produce interpretable networks without either a large excess or total lack of edges. Similarly, when the one-step method was used, *t* was set to 1.6, as this value also produced interpretable networks when used on MC66 subcluster 2.1.

One- and two-step networks were produced for each perturbed dataset. These were compared to the one- and two-step networks produced for the original 401-gene subcluster using a Perl script. This determined how many links between nodes were present in both networks, as a percentage of links that were present in at least one network. This quantity acted as a metric of network stability when presented with noisy data. T-tests were used to compare these percentages between the 1 and 2 step networks.

The two-step coexpression network construction method was used upon MC66 2546 and Dobrin 3093 subclusters with significant functional analysis enrichments in categories that could plausibly play a role in bipolar disorder or schizophrenia aetiology. The correlations from these networks were included when building networks with MetaCore (see above).The effect of this on the number of genes included on the network and the most significant functional category enrichment of the network was also examined.

4.2.8 Correction for gene length

Subsequent to the other analyses, a second round of enrichment analysis was performed. In the previous analysis, genes were selected based upon the most associated SNP they contained. However, this biased toward long genes, as these are likely to contain a larger number of independent SNPs, and so will have multiple opportunities to obtain a highly associated SNP. This may induce a bias toward genes expressed in neurons, as they tend to be long (163).

In this second analysis, the p-value of the most associated SNP was corrected by gene length. This was achieved by creating 1000 versions of the GWAS dataset with randomly permuted case/control status, and calculating association values for every SNP in each permuted dataset. Again, genes are assigned association p-values based upon the most significant association of any SNP within their sequence. The length-adjusted p-value for each gene is equal to the number of permuted runs where the association p-value of the gene is more significant than the association p-value in the original run, divided by 1000. This

corrects for different gene lengths and patterns of linkage disequilibrium between genes.

Lists of gene length adjusted associated genes were produced based upon three sources – the WTCCC bipolar disorder sample, the UK schizophrenia sample and the International Schizophrenia Consortium (ISC) sample (29). The enrichment of clusters and subclusters was calculated using EASE.

4.3 Results

4.3.1 Enrichment of clusters for schizophrenia related genes

The clusters produced from the combined k-means/ISA/memISA method on the Dobrin dataset were tested for enrichment with 607 genes associated with schizophrenia according to a recent genome-wide association study (160), using the program EASE (159). These 607 genes each contained at least one SNP associated with schizophrenia at an Armitage p-value of 0.005 or under. One cluster, containing 3093 genes and originally found by memISA, was enriched (p=0.0004 before correction, p=0.0104 after Bonferroni correction).

This cluster was also tested for enrichment with 352 genes found to be differentially expressed between schizophrenics and controls in the analysis of the Stanley Medical Research Institute Online Genomics Database (81) at an uncorrected p-value of 0.02 or lower. The cluster showed a non-significant trend toward enrichment, at a nominal p-value of 0.09.

Clusters from combined k-means/ISA/memISA in the independent MC66 dataset that shared a high proportion of their genes with this enriched cluster were then identified. Two clusters were found (containing 2546 and 436 genes respectively), both of which were nominally enriched for both schizophrenia-associated genes (2546-gene cluster at p=0.00844, 436-gene cluster at p=0.0117) and genes differentially expressed in schizophrenia (2546-gene cluster at p=0.004, 436-gene cluster at p=0.00047) (Table 4.1). 52% of the genes in the 2546-gene cluster were present in the Dobrin 3093 cluster, as were 47% of the genes in the 436-gene

Table 4.1. Enrichment of parent clusters for genes associated with or differentially expressed in schizophrenia or bipolar disorder

Cluster name	Cluster size	Average GIFtS value of genes	Enrichment p-value for UK SCZ associated genes	Enrichment p-value for genes diff. expressed in SCZ	Enrichment p-value for WTCCC BP associated genes	Enrichment p-value for genes diff. expressed in BP
Dobrin 3093	3093	49.54	0.0004	0.062	4.00E-07	4.42E-08
MC66 2546	2546	51.75	0.00844	0.004	0.0478	1.3E-12
MC66 436	436	57.76	0.0117	0.0005	0.0140	0.0843

cluster. However, these enrichments may have been due to their overlap with the 3093-gene Dobrin cluster, and so cannot be considered independent replications of the original cluster.

To avoid this confounding effect, the enrichment of the clusters for schizophrenia-associated genes and genes differentially expressed in schizophrenia was determined using a permutation-based method. The 436-gene cluster was significantly enriched for the schizophrenia associated genes ($p=0.0255$), while the 2546-gene cluster was not ($p =0.169$). Both clusters were significantly enriched for genes differentially expressed in schizophrenia (permutation $p = 0.0053$ for the 2546-gene cluster, permutation $p = 0.0005$ for the 436-gene cluster).

These two clusters and the Dobrin 3093-gene cluster were tested for enrichment with genes associated with bipolar disorder according to the WTCCC study, and also for genes differentially expressed in bipolar disorder according to the Stanley Medical Research Institute Online Genomics Database (81, 154). The Dobrin 3093-gene cluster was strongly enriched for both the WTCCC bipolar association gene list ($p=1.54e^{-4}$ after Bonferroni correction) and the Stanley differentially expressed gene list ($p=1.15e^{-6}$ after Bonferroni correction).

The MC66 2546-gene and 436-gene clusters were again tested using permutation. The 2546-gene cluster was enriched for the Stanley bipolar disorder differentially expressed genes (perm $p = 0$), but not for the bipolar disorder association genelist. The 436-gene cluster was enriched for the WTCCC bipolar

disorder association list (perm p=0.0215) but not for the Stanley bipolar disorder differentially expressed genes.

These clusters were also examined for enrichment in KEGG and BioCarta pathways, using the Composite Regulatory Signature Database (161) (http://140.120.213.10:8080/crsd/main/home.jsp). The top hit for the Dobrin cluster and the 2546-gene MC66 cluster was the KEGG entry for the MAPK signalling pathway ($p=1.12e^{-7}$, FDR q=0.00024 in Dobrin, $p=6.95e^{-10}$, FDR $q=1.46e^{-6}$ in MC66). The only significant hit for the MC66 436-gene cluster was from the BioCarta Synaptic Junction pathway ($p=3.88e^{-5}$, FDR $q=2.71e^{-2}$) (Table 4.2).

The MC66 436-gene cluster was also examined using GOstat, where the best hit was for the 'nervous system development' GO category (p=0.044 after FDR correction). There was also a near-significant hit for serine / threonine kinases (p=0.07 after FDR).

The three clusters were also tested for enrichment with genes found to be ten-fold or more upregulated in specific cell types within brain tissue according to Cahoy *et al* (162)– specifically, neurons, oligodendrocytes and astrocytes. All three clusters were found to be highly significantly enriched with genes upregulated in neurons ($p=2.5e^{-21}$ in Dobrin, $p=1.55e^{-16}$ in MC66, Bonferroni corrected). There was also some enrichment for genes upregulated in oligodendrocytes (Dobrin p=0.06, MC66 $p=2.4e^{-4}$, Bonferroni corrected) and astrocytes (Dobrin $p=5.13e^{-22}$, MC66 $p=2.26e^{-10}$, Bonferroni corrected).

Table 4.2. Enrichment of parent clusters for KEGG, BioCarta, MetaCore and GO functional categories, and genes upregulated in brain cell types

Cluster name	Heavily enriched functional categories	Database	P value	Heavily enriched cell type lists
Dobrin 3093	MAPK signaling	KEGG	$1.12e^{-7}$	Astrocytes, neurons
	Amphoterin signaling	MetaCore	$3.42e^{-6}$	
	Angiogenesis regulation	MetaCore	$2.59e^{-4}$	
	Acetyltransferases	GO	$1.13e^{-8}$	
	Negative regulation of development	GO	$1.21e^{-8}$	
	Cell communication	GO	$1.71e^{-7}$	
	Anti-apoptosis	GO	$2.25e^{-7}$	
	Angiogenesis	GO	$4.24e^{-7}$	
	Kinase regulation	GO	$2.09e^{-6}$	
MC66 2546	MAPK signaling	KEGG	$6.95e^{-10}$	Neurons, astrocytes
	Cytoplasmic microtubules	MetaCore	$2.98e^{-5}$	
	Synaptic contact	MetaCore	$1.25e^{-4}$	
	CNS development	GO	$1.1e^{-11}$	
	Regulation of synaptic transmission	GO	$9.28e^{-10}$	
	Acetyltransferases	GO	$7.11e^{-9}$	
	Regulation of exocytosis	GO	$2.24e^{-8}$	
MC66 436	Synaptic junction	BioCarta	$3.88e^{-5}$	Neurons, oligodendrocytes
	Synaptic contact	MetaCore	$9.5e^{-5}$	
	Cell surface receptor linked signaling pathway	GO	$1.76e^{-7}$	
	Regulation of cell projection organization	GO	$4.85e^{-7}$	
	CNS development	GO	$8.06e^{-7}$	

Three overlapping clusters, enriched to varying degrees for either schizophrenia-associated genes or genes differentially expressed in schizophrenia were found from the two independent dorsolateral prefrontal cortex datasets (82, 116). The apparent excess of schizophrenia-associated genes in the 2546-gene MC66 cluster could be explained by it being selected based upon overlap with the Dobrin cluster. Thus, although the cluster itself appears independently in both datasets, it does not constitute independent evidence for schizophrenia-associated genes clustering together with respect to their expression levels. However, the 436-gene MC66 cluster remained significantly enriched when assessed by the permutation method and so does constitute independent evidence of this. Also, both MC66 clusters did show significant over-representation for genes differentially expressed in schizophrenia, even after correction for the overlap with the Dobrin cluster. This demonstrates the ability of the clustering methods (originally memISA for all these clusters) in finding potentially disease-related functional clusters.

4.3.2 Two-step coexpression network construction and testing

One-step and two-step networks were produced from the 3000 perturbed datasets, and these compared to the non-perturbed networks. A significantly higher percentage of links were present in both original and perturbed networks when 2-step network production was used (in all three perturbed datasets - Table 4.3). This indicates that 2-step network production is more robust to random noise than the 1-step network.

The effect on functional analysis enrichment of adding the two-step coexpression network interactions to two of the MetaCore networks was also examined. The Dobrin 2.3
network acquired 23 additional genes, an increase of 31%, and its most significant GO category was of a similar order of magnitude ($p=3.9e^{-6}$, compared to $p=1.5e^{-6}$ previously). The MC66 1.3 network acquired 80 additional genes, an increase of 48%, and its most significant GO category was also of a similar order of magnitude ($p=3.8e^{-16}$, compared to $p=1.2e^{-18}$ previously). The expansion of the network was felt to be worth the slight reduction in significance, so two-step gene coexpression interactions were added to all interesting clusters in MetaCore.

Table 4.3. Effect of using perturbed datasets on one-step and two-step expression correlation network construction

Percentage chance of a data point being perturbed	Maximum perturbation of a data point	Mean percentage of links present in both original and perturbed one-step networks	Mean percentage of links present in both original and perturbed two-step networks	t-test p-value
20%	20%	82.46804	84.45711	$<2.2e^{-16}$
40%	40%	55.42918	60.28294	$<2.2e^{-16}$
60%	60%	31.44726	37.52607	$<2.2e^{-16}$

4.3.3 Functional analysis of subclusters using MetaCore and EASE

Subclusters derived from the Dobrin 3093-gene cluster were examined for overlapping gene content with the subclusters derived from the MC66 2546-gene cluster. Overlap was defined as the smaller cluster sharing 30% or more of its genes with the larger cluster. The majority of the subclusters overlapped with a subcluster from the other dataset, although some did not. Several of the clusters were also enriched for genes associated with, or differentially expressed in, schizophrenia or bipolar disorder. The clusters were examined for enrichment with five types of MetaCore functional categories, as well as the Cahoy cell type lists.

4.3.4 Dobrin 3093-gene and MC66 2546-gene cluster

Both the Dobrin 3093-gene and MC66 2546-gene cluster were examined for enrichment with MetaCore. Neither was enriched for any MetaCore maps, although the Dobrin cluster was significantly enriched for the MetaCore amphoterin signalling network. The two clusters have very similar enrichments for GO molecular function (acetyltransferases, angiotensin, alpha adrenergic receptor, G-proteins) and GO cellular compartment (cytoplasm, synapse, plasma membrane) categories (Table 4.2). The similarities in the GO biological process categories between them were not as great – the MC66 2546-gene cluster was much less enriched for kinases, in particular. However, both clusters contain concentrations of development-related genes, transport-related genes and anti-apoptotic genes.

4.3.5 MC66 subcluster 1 and 1.3 and Dobrin subclusters 2, 1.1 and 1.2

Subcluster 1 of the MC66 2546-gene cluster, and its constituent subclusters (1.1, 1.2 and 1.3) were examined for overlap with the Dobrin subclusters (Tables 4.4 and 4.5).

MC66 subcluster 1 and Dobrin subcluster 2 are both enriched for genes associated with and differentially expressed in bipolar disorder, so these pathways may be important in bipolar aetiology (Table 4.6). They also have some enrichment for schizophrenia associated genes, although less significantly. They contain genes related to GABA neurotransmission, calcium signalling, synaptic contact, and synaptic vesicle exocytosis (Table 4.7).

MC66 subcluster 1.3 is particularly interesting. Both it and Dobrin subcluster 2.2, with which it overlaps heavily, are enriched for SCZ associated, BP associated

and BP differentially expressed genes. When examined with MetaCore, it is also enriched for genes linked with schizophrenia in MeSH (MetaCore $p=3.09e^{-7}$) (Table 4.8). Few other subclusters showed any enrichment for SCZ MeSH terms, so further attention was focused on this cluster. However, Dobrin subcluster 2.2 is not enriched for the SCZ MeSH terms (Table 4.9). Genes in these two subclusters do not have significantly higher GIFtS than their parent clusters (Dobrin t-test p = 0.43, MC66 t-test p = 0.17).

Tables 4.4 and 4.5 Dobrin 3093 and MC66 2546 subcluster overlap

Dobrin 3093	MC66 2546	Overlap
1	1	34.3%
1.1	1.1	43.9%
1.2	1.2	46.2%
1.3	2.2	30.4%
2	1	50.2%
2.1	2.1	44.7%
2.2	1.3	70.6%
2.3	N/A	N/A
3	2	52.4%
3.1	2.1	39.6%
3.2	2.2	52.1%
3.3	N/A	N/A

MC66 2546	Dobrin 3093	Overlap
1	2	50.2%
1.1	1.1	43.9%
1.2	1.2	46.2%
1.3	2.2	70.6%
2	3	52.4%
2.1	3.1	39.6%
2.2	3.2	52.1%
2.3	1.2	31.1%
3	N/A	N/A
3.1	N/A	N/A
3.2	N/A	N/A
3.3	1.3	30.4%

Tables show subcluster from other dataset with highest overlap (N/A indicates that no subcluster shared over 30% of its genes with the subcluster)

Table 4.6. Dobrin 3093 and MC66 2546 expression clusters and subclusters - cluster size, GIFtS value, overlap and enrichment for schizophrenia or bipolar disorder associated or differentially expressed genes

Cluster name	Cluster size	Average GIFtS value of genes	Best overlap (N/A = no overlap above 30%)	Enrichment p-value for WTCCC SCZ associated genes	Enrichment p-value for genes diff. expressed in SCZ	Enrichment p-value for WTCCC BP associated genes	Enrichment p-value for genes diff. expressed in BP
Dobrin 1	697	48.26	MC66 1	0.84	0.164	0.316	**2.27E-08**
Dobrin 1.1	215	47.99	MC66 1.1	0.736	0.661	0.75	**7.84E-06**
Dobrin 1.2	148	50.31	MC66 1.2	0.943	0.815	0.317	0.467
Dobrin 1.3	334	47.50	MC66 3.3	0.597	**0.0356**	0.279	**7.07E-05**
Dobrin 2	868	49.51	MC66 1	**6.12E-05**	0.899	**4.21E-08**	**0.00104**
Dobrin 2.1	208	49.87	MC66 1.3	0.433	1	0.152	0.556
Dobrin 2.2	274	50.04	MC66 1.3	**0.000298**	0.468	**0.000101**	**0.000917**
Dobrin 2.3	386	48.95	N/A	**0.0135**	0.804	**0.000121**	0.0607
Dobrin 3	1209	50.36	MC66 2	**0.035**	**0.0115**	0.303	0.479
Dobrin 3.1	317	51.45	MC66 2.1	**0.00367**	**1.4E-05**	0.396	**0.0217**
Dobrin 3.2	484	52.23	MC66 2.2	0.745	0.45	0.524	0.925
Dobrin 3.3	408	47.18	N/A	0.187	0.931	0.316	0.79
MC66 1	843	51.43	Dobrin 2	0.058	0.798	**0.00195**	**1.79E-18**
MC66 1.1	377	50.13	Dobrin 1.1	0.86	0.919	0.905	**5.88E-11**
MC66 1.2	53	53.06	Dobrin 1.2	1	1	**0.0451**	0.311
MC66 1.3	413	52.42	Dobrin 2.2	**0.00126**	0.437	**8.3E-05**	**2.08E-08**
MC66 2	737	53.63	Dobrin 2	0.166	**0.00031**	0.728	0.113
MC66 2.1	349	54.38	Dobrin 3.1	**0.00542**	**8.872E-05**	0.826	0.0844
MC66 2.2	96	56.41	Dobrin 3.2	1	0.433	0.945	**0.0295**
MC66 2.3	292	51.82	Dobrin 1.2	0.78	0.295	0.267	0.935
MC66 3	856	50.69	N/A	0.0912	0.101	0.398	0.585
MC66 3.1	78	51.32	N/A	**0.00394**	**0.0104**	**0.0239**	0.403
MC66 3.2	305	52.64	N/A	0.727	0.983	0.664	0.99
MC66 3.3	472	49.34	Dobrin 1.3	0.224	0.0482	0.698	0.137

Bold type indicates nominally significant clusters and subclusters.

Table 4.7. Cahoy cell type and MetaCore functional categories enriched in the top-level subclusters of the Dobrin 3093 and MC66 2546 clusters

Subcluster name	MetaCore/GO biological process categories enriched in cluster	Database	P value	Cell type lists enriched in cluster
Dobrin 1	Helicases	GO	6.67e-7	Oligodendrocyte
	Protein localisation	GO	6.10e-6	Astrocyte
	Anti-apoptosis	GO	1.23e-5	
Dobrin 2	GABA neurotransmission	MetaCore	4.12e-12	Neuron
	Synaptic vesicle exocytosis	MetaCore	1.24e-7	
	Synaptic contact	MetaCore	6.57e-7	
	Calcium signaling	MetaCore	6.66e-8	
	Neuroendocrine-macrophage connector	MetaCore	2.42e-5	
	Synaptic transmission	GO	1.00e-7	
	Neuron development	GO	9.75e-11	
	Synaptic vesicle transport	GO	2.79e-9	
	Memory	GO	1.61e-9	
Dobrin 3	Angiogenesis	MetaCore	1.41e-9	Astrocyte
	Amphoterin signaling	MetaCore	4.69e-8	
	Th17-derived cytokines	MetaCore	4.52e-8	
	Platelet-endothelium-leukocyte interactions	MetaCore	4.75e-7	
	IL-1 signaling	MetaCore	3.99e-7	
	Wounding response	GO	3.51e-21	
	Development	GO	4.45e-20	
	Angiogenesis	GO	3.44e-20	
	Kinases	GO	1.06e-14	
	Apoptosis regulation	GO	6.04e-13	
MC66 1	GABA neurotransmission	MetaCore	7.13e-7	Neuron
	Synaptic contact	MetaCore	5.72e-6	
	GABA-A receptor life cycle	MetaCore	2.86e-5	
	Nerve impulse transmission	MetaCore	5.76e-6	
	Synaptic transmission	GO	1.75e-7	
	Transport	GO	5.29e-7	
	CNS development	GO	1.05e-6	
MC66 2	Platelet-endothelium-leukocyte interactions	MetaCore	2.01e-6	Astrocyte
	Angiogenesis regulation	MetaCore	3.38e-6	
	Chemotaxis	MetaCore	5.41e-6	
	Wound response	GO	9.25e-20	
	Inflammation	GO	2.57e-13	
MC66 3	Neurotransmitter transport	GO	3.93e-9	Neuron,
	Synaptic transmission	GO	4.21e-9	oligodendrocyte
	GABA secretion	GO	4.75e-8	
	Synaptic plasticity regulation	GO	1.9e-6	

Table 4.8. Cahoy cell type and MetaCore functional categories enriched in the 9 low-level subclusters of the MC66 2546 expression cluster

Subcluster name	MetaCore GO biological process categories enriched in cluster	Database	P-value	Cell type lists enriched in subcluster
MC66 1.1	DNA damage and BRCA1	MetaCore	8.44e-5	Oligodendrocyte
	Cellular biopolymer metabolic process	GO	3.41e-6	
MC66 1.2	Connective tissue degradation	MetaCore	2.79e-5	Oligodendrocyte
	Myelination	GO	4.88e-9	
	Neuronal action potential regulation	GO	3.68e-8	
	Vasoconstriction	GO	7.31e-6	
	Ion homeostasis	GO	8.47e-6	
	Beta amyloid metabolism	GO	1.07e-5	
MC66 1.3	ERK inhibition	MetaCore	1.88e-6	Neuron
	CDK5 presynaptic signaling	MetaCore	3.36e-6	
	Neuroendocrine-macrophage connector	MetaCore	5.09e-6	
	GABA neurotransmission	MetaCore	3.42e-6	
	Synaptic contact	MetaCore	1.82e-6	
	Cytoplasmic microtubules	MetaCore	3.58-5	
	MeSH schizophrenia terms	MetaCore	3.09e-7	
	Transport	GO	2.23e-13	
	Synaptic transmission	GO	4.56e-11	
	CNS development	GO	1.92e-10	
	Ribonucleotide biosynthesis	GO	8.45e-10	
	ATP metabolism	GO	5.81e-9	
MC66 2.1	Skeletal muscle development	MetaCore	6.02e-6	Astrocyte
	Actin	MetaCore	4.32e-5	
	Developmental regulation	GO	5.71e-11	
	Wounding response	GO	8.32e-10	
	Muscle contraction	GO	2.57e-8	
	CNS development	GO	2.12e-8	
	Angiogenesis	GO	6.86e-7	
MC66 2.2	Th17 cytokines	MetaCore	1.56e-10	Astrocyte
	Interferon	MetaCore	2.58-9	
	Complement	MetaCore	2.96e-9	
	Immunity	GO	7.96e-25	
	Wound response	GO	1.95e-23	
	Inflammation	GO	2.85e-22	
	Leukocyte chemotaxis	GO	5.91e-15	
MC66 2.3	Muscle development	GO	8.18e-7	Astrocyte
	Nucleosome assembly	GO	3.48e-6	
	Chromatin silencing	GO	4.03e-6	
MC66 3.1	Nuclear protein export	GO	2.88e-5	Neuron
	Transcription upregulation	GO	3.06e-5	
MC66 3.2	Synaptogenesis	MetaCore	6.47e-7	Neuron
	RAB3 regulation	MetaCore	7.25e-6	
	Transmission of nerve impulse	MetaCore	7.72e-6	
	Nerve impulse transmission regulation	GO	6.11e-13	
	Neurotransmitter secretion	GO	4.86e-12	
	Cellular localisation regulation	GO	2.12e-8	
	Synaptic plasticity regulation	GO	2.41e-8	
	Regulation of presynaptic vesicle fusion	GO	1.58e-7	
MC66 3.3	Bleb formation	GO	4.94e-6	Oligodendrocyte
	GABA secretion	GO	2.08e-5	
	K+ transport	GO	2.12e-5	
	Dopamine	GO	4.73e-5	
	Serotonin	GO	6.12e-5	

Table 4.9. Cahoy cell type and MetaCore functional categories enriched in the 9 low-level subclusters of the Dobrin 3093 expression cluster

Cluster name	MetaCore GO biological process categories enriched in cluster	Database	P-value	Cell type lists enriched in subcluster
Dobrin 1.1	Protein localization	GO	2.66e-5	Oligodendrocyte
Dobrin 1.2	Cell adhesion (amyloid proteins)	MetaCore	1.65e-4	Astrocyte
	EGF pathway regulation	GO	3.23e-7	Oligodendrocyte
	Nucleosome assembly	GO	2.10e-6	
	Memory	GO	3.39e-6	
	Anti-apoptosis	GO	4.68e-5	
Dobrin 1.3	Biopolymer metabolism	GO	3.60e-6	Oligodendrocyte
	Helicases	GO	2.16e-5	
Dobrin 2.1	Neurogenesis	MetaCore	1.87e-5	Neuron
	Angiogenesis	MetaCore	7.37e-5	
	Muscle contraction	GO	5.85e-8	
	Activation of phospholipase C by acetylcholine	GO	1.65e-6	
	Axonogenesis	GO	1.70e-6	
	Negative regulation of adenylate cyclase	GO	1.95e-6	
	Kinase regulation	GO	2.38e-6	
Dobrin 2.2	GABA neurotransmission	MetaCore	1.31e-4	Neuron
	GABA-A receptor life cycle	MetaCore	1.15e-6	
	Synaptic vesicle exocytosis	MetaCore	6.93e-5	
	CNS development	GO	1.33e-7	
	Neurotransmitter transport	GO	2.35e-7	
	Synaptic transmission	GO	5.70e-7	
	Microtubule-based movement	GO	1.14e-6	
	Memory	GO	1.61e-6	
Dobrin 2.3	MIF signaling	MetaCore	7.45e-6	Neuron
	Transmission of nerve impulse	MetaCore	7.39e-5	
	Synaptic transmission	GO	5.40e-10	
	Ion transport	GO	8.63e-8	
	Synaptogenesis	GO	1.33e-7	
	Vesicle fusion	GO	5.24e-7	
	Cellular insulin response	GO	5.26e-7	
Dobrin 3.1	Lysine metabolism	MetaCore	7.32e-6	Astrocyte
	Angiogenesis	MetaCore	5.36e-5	
	Angiogenesis	GO	8.12e-13	
	CNS development	GO	2.69e-7	
	Cell proliferation regulation	GO	6.13e-7	
	Inflammatory response	GO	4.14e-6	
	Nitric oxide mediated signal transduction	GO	4.86e-6	
Dobrin 3.2	Platelet-endothelium-leukocyte interactions	MetaCore	8.38e-11	Astrocyte
	Angiogenesis	MetaCore	1.44e-8	
	Interferon	MetaCore	1.21e-7	
	Th17 cytokines	MetaCore	4.62e-7	
	Wounding response	GO	2.42e-23	
	Development	GO	7.29e-22	
	Inflammatory response	GO	2.24e-19	
	Leukocyte migration	GO	2.96e-15	
	Cell proliferation regulation	GO	1.45e-14	
Dobrin 3.3	Interleukin	MetaCore	4.90e-6	Oligodendrocyte
	HGF signalling	MetaCore	4.58e-5	
	Transcription regulation	GO	2.87e-10	
	Kinases	GO	3.84e-6	
	Regulation of neuron apoptosis	GO	4.86e-5	

Dobrin subcluster 2.2 and MC66 subcluster 1.3 are also enriched for genes linked to GABA neurotransmission, nervous system development, synaptic vesicles and synaptic transmission, as are their parent subclusters.

MC66 subcluster 1.3 was used with the Build Network function in MetaCore. As the number of genes was large (392 genes recognised in MetaCore), only direct interactions between genes were included on the network. In addition, interactions from the 2-step correlation network for the subcluster were included in the network (see Section 4.3.2). Those genes which interacted with each other were left in the network, and other genes removed (Figures 4.6 and 4.8).

In these figures, green circles indicate genes linked to schizophrenia or bipolar disorder according to MeSH, red circles indicate genes associated with schizophrenia or bipolar disorder in the WTCCC study (153, 154) and blue circles indicate genes differentially expressed between cases and controls in the Stanley database (116). Although all three are spread throughout the network in schizophrenia, they appear to be particularly concentrated in the top left area of Figure 4.6. This coincides with the area where genes annotated as involved in synaptic transmission in GO are found.

However, this may also reflect the difference in the types of data used – the top left mainly contains genes with interactions drawn from MetaCore, while the bottom right is mainly interactions from the two-step expression correlation network. It is possible that the genes with links in MetaCore are better studied than the expression correlation network genes, but genes annotated with the 'transport' GO category were spread much more evenly around the network. This suggests that both areas have some level of GO annotation, and so the placement of genes annotated as synaptic transmission related alongside genes associated with or differentially expressed in schizophrenia may have functional significance.

The genes in the network also show enrichment for central nervous system development genes ($p=3.4e^{-6}$). Again, these show some bias for the genes with links in MetaCore, but this is not so pronounced as with the synaptic transmission genes.

Dobrin subcluster 2.2, which overlaps with MC66 subcluster 1.3 (70% genes shared – Tables 4.4 and 4.5) was also used to build a network in MetaCore (Figures 4.7 and 4.9). Again, interactions from the 2-step correlation network for the

subcluster were included. Red circles indicate schizophrenia associated genes in the WTCCC study, green circles indicate genes annotated as schizophrenia-linked genes in MeSH, and blue circles indicate genes listed as being differentially expressed in schizophrenia in the Stanley database.

Most SCZ-related genes are in the top half of the network. This coincides with the placing of several GO categories, most obviously including CNS development, synaptic transmission and synaptic vesicle transport. Although the lower half of the network has fewer genes than the upper half, all genes there were included in both the WTCCC and Stanley studies, and so had the opportunity to be considered. Unlike schizophrenia, genes relevant to bipolar disorder are not concentrated in a particular part of the network (Figures 4.7 and 4.9). They appear both in regions linked together by edges based upon coexpression and regions linked by MetaCore interactions. It is also notable that several genes from this cluster are both differentially expressed in BP and associated with BP (AUMH, secretogranin, IDH3A, KAP3, AP180, UQCRFS1, neuroserpin, carbonic anhydrase X). This congruence strengthens the evidence that they underlie BP aetiology.

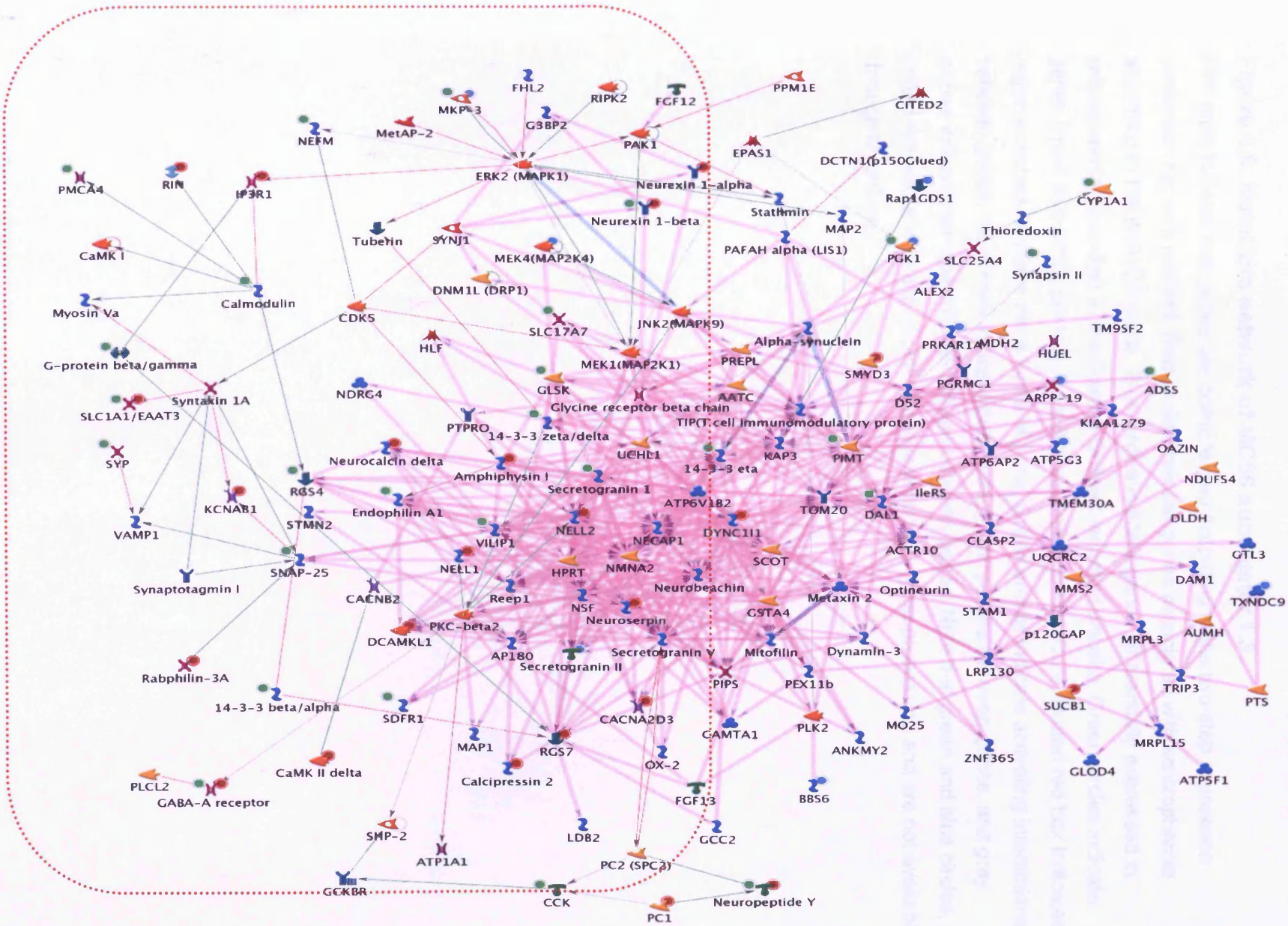

FHL2
RIPK2
FGF12
PPM1E
CITED2
MKP-3
G3BP2
NEFM
MetAP-2
PAK1
EPAS1
DCTN1(p150Glued)
PMCA4
RIN
IP3R1
ERK2 (MAPK1)
Neurexin 1-alpha
Rap1GDS1
CYP1A1
Neurexin 1-beta
Stathmin
MAP2
Thioredoxin
Tubulin
SYNJ1
PAFAH alpha (LIS1)
PGK1
CaMK I
MEK4(MAP2K4)
SLC25A4
Synapsin II
DNM1L (DRP1)
ALEX2
Myosin Va
Calmodulin
CDK5
JNK2(MAPK9)
TM9SF2
SLC17A7
HLF
Alpha-synuclein
PRKAR1A
MDH2
HUEL
MEK1(MAP2K1)
PREPL
G-protein beta/gamma
SMYD3
PCRMC1
Syntaxin 1A
GLSK
AATC
D52
ARPP-19
ADSS
NDRG4
Glycine receptor beta chain
SLC1A1/EAAT3
TIP(T cell immunomodulatory protein)
KIAA1279
PTPRO
14-3-3 zeta/delta
OAZIN
SYP
UCHL1
KAP3
PIMT
ATP6AP2
ATP5G3
Neurocalcin delta
Amphiphysin I
14-3-3 eta
NDUFS4
RGS4
Secretogranin 1
IleRS
KCNAB1
ATP6V1B2
TOM20
DAL1
TMEM30A
Endophilin A1
STMN2
DLDH
VAMP1
NELL2
DYNC1I1
CLASP2
VILIP1
NECAP1
ACTR10
UQCRC2
GTL3
NELL1
SCOT
SNAP-25
HPRT
NMNA2
DAM1
Synaptotagmin I
Neurobeachin
Optineurin
MMS2
CACNB2
Reep1
Metaxin 2
TXNDC9
NSF
GSTA4
STAM1
PKC-beta2
Neuroserpin
p120GAP
MRPL3
AUMH
DCAMKL1
AP180
Secretogranin V
Mitofilin
Dynamin-3
Secretogranin II
LRP130
Rabphilin-3A
PIPS
PEX11b
TRIP3
14-3-3 beta/alpha
SUCB1
PTS
SDFR1
CACNA2D3
CAMTA1
PLK2
MO25
MAP1
RGS7
ANKMY2
MRPL15
CaMK II delta
OX-2
ZNF365
GLOD4
PLCL2
Calcipressin 2
BBS6
ATP5F1
GABA-A receptor
FGF13
SHP-2
LDB2
GCC2
ATP1A1
PC2 (SPC2)
GCKBR
CCK
Neuropeptide Y
PC1

Figure 4.6. MetaCore network of MC66 subcluster 1.3

Pink lines indicate that genes are coexpressed according to the two-step expression correlation network method. Red circles indicate genes associated with schizophrenia according to the WTCCC study. Blue circles indicate genes differentially expressed in schizophrenia according to the Stanley Online Genomics website. Green circles indicate genes linked with schizophrenia in MeSH, according to MetaCore. Dotted red box indicates region enriched for these three types of gene. Green arrows indicate activating interactions between genes, red arrows indicate deactivating interactions between genes, and grey arrows indicate non-specific interactions between genes. Note that green and blue circles, dotted red box, and fading of pink lines were added in post-processing and are not available through MetaCore.

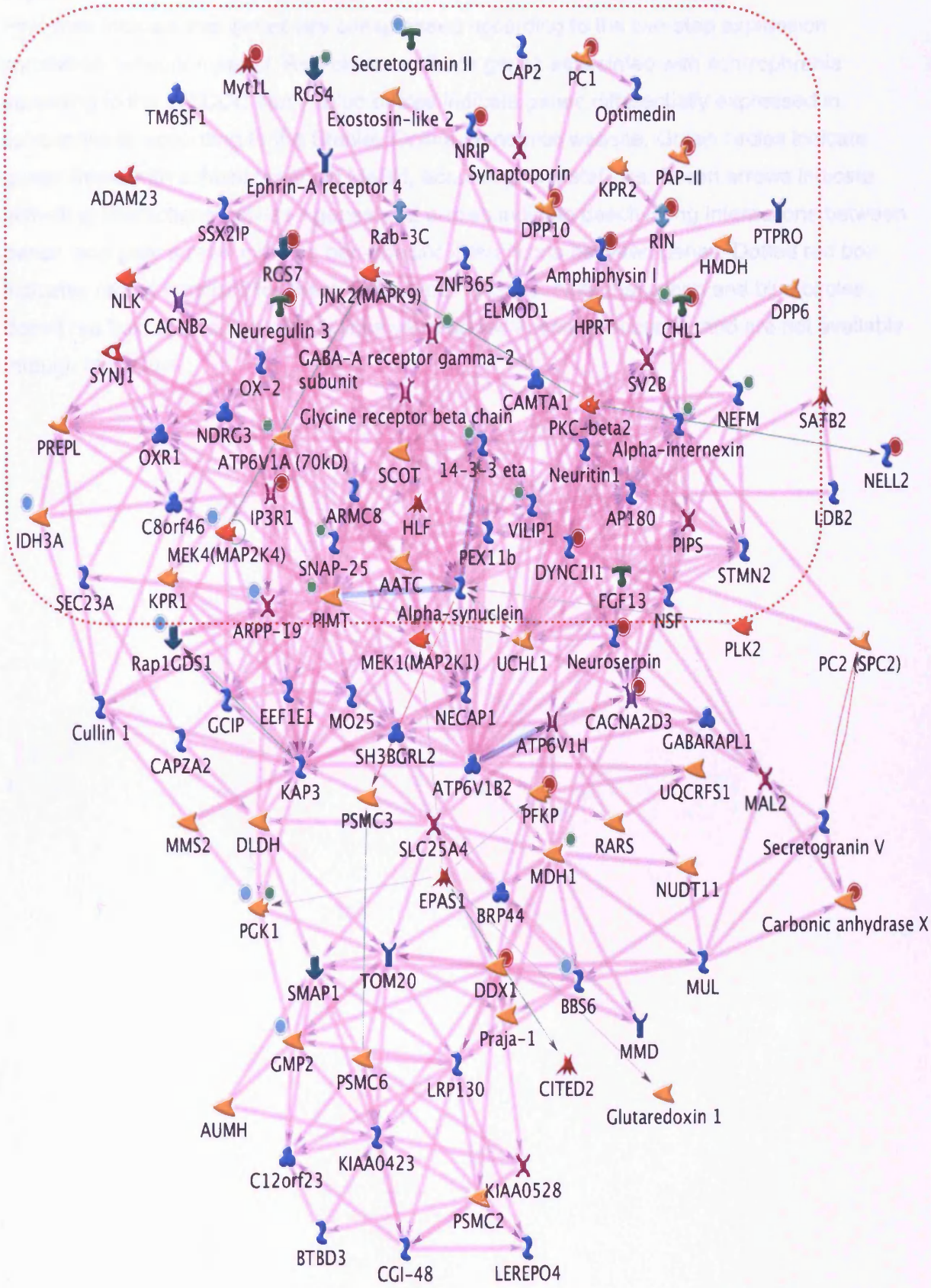

Secretogranin II
CAP2
PC1
Myt1L
RGS4
TM6SF1
Exostosin-like 2
Optimedin
NRIP
Synaptoporin
KPR2
PAP-II
ADAM23
Ephrin-A receptor 4
SSX2IP
Rab-3C
DPP10
RIN
PTPRO
RGS7
ZNF365
Amphiphysin I
HMDH
NLK
CACNB2
JNK2(MAPK9)
ELMOD1
HPRT
CHL1
DPP6
Neuregulin 3
SYNJ1
GABA-A receptor gamma-2 subunit
OX-2
SV2B
CAMTA1
NEFM
SATB2
Glycine receptor beta chain
PREPL
NDRG3
PKC-beta2
OXR1
ATP6V1A (70kD)
Alpha-internexin
NELL2
SCOT
14-3-3 eta
Neuritin1
IDH3A
C8orf46
IP3R1
ARMC8
HLF
VILIP1
AP180
LDB2
MEK4(MAP2K4)
PEX11b
PIPS
SNAP-25
DYNC1I1
STMN2
SEC23A
KPR1
AATC
FGF13
ARPP-19
PIMT
Alpha-synuclein
NSF
PLK2
Rap1GDS1
MEK1(MAP2K1)
UCHL1
Neuroserpin
PC2 (SPC2)
Cullin 1
GCIP
EEF1E1
MO25
NECAP1
CACNA2D3
GABARAPL1
CAPZA2
SH3BGRL2
ATP6V1H
KAP3
ATP6V1B2
UQCRFS1
MAL2
PSMC3
PFKP
MMS2
DLDH
SLC25A4
RARS
Secretogranin V
MDH1
NUDT11
EPAS1
BRP44
PCK1
Carbonic anhydrase X
SMAP1
TOM20
DDX1
BBS6
MUL
Praja-1
MMD
GMP2
PSMC6
LRP130
CITED2
Glutaredoxin 1
AUMH
KIAA0423
C12orf23
KIAA0528
PSMC2
BTBD3
CGI-48
LEREPO4

Figure 4.7. MetaCore network of Dobrin subcluster 2.2

Pink lines indicate that genes are coexpressed according to the two-step expression correlation network method. Red circles indicate genes associated with schizophrenia according to the WTCCC study. Blue circles indicate genes differentially expressed in schizophrenia according to the Stanley Online Genomics website. Green circles indicate genes linked with schizophrenia in MeSH, according to MetaCore. Green arrows indicate activating interactions between genes, red arrows indicate deactivating interactions between genes, and grey arrows indicate non-specific interactions between genes. Dotted red box indicates region enriched for these three types of gene. Note that green and blue circles, dotted red box, and fading of pink lines were added in post-processing and are not available through MetaCore.

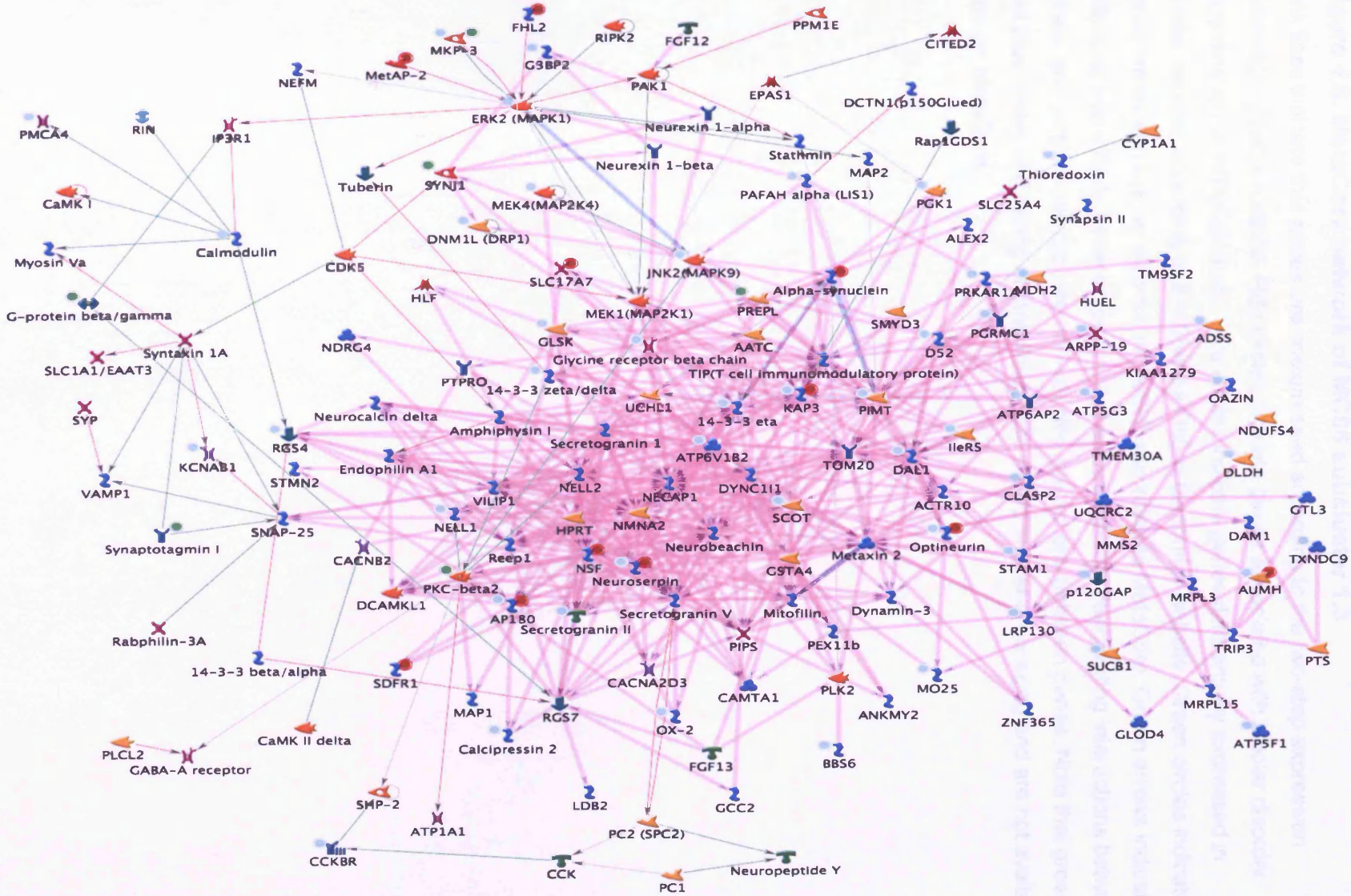

FHL2
RIPK2
FGF12
PPM1E
CITED2
MKP-3
G3BP2
NEFM
MetAP-2
PAK1
EPAS1
DCTN1(p150Glued)
PMCA4
RIN
IP3R1
ERK2 (MAPK1)
Neurexin 1-alpha
Rap1GDS1
CYP1A1
Stathmin
MAP2
Neurexin 1-beta
Thioredoxin
TubeHn
SYNJ1
PAFAH alpha (LIS1)
CaMK I
MEK4(MAP2K4)
PGK1
SLC25A4
Synapsin II
DNM1L (DRP1)
ALEX2
Myosin Va
Calmodulin
CDK5
JNK2(MAPK9)
TM9SF2
SLC17A7
Alpha-synuclein
PRKAR1A
MDH2
HUEL
HLF
MEK1(MAP2K1)
PREPL
G-protein beta/gamma
SMYD3
PGRMC1
Syntaxin 1A
GLSK
AATC
D52
ARPP-19
ADSS
NDRG4
Glycine receptor beta chain
TIP(T cell immunomodulatory protein)
KIAA1279
SLC1A1/EAAT3
PTPRO
14-3-3 zeta/delta
OAZIN
UCHL1
KAP3
PIMT
ATP6AP2
ATP5G3
SYP
Neurocalcin delta
14-3-3 eta
NDUFS4
Amphiphysin I
RGS4
Secretogranin 1
ileRS
TMEM30A
KCNAB1
ATP6V1B2
TOM20
DAL1
DLDH
Endophilin A1
STMN2
NELL2
DYNC1I1
VAMP1
VILIP1
NECAP1
ACTR10
CLASP2
UQCRC2
GTL3
SCOT
NELL1
HPRT
NMNA2
DAM1
SNAP-25
Neurobeachin
Optineurin
MMS2
Synaptotagmin I
CACNB2
Reep1
Metaxin 2
STAM1
TXNDC9
NSF
GSTA4
p120GAP
MRPL3
AUMH
PKC-beta2
Neuroserpin
DCAMKL1
AP180
Secretogranin V
Mitofilin
Dynamin-3
LRP130
Rabphilin-3A
Secretogranin II
PIPS
PEX11b
TRIP3
PTS
14-3-3 beta/alpha
SDFR1
SUCB1
CACNA2D3
MO25
MAP1
CAMTA1
PLK2
MRPL15
RGS7
ANKMY2
ZNF365
CaMK II delta
OX-2
GLOD4
PLCL2
Calcipressin 2
ATP5F1
GABA-A receptor
FGF13
BBS6
SMP-2
LDB2
GCC2
ATP1A1
PC2 (SPC2)
CCKBR
CCK
Neuropeptide Y
PC1

Figure 4.8. MetaCore network of MC66 subcluster 1.3

Pink lines indicate that genes are coexpressed according to the two-step expression correlation network method. Red circles indicate genes associated with bipolar disorder according to the WTCCC study. Blue circles indicate genes differentially expressed in bipolar disorder according to the Stanley Online Genomics website. Green circles indicate genes linked with bipolar disorder in MeSH, according to MetaCore. Green arrows indicate activating interactions between genes, red arrows indicate deactivating interactions between genes, and grey arrows indicate non-specific interactions between genes. Note that green and blue circles and fading of pink lines were added in post-processing and are not available through MetaCore.

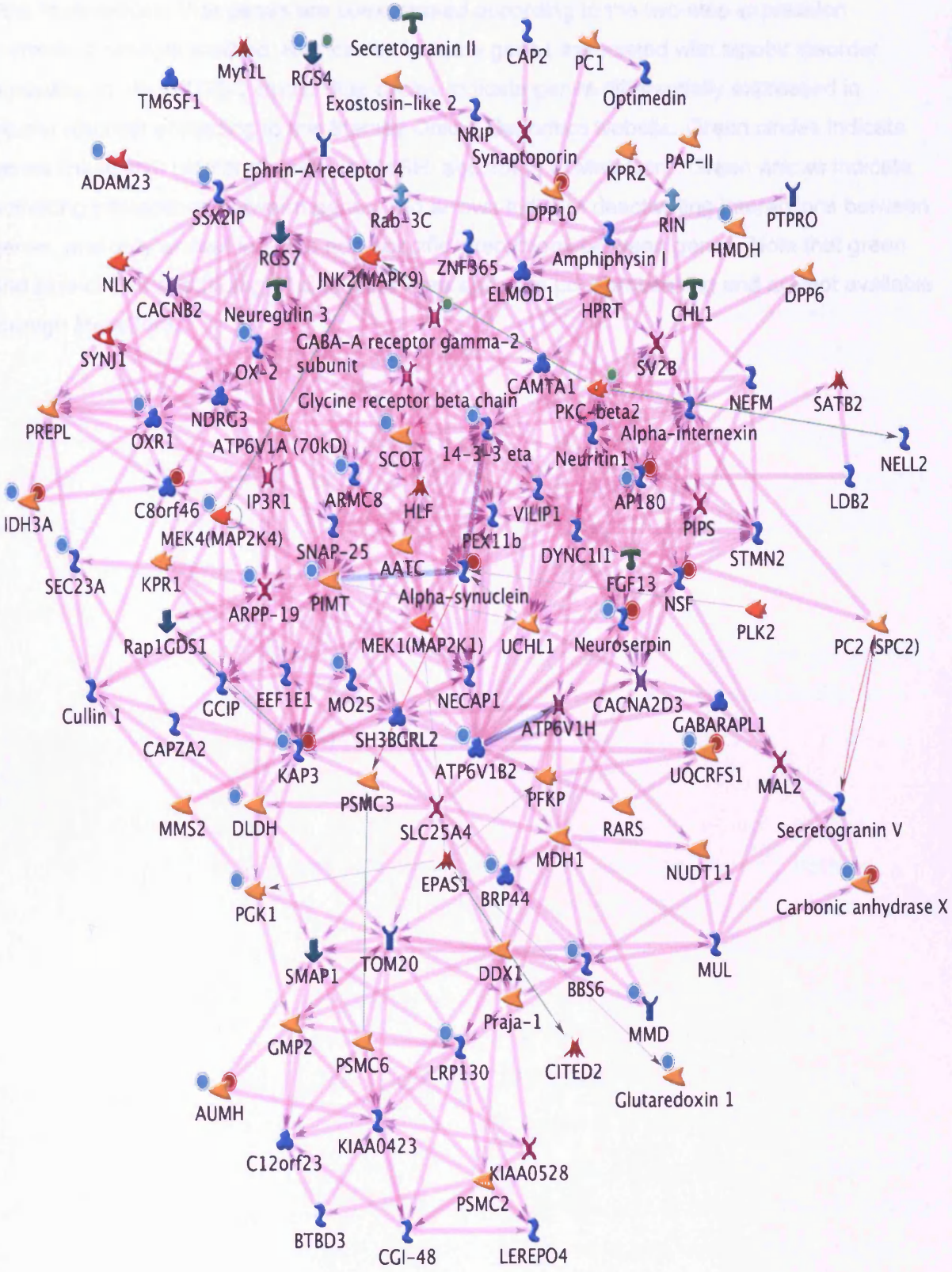

Secretogranin II
CAP2
PC1
Myt1L
RGS4
Optimedin
TM6SF1
Exostosin-like 2
NRIP
Synaptoporin
PAP-II
KPR2
ADAM23
Ephrin-A receptor 4
SSX2IP
Rab-3C
DPP10
RIN
PTPRO
HMDH
RGS7
ZNF365
Amphiphysin I
NLK
JNK2(MAPK9)
ELMOD1
DPP6
CACNB2
Neuregulin 3
HPRT
CHL1
GABA-A receptor gamma-2 subunit
SYNJ1
OX-2
SV2B
CAMTA1
NEFM
SATB2
Glycine receptor beta chain
PREPL
OXR1
NDRG3
PKC-beta2
Alpha-internexin
NELL2
ATP6V1A (70kD)
SCOT
14-3-3 eta
Neuritin1
IDH3A
C8orf46
IP3R1
ARMC8
HLF
VILIP1
AP180
PIPS
LDB2
MEK4(MAP2K4)
SNAP-25
PEX11b
DYNC1I1
STMN2
SEC23A
KPR1
AATC
FGF13
NSF
ARPP-19
PIMT
Alpha-synuclein
PLK2
Rap1GDS1
MEK1(MAP2K1)
UCHL1
Neuroserpin
PC2 (SPC2)
Cullin 1
GCIP
EEF1E1
MO25
NECAP1
CACNA2D3
GABARAPL1
CAPZA2
SH3BGRL2
ATP6V1H
KAP3
ATP6V1B2
UQCRFS1
MAL2
PSMC3
PFKP
MMS2
DLDH
SLC25A4
RARS
Secretogranin V
MDH1
EPAS1
NUDT11
BRP44
Carbonic anhydrase X
PGK1
SMAP1
TOM20
DDX1
MUL
BBS6
Praja-1
MMD
GMP2
PSMC6
LRP130
CITED2
Glutaredoxin 1
AUMH
KIAA0423
C12orf23
KIAA0528
PSMC2
BTBD3
CGI-48
LEREPO4

Figure 4.9. MetaCore network of Dobrin subcluster 2.2

Pink lines indicate that genes are coexpressed according to the two-step expression correlation network method. Red circles indicate genes associated with bipolar disorder according to the WTCCC study. Blue circles indicate genes differentially expressed in bipolar disorder according to the Stanley Online Genomics website. Green circles indicate genes linked with bipolar disorder in MeSH, according to MetaCore. Green arrows indicate activating interactions between genes, red arrows indicate deactivating interactions between genes, and grey arrows indicate non-specific interactions between genes. Note that green and blue circles and fading of pink lines were added in post-processing and are not available through MetaCore.

4.3.6 MC66 subcluster 2 and Dobrin subcluster 3

Another interesting pair of clusters is MC66 subcluster 2 and Dobrin subcluster 3, which share over 50% of their genes. Both of these clusters show enrichment for genes differentially expressed in schizophrenia (MC66 p=0.0003, Dobrin p=0.011), and Dobrin subcluster 3 is also enriched for genes associated with schizophrenia (p=0.035). Both clusters are also enriched for genes upregulated in astrocytes, and are enriched for functional categories relating to angiogenesis, CNS development and immune response (Table 4.5).

Of the three component subclusters of Dobrin subcluster 3, subcluster 3.1 is the most interesting. It has increased enrichment for SCZ associated (p=0.0037) and differentially expressed (p=1.4e-5) genes compared to its parent subcluster, and also shows some enrichment for BP differentially expressed genes (p=0.022). It shares 39.6% of its genes with MC66 subcluster 2.1, which is also heavily enriched for SCZ associated genes and SCZ differentially expressed genes. The other subclusters from MC66 subcluster 2 and Dobrin subcluster 3 are not enriched for SCZ or BP related genes.

Both MC66 subcluster 2.1 and Dobrin subcluster 3.1 are enriched for genes relating to angiogenesis and CNS development. The Dobrin subcluster is also enriched for immune response genes, like the parent subclusters.

4.3.7 MC66 subcluster 3.1

Few other subclusters show significant enrichment for schizophrenia or bipolar disorder differentially expressed or associated genes. One that does is MC66 subcluster 3.1. This cluster is enriched for schizophrenia associated genes (p=0.0039), schizophrenia differentially expressed genes (p=0.01) and bipolar disorder associated genes (p=0.024). It does not show any overlap with any other subcluster, although this may be because this subcluster is small (78 genes).

The small size would make this subcluster suitable for network manipulation in GeneGO, but there are only two pairs of genes with canonical links between them in the subcluster. The genes in this subcluster are less well annotated than the other subclusters, with an average GIFtS value for genes in this cluster of 51.3. This is below the average GIFtS value for the top-level MC66 2546 cluster (51.7). However,

the difference between the two sets of GIFtS values was not significant (t-test p = 0.68).

This suggests that the enrichment for disease-related genes is unlikely to be due to a concentration of highly investigated genes, and may explain the lack of links between them. The cluster is enriched for genes upregulated in neurons, and slightly enriched for a small number of GO categories (nuclear protein export, transcription upregulation).

4.3.8 Enrichment for gene-length adjusted association lists

Subsequent to the other analyses, the Dobrin 3093 and MC66 2546 clusters and their subclusters were analysed for enrichment with lists of bipolar disorder (WTCCC) and schizophrenia (ISC and UK) associated genes, corrected for the effects of gene length (Table 4.10). The clusters and subclusters were much less enriched for these gene lists than they were for those not adjusted for gene length, suggesting that bias due to gene length may have been one reason for the previous enrichments for disease associated genes.

However, some subclusters were nominally significantly enriched for WTCCC bipolar associated genes (MC66 1.3, MC66 2.3), although these do not survive Bonferroni correction for multiple testing with 26 subclusters (MC66 1.3 Bonferroni corrected p = 0.488, MC66 2.3 corrected p = 0.7). The MC66 1 and MC66 1.1 subclusters were nominally significantly enriched for ISC schizophrenia associated genes, and the MC66 1.1 subcluster remains significantly enriched after Bonferroni correction (Bonferroni corrected p = 0.017). Other subclusters were close to nominal significance for WTCCC bipolar associated genes (Dobrin 3093, Dobrin 3093 subcluster 2) or ISC schizophrenia associated genes (MC66 2546). No clusters or subclusters were enriched for genes associated with schizophrenia according to the UK schizophrenia study. This may be because the small size of this study limits its power to detect true associations.

MC66 subcluster 1.1 may be worth investigating further for relevance to schizophrenia. It is enriched for relatively few MetaCore or GO functional categories. However, it is significantly enriched for a MetaCore map relating to BRCA1 and DNA damage (Table 4.8). It overlaps with Dobrin subcluster 1.1, which has a similarly low number of enriched functional categories (Tables 4.4 and 4.5), but which is not

significantly enriched for schizophrenia associated genes. The average GIFtS value of genes from both these subclusters is lower than the average GIFtS value of genes in the parent Dobrin 3093 and MC66 2546 expression clusters, perhaps suggesting that these subclusters contain relatively poorly annotated genes that could benefit from more study.

Table 4.10. Enrichment of clusters and subclusters for schizophrenia and bipolar disorder associated genes, adjusted for length

Cluster name	ISC schizophrenia associated genes, p<0.05	UK schizophrenia associated genes, p<0.05	WTCCC bipolar disorder associated genes, p<0.05
Dobrin 3093	0.281	0.907	0.061
Dobrin 3093 1	0.159	0.831	0.431
Dobrin 3093 1.1	0.168	0.871	0.674
Dobrin 3093 1.2	0.745	0.868	0.105
Dobrin 3093 1.3	0.210	0.514	0.725
Dobrin 3093 2	0.565	0.450	0.068
Dobrin 3093 2.1	0.534	0.616	0.197
Dobrin 3093 2.2	0.586	0.329	0.339
Dobrin 3093 2.3	0.587	0.593	0.150
Dobrin 3093 3	0.540	0.933	0.254
Dobrin 3093 3.1	0.963	0.481	0.752
Dobrin 3093 3.2	0.556	0.993	0.483
Dobrin 3093 3.3	0.111	0.630	0.074
MC66 2546	0.058	0.532	0.166
MC66 2546 1	**0.009**	0.874	0.190
MC66 2546 1.1	**0.0007**	0.734	0.970
MC66 2546 1.2	0.491	N/A	0.164
MC66 2546 1.3	0.530	0.738	**0.019**
MC66 2546 2	0.538	0.623	0.378
MC66 2546 2.1	0.827	0.342	0.939
MC66 2546 2.2	0.934	N/A	0.684
MC66 2546 2.3	0.115	0.572	**0.027**
MC66 2546 3	0.716	0.350	0.311
MC66 2546 3.1	0.644	0.588	0.316
MC66 2546 3.2	0.676	0.415	0.460
MC66 2546 3.3	0.668	0.408	0.421

4.4 Discussion

4.4.1 Recurrent themes in the functional categories of the clusters

A number of functional categories appear in the clusters and subclusters from both the Dobrin and MC66 datasets that are enriched for genes relevant to schizophrenia and bipolar disorder (Tables 4.2 and 4.4-4.9). The large size of MC66 2546 and Dobrin 3093 clusters makes inference about individual genes difficult. However, both the larger clusters are enriched for genes present in the KEGG MAP kinase pathway, suggesting that this pathway may relate to their function, and possibly to the aetiology of schizophrenia. Members of this pathway have also been found to be differentially expressed between controls and schizophrenics in other brain regions (164). In addition, when structural variants such as microdeletions occur in the genomes of schizophrenics, some evidence suggests that they are particularly likely to occur in the genes of the MAP kinase pathway (165).

However, the MAP kinase-related genes present in the two large clusters do not overlap with the schizophrenia associated gene set or the differentially expressed in schizophrenia gene set (they share no genes at all in either the MC66 or Dobrin cluster). This might suggest the MAP kinase function of the clusters may be incidental to their roles in schizophrenia aetiology. Further investigation with other functional analysis tools (both free and commercial) may reveal more about these clusters, and would be a good avenue for further study

Of the subclusters, the Dobrin 2.2 and MC66 1.3 subcluster pair are both enriched for genes associated with schizophrenia and bipolar disorder according to the original analysis. The MC66 1.3 subcluster is also enriched for genes annotated as schizophrenia related, according to MeSH and MetaCore.

These results were somewhat undermined by the subsequent enrichment analysis using association values that took into account varying gene lengths. Although the Dobrin 2.2 subcluster was no longer significantly enriched for bipolar or schizophrenia associated genes in this analysis, the MC66 1.3 subcluster remained nominally enriched for bipolar disorder associated genes. As both subclusters are also highly significantly enriched for genes differentially expressed in bipolar disorder (Tables 4.8 and 4.9), they retain some of their potential relevance to bipolar aetiology despite the reduction in significance for association enrichment.

The two subclusters are also both annotated with several MetaCore and GO terms that could plausibly play a role in the aetiology of schizophrenia and bipolar disorder. These include GABA neurotransmission (166, 167), nervous system development and synaptic vesicles (168-170).

Evidence for the GABA hypothesis of schizophrenia has mounted over the past few years, especially expression-based evidence that shows a reduction in GABA receptors across multiple brain areas (171). Reductions in GAD1, in particular, have been found in dorsolateral prefrontal cortex, anterior cingulate region, and hippocampus (166). GABA also has strong influence over dopamine neurotransmitter levels, allowing the GABA hypothesis of schizophrenia to exist alongside the dopamine hypothesis.

Study of synaptic vesicles in schizophrenia and bipolar disorder has primarily focused on particular genes involved in synaptic vesicle function, like SNARE and SVMT (168, 169). However, some putative schizophrenia-related genes, such as dysbindin and DISC1, are also linked with synaptic vesicle function (172-174). Furthermore, the synaptic vesicle exocytosis MetaCore network is extremely heavily enriched with genes annotated as schizophrenia genes in MeSH (43.75% of the genes in the network, MetaCore $p=1e^{-17}$). This enrichment is likely to be at least partly due to annotation bias, as a neuronal process like synaptic vesicle exocytosis is an obvious place to look for a relationship to schizophrenia aetiology. However, this may not be the only cause of the enrichment, as the existence of gene coexpression clusters derived from two independent datasets which are enriched for both synaptic vesicle genes and genes related to schizophrenia suggests.

The Dobrin subcluster 3.1 and the MC66 subcluster 2.1 are enriched for genes differentially expressed in or associated with schizophrenia, and also show enrichment for CNS development genes (Tables 4.8 and 4.9). In addition to this, they are enriched for angiogenesis genes. There is little evidence linking angiogenesis and schizophrenia, but one study has found that SNPs near BAI3, an angiogenesis inhibitor, are associated with the severity of disorganised symptoms in schizophrenia (175).

This work could be expanded and extended in a number of ways. The Dobrin 3093 and MC66 2546 clusters and their subclusters could be further tested for a functional relationship with schizophrenia by testing for enrichment with the results of

other GWAS studies (such as the ISC (29)) or other differential expression studies. The clusters could also be used to define eQTLs for polygenic score analysis, as was done with the MC66 1.3 subcluster (see Chapter 3, Section 3.2.5), although the top eQTL SNP lists for that subcluster were not generally superior to the bottom eQTL SNP lists at discriminating between cases and controls through polygenic.

4.4.2 MetaCore

MetaCore demonstrated several strengths in this study. In some subclusters, the MetaCore maps and networks showed enrichments for functional categories that the standard GO categories did not detect or did not rank so highly. The most obvious example of this is in the MC66 subcluster 1.3 and the Dobrin subcluster 2.2. In these overlapping subclusters, the significant enrichment for genes relating to GABA neurotransmission only appeared in the MetaCore maps and networks, not the GO categories (Tables 4.6 and 4.7, Supplemental Data 1).

The MetaCore networks themselves were also useful (Figures 4.6 to 4.9). Their capacity to be annotated with external data (such as the lists of schizophrenia associated genes) and internal MetaCore data (such as the MeSH schizophrenia genes) enables the identification of subsets of the network enriched for these disease-relevant genes through visual inspection. However, as this analysis is not statistically founded, care must be taken to ensure that such areas do not appear enriched simply because of the arrangement of genes in the network (e.g. due to some areas having a higher density of genes than others).

MetaCore does have some weaknesses. The visualisation options when viewing networks do not give the user much control over the appearance of the graph. Although this is usually only a cosmetic problem, it can sometimes hamper interpretation, such as forcing the use of thick, pink lines to represent user-defined interactions (concentrations of which tend to obscure other interactions). Also, it is only possible to annotate the network with one set of genes at a time. In Figures 4.6 to 4.9, it was necessary to use an image editor to fade the pink lines, brighten the other lines, and to add additional annotation through the use of green and blue circles.

Also, previous methods used by MetaCore to determine enrichment significance appeared to be biased towards finding significant results, even finding near-significant functional categories when the same gene list was used as target and background. However, this issue has been solved in the latest version of MetaCore – enrichment analyses using the same gene list for target and background now find no results.

MetaCore has a wide array of capabilities, and this study has only used a small proportion of them. It would be interesting to investigate to what extent the disease-related genes in the Dobrin and MC66 clusters and subclusters overlap with the different functional categories for which they are enriched. Currently it is a challenge to do this systematically, due to the inability of MetaCore to output the genes of networks as plain text or to annotate networks with freely chosen functional categories. However, it may be possible to use its ability to export to the external network visualisation program Cytoscape to circumvent these limitations.

4.4.3 GIFtS

Ascribing mean GIFtS values to clusters enabled me to exclude annotation bias as a probable reason for enrichment (e.g. MC66 subcluster 3.1, see Section 4.3.7 above). The lack of interactions present between genes of this cluster in MetaCore might suggest that the cluster is a chance finding with no biological relevance, especially as it does not overlap heavily with any Dobrin subclusters. However, a lower mean GIFtS value than the MC66 2546 cluster implies that it cluster contains less well annotated and studied genes, explaining the lack of MetaCore interactions and alternatively suggesting that focusing future investigations on the genes of this cluster may be particularly fruitful.

The lack of significant enrichments for this cluster when using association gene lists adjusted for gene length undermines this conclusion, though, and suggests that the enrichments of MC66 subcluster 3.1 may be due only to chance. However, MC66 subcluster 1.1, which is enriched for schizophrenia associated genes after adjustment for gene length and which also lacks many enrichments for genes from particular functional categories, has a lower GIFtS value than the parent

MC66 2546 cluster. This again suggests that genes in this subcluster could benefit from further study.

The work on GIFtS could be continued in several ways. Firstly, a more structured study examining the effects of using mean GIFtS to define the annotation level of expression clusters and other gene groups would be a necessary first step before GIFtS could be used like this on a larger scale. Comparing and contrasting GIFtS to other measures of annotation level, such as a count of the number of PubMed abstracts a gene is mentioned in, would also be useful.

Many of the data types which underlie GIFtS values are likely to be correlated. For example, it would be highly unlikely to have data on the SNPs a gene contains without first knowing the sequence of the gene. Using principle component analysis to analyse the underlying binary annotation data would allow the determination of which data types best distinguish between levels of annotation. This may also allow for improved definitions of annotation level, and possibly the creation of meaningful dimensions or subtypes of annotation level.

4.4.4 Enrichment for length adjusted associated gene lists

Generally, the enrichments for length adjusted associated gene lists were much less significant than the enrichments for the non-length adjusted gene lists (Table 4.10). As both the Dobrin 3093 and MC66 2546 clusters are enriched for genes with upregulated expression in neurons, it is possible that the presence of a large number of long neuronal genes biased the clusters and subclusters toward enrichment.

However, some subclusters retained some enrichment for disorder-associated genes, such as MC66 subcluster 1.3 (see Section 4.4.1 above). Others, such as MC66 subcluster 1 and 1.1 were much more enriched when gene length was corrected for and associated genes based upon the ISC were used. MC66 subcluster 1.1 has few enrichments for functional annotations, but MC66 subcluster 1 is enriched for GABA receptors, nervous system development and synaptic contact (Table 4.7).

It is also possible that the permutation-based adjustment for gene length biases too heavily against long genes (see Section 4.2.8 above). For example, a true association signal which exists at one end of a gene could be diluted more heavily if that gene was long. This is because a long gene is likely to contain more non-

associated, independent SNPs, each of which will have the opportunity to obtain a more significant p-value than the true signal during the permutations.

It would be interesting to examine the correlation between average gene length of a cluster or subcluster and significance of enrichment for disorder associated genes (using adjustment for gene length). A negative correlation between gene length and significance might indicate that the adjustment for gene length was too stringent, and is biasing the method in favour of shorter genes.

4.4.5 Two-step correlation network construction

The results for the two-step correlation networks supported the hypothesis that limiting network edge addition to regions of the network where reliable edges already exist leads to networks that are more robust to noise. The similarities between the networks produced from randomly perturbed and unperturbed data were much greater for the two-step correlation networks than the one-step correlation networks. Also, using the two-step correlation networks allowed a considerable increase in the number of genes on the network for a relatively modest reduction in significance of the most significant GO category. This suggests that two-step expression correlation network construction should be considered as an alternative to the standard one-step method in future work.

The method could be extended in a number of ways. Adding further steps to the method is a possibility – permitting increasingly lax thresholds as the number of already-linked transcripts the expression of a gene correlates with increases. Using alternative data sources (such as MetaCore interactions) to seed the network with reliable edges (in addition to those seeded by the more stringent correlation threshold) might also increase the quality of the network produced.

The effect of using correlation thresholds based upon standard deviations above the mean absolute correlation would also be interesting to examine further. Using standard deviations allows for maximum comparability between two-step expression correlation networks produced from different sources, as regardless of the absolute extent of correlation, similar proportions of nodes will be linked by edges. However, it does not allow for the comparison of absolute correlation

between networks, and means that adding or removing nodes from a network could affect the presence or absence of other edges.

Using highly significant negative correlations to link genes is another way in which the method could be extended. However, although this would allow the inclusion of connections where a negative correlation does indicate shared function, it may also include connections between genes that have opposing functions (e.g. genes from opposite sides of the cell cycle). It may be a better strategy to use positive correlations with the two-step method to determine low-level connections between genes and to form these genes into groups, then to use one-step positive and negative correlations between average expression values to show relationships among the groups. Positive and negative correlation of average expression could also be used to form a network from the clusters produced by memISA and k-means.

Extending the two-step method to report the average correlation among genes of a network, or giving the user a choice between absolute and relative correlation thresholds, might be useful. Alternatively, the method could also calculate the significance of each correlation, giving the option of using a significance threshold.

Chapter Five

General Discussion

5.1 Expression data and psychiatric disease

A very large quantity of expression data has been produced through microarray analysis in the last 15 years (176). However, it has proven difficult to translate these data into reproducible results that have an impact on the understanding of human psychiatric disorders such as schizophrenia (177).

Part of the problem in the early days of large-scale expression analysis was the focus on finding genes that are differentially expressed between disease states. Individual genes found through differential expression analysis often fail to replicate in other studies or between different microarray platforms (178). This is partly due to the multiple testing burden inherent to testing thousands or tens of thousands of genes, and partly due to the relatively small sample sizes of microarray experiments. This leads to a high proportion of false positive results.

This is issue is further compounded by the diagnostic uncertainty affecting some neuropsychiatric disorders, especially schizophrenia (36). Currently, it is uncertain the extent to which schizophrenia represents a unified disease state with a common aetiology or a cluster of syndromes with related symptoms but a variety of causes. Some evidence suggests the former, such as the age of onset occurring in adolescence or early adulthood in the majority of schizophrenia cases. Other evidence points toward the latter, such as the range of symptoms experienced by schizophrenia patients or the existence of schizophrenia symptoms as part of rare developmental syndromes with specific genetic causes, such as velocardiofacial syndrome, caused by a deletion of chromosome 22q11.2 (179).

This diagnostic uncertainty reduces the power of microarray experiments to detect differentially expressed transcripts, as the expression of a gene may only vary in a specific subset of schizophrenia cases. It also reduces the reproducibility of differential expression results, as a result may depend upon the mix of subtypes of schizophrenia cases present in a dataset, which may not exist in independently

derived datasets. Examining methods that look beyond the individual differentially expressed genes was therefore thought to be particularly appropriate in psychiatric diseases.

5.2 Comparison of clustering methods

These issues with differential expression analysis, have led to a greater focus upon clustering genes according to the similarity of their expression profile. Clustering methods can aid understanding of a dataset by simplifying the data from tens of thousands of genes to a much smaller number of gene expression clusters. They can also be a source of inference in their own right, as similarities in expression can indicate similarity in biological function (180).

In Chapter 2, I examined the utility of four different clustering methods when applied to three brain gene expression datasets – k-means clustering, Chinese Restaurant Clustering (CRC), the Iterative Signature Algorithm (ISA), and the Memory Iterative Signature Algorithm (memISA) (66, 67, 181). Of the four methods, memISA produced the highest percentage of enriched clusters, but these clusters only included a relatively small percentage of available genes. K-means clustering produced a slightly higher percentage of enriched clusters than CRC while assigning a cluster to every gene. The failure of CRC to outperform the simpler k-means clustering suggests that relatively simple methods may be more effective when used upon data with a particularly complex coexpression structure, such as brain expression data (42).

CRC and k-means clustering found similar sets of clusters, while memISA and ISA produced different cluster sets to them. This was primarily due to the ability of memISA and ISA to find clusters that only exist in subsets of the available samples. Therefore, combining the clusters found by k-means clustering and memISA into a single set is a good strategy to find the widest possible selection of expression clusters in a dataset.

It is notable that one of the simplest clustering methods, k-means clustering, was the most effective upon brain gene expression data. It is possible that the simplicity of this method, which makes minimal assumptions about the structure of the data, is particularly well suited to dealing with highly complex brain expression data.

This work could be expanded to examine a variety of other clustering methods. In particular, network-based methods, such as the Generalised Topological Overlap Matrix (GTOM) were not considered (182). It would also be interesting to see what effects the two-step network calculation method discussed in Chapter 4 would have upon the effectiveness of these methods.

5.3 eQTLs and the relevance of expression data to psychiatric disease

In addition to clustering and differential expression analysis, expression data can be used with SNP data from the same samples to calculate 'expression quantitative trait loci' (eQTLs). eQTLs are SNPs where there is evidence that the allele present has an effect on the expression of a gene transcript (112). They are found by regressing expression transcript level on genotype.

In Chapter 3, I investigated whether SNPs which affect gene expression were more effective at predicting the schizophrenia affected status of samples from a separate dataset through polygenic score analysis. This method aggregates schizophrenia risk information across a large number of common SNPs associated with affected status at a lax threshold ($p<0.5$). Previous studies show that polygenic score is significantly higher in cases than in controls, suggesting that these common alleles of small effect play a role in schizophrenia aetiology (29, 104).

My results showed that these expression-affecting SNPs were in fact superior at predicting case/control status. Demonstrating that SNPs which affect gene expression are better predictors of schizophrenia implies that expression may be a mechanism by which common alleles influence schizophrenia risk. This has several implications for research into psychiatric disorders. Firstly, it supports the hypothesis that common SNP variants of limited effect are the primary genetic driver of schizophrenia development, rather than rarer variants of greater effect. Secondly, it validates the potential utility of expression data in the study of schizophrenia, suggesting that expression clustering and differential expression may also have relevance to the aetiology of the disorder.

Lastly, it shows that it is possible to use expression data to enhance the analysis of GWAS data for neuropsychiatric conditions. Although the analysis in this study deals with aggregated association and expression data, it may also be possible to use them together to investigate specific genes. For example, if a gene

which contained SNPs nominally, but not genome-widely, significantly associated with schizophrenia could be shown to affect the expression of a gene that is genome-wide significantly associated with schizophrenia, the evidence for both genes being true positives is strengthened.

Other authors have combined expression data with GWAS data to investigate a variety of medical conditions. Schadt *et al* used liver expression data to provide evidence to prioritise candidate genes for coronary artery disease and type I diabetes found through GWAS (122). This differed from the work here in that it focused on individual association loci, rather than using a method like polygenic score analysis to aggregate association data from across the genome. Nicolae *et al* looked at several medical conditions, examining the 10,000 most disorder associated SNPs for enrichment with SNPs with a high eQTL score (183). They found enrichment when considering autoimmune related conditions, which may be due to their use of lymphoblastoid cell line expression data to derive eQTLs.

However, neither of these studies took as large a selection of SNPs as my study, which combined eQTL data with SNPs from anywhere in the genome associated with schizophrenia at a lax p-value threshold of $p<0.5$. Hence, my study provides evidence that expression data can be combined usefully with GWAS data, whether using SNPs that are reliably associated with affected status or SNPs that only have the most marginal evidence for association.

5.4 Functional analysis of expression clusters using enrichment analysis

The clusters produced by using the memISA and k-means clustering methods described in Chapter 2 on the Dobrin prefrontal cortex brain expression dataset were examined for enrichment for genes containing SNPs associated with schizophrenia according to a UK GWAS study (116, 184). A 3093-gene coexpression cluster, found using memISA, was found to be significantly enriched for these genes.

Two clusters produced by using memISA on the McLean 66 (MC66) prefrontal cortex expression dataset were identified as overlapping with the Dobrin 3093-gene coexpression cluster, one containing 2546 genes and the other containing 436 genes. These clusters were significantly enriched for schizophrenia associated genes, and also for genes differentially expressed in schizophrenia according to the Stanley database.

The Dobrin 3093-gene cluster and the two MC66 clusters were also examined for enrichment with genes differentially expressed in, or containing SNPs associated with, bipolar disorder. All three clusters were significantly enriched for both associated and differentially expressed genes, except for the MC66 436-gene cluster which was only significantly enriched for associated genes.

K-means clustering was used to subdivide the Dobrin 3093-gene and MC66 2546-gene cluster into three subclusters each. Each of these subclusters was also divided into three using k-means clustering, creating a second layer of subclusters. The subclusters were also tested for enrichment for genes upregulated in brain cell types. Additionally, Dobrin subclusters which shared a high proportion of their genes with an MC66 subcluster were identified.

The subclusters were examined for enrichment with genes associated with or differentially expressed in schizophrenia or bipolar disorder. Several particularly heavily enriched subclusters were identified. One of the most enriched was MC66 subcluster 1.3, which shared a high proportion of genes with Dobrin subcluster 2.2. Both of these subclusters were significantly enriched for schizophrenia associated genes, bipolar disorder associated genes, and genes differentially expressed in bipolar disorder.

However, this analysis did not take differences in gene length into account when determining enrichment for disorder associated genes. As long genes will on average contain more independent SNPs than short genes, there will be a greater likelihood of a long gene containing a particularly significantly associated SNP. This will bias lists of disorder associated genes toward including long genes. The analysis was therefore repeated using an external set of disorder associated genes which had been corrected for this bias.

The analysis using the length-adjusted disorder associated genes found that most of the clusters and subclusters were no longer significantly enriched for schizophrenia or bipolar disorder associated genes. There were some exceptions – in particular, MC66 subcluster 1.3 remained significantly enriched for bipolar disorder associated genes. However, the Dobrin subcluster 2.2, which shares a high proportion of genes with MC66 subcluster 1.3, was no longer significantly enriched for these genes.

MetaCore was used to examine the clusters and subclusters for enrichment with several different functional categories. The MC66 1.3 and Dobrin 2.2 subclusters were particularly enriched for genes relating to GABA neurotransmission and synaptic vesicles. This may suggest a link between these functional categories and the aetiology of bipolar disorder.

Overall, these results show that enrichment analysis can be a powerful tool for gaining insight into the possible functions of coexpression clusters. However, the analyses also demonstrate the effect a serious bias can have, with far fewer significant results after gene length was taken into account.

This work could be expanded in a number of directions. Firstly, it would be useful to return to the cluster sets produced by memISA and k-means clustering, and examine them for enrichment for the sets of genes associated with schizophrenia or bipolar disorder after correcting for gene length. If a strongly enriched cluster was found, it could also be further analysed using MetaCore.

Correlations between the clusters and subclusters could also be calculated, to determine relationships between them within a cluster set. Additionally, correlations could function as an alternative to the proportion of genes shared when finding related clusters and subclusters between cluster sets derived from different expression datasets. The best method of calculating consensus expression values across the genes of a whole cluster would need to be determined, although simply taking the mean expression of all genes in a cluster could be used as a starting point.

MetaCore has a wide range of capabilities not touched upon in this study. In particular, using it to identify genes which multiple strands of evidence suggest are related to schizophrenia or bipolar disorder and which belong to functional categories of relevance to brain function would be useful. Direct interactions between these genes and other well-supported putative disorder related genes would also be interesting. Such genes are likely to be worthy of further investigation, using either bioinformatics or laboratory based techniques.

Similar approaches integrating information from a range of data types have proved useful in the study of other disorders, such as atherosclerosis or cancer (185). The results that these systems biology techniques have produced include specific insights into the relationship between a particular gene and disease (186).

They also include broader findings that connect atherosclerosis to functional categories, and have helped to emphasise the active role of immunity in the disease (187).

The study of these two diseases is different to the situation in the neuropsychiatric disorders examined here. They do not suffer the problems of tissue availability and diagnostic uncertainty to the same degree as schizophrenia or bipolar disorder. Furthermore, even before the advent of large-scale expression and GWAS datasets, the aetiology of both cancer and atherosclerosis was well understood, giving systems biology methods a reliable framework to add to. Despite these factors, integrative approaches offer a powerful way to increase our understanding of mental disorders.

5.5 Further work

The work in this study could be expanded in a number of directions. The comparison of clustering methods in Chapter 2, for example, could be expanded by increasing the number and type of methods studied. The eQTL and polygenic score analysis in Chapter 3 would benefit from replication using alternative GWAS datasets, or alternative datasets containing both expression and association data to derive eQTLs.

A potential use of eQTL data is to demonstrate a connection between a GWAS locus associated with a disorder and the expression values of a gene in the vicinity of the SNP. This can sometimes show that the mode of action of a disorder-associated SNP is not always through the closest gene to it. For example, Schadt *et al* found that eQTL and expression evidence suggested that RPS26, not ERBB3 as previously thought, was responsible for a novel type I diabetes association signal on chromosome 12q3 (122).

However, these relationships are not always simple to dissect. When Cookson *et al* examined an obesity associated missense SNP in SH2B1, they found it also affected the expression values of the nearby genes EIF3C and TUFM (109). It is uncertain whether this SNP is in linkage disequilibrium with another SNP that directly affects EIF3C and TUFM expression, or whether SH2B1 protein has a regulatory role in EIF3C and TUFM expression that the missense mutation affects.

It might be useful to extend the eQTL and polygenic score analysis work in Chapter 3 by attempting to use expression data alone to predict affected status. A training dataset could be used to establish mean expression levels for cases and controls, and find the significance of any difference in expression between them. For each sample in the target dataset, the number of transcripts with expression levels closer to the mean expression for cases than controls could be assessed, and affected status regressed upon this.

The clustering methods described in Chapter 2 could also be used to group similar schizophrenia or bipolar disorder cases together, effectively defining expression-based subtypes of these disorders. Defining such subtypes could help to better link these diagnostic categories to the underlying biology, in a similar way to which expression data is used to define subtypes of some types of cancer (188). The subtypes would be particularly useful if they could be shown to also have relevance to external symptoms of the disorder (for example, an expression subtype whose members all suffer from paranoid delusions) or to the genotypes of the samples (e.g. an expression subtype whose members all share a risk allele in a specific gene).

The second two extensions share a common weakness. As they both use expression data from schizophrenia or bipolar disorder cases, they both may be affected by changes in expression caused by the treatment of neuropsychiatric disorders, rather than changes in expression caused by the disorders themselves. Although this may be corrected for by using linear regression with covariates that represent the extent of treatment in each sample, such as lifetime dosage of antipsychotic drugs in the case of schizophrenia, it is possible that the effects of such treatments are too complex to be simply accounted for.

5.6 Conclusions

Overall, this study demonstrates a variety of ways in which expression data can be relevant to research into neuropsychiatric disorders. In Chapter 2, the clustering methods detailed provide structure to the data, moving from the mass of expression data to more specific clusters of coexpressed genes. These clusters were further subdivided in Chapter 4, and enrichment analysis used to suggest links between the subclusters and schizophrenia or bipolar disorder. Enrichment analysis was also used to annotate the clusters and subclusters with functional categories. Clusters

and subclusters that are enriched both for disorder-related genes and genes from a functional category suggest a possible link between that disorder and that functional category.

In contrast to this increasing specificity, in Chapter 3, entire brain expression datasets were used to categorise SNPs into sets with greater and lesser effect on global expression. These ability of these SNP sets to predict schizophrenia affected status through polygenic score analysis was then examined, and it was shown that the SNPs which affected expression to a greater extent also predicted schizophrenia affected status significantly better. This is a useful and novel finding which demonstrates that, despite the probable neurodevelopmental origin of schizophrenia, gene expression data from adult human brain can have relevance to the study of neuropsychiatric disease.

References

1. Eddy SR. Non-coding RNA genes and the modern RNA world. Nature Reviews Genetics2001 Dec;2(12):919-29.
2. Guo YF, Xiao P, Lei SF, Deng FY, Xiao GG, Liu YZ, et al. How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. Acta Biochimica Et Biophysica Sinica2008 May;40(5):426-36.
3. Wheelan SJ, Martinez Murillo F, Boeke JD. The incredible shrinking world of DNA microarrays. Mol Biosyst2008 Jul;4(7):726-32.
4. Southern EM. Detection of Specific Sequences among DNA Fragments Separated by Gel-Electrophoresis. Journal of Molecular Biology1975;98(3):503-&.
5. Burnette WN. Western Blotting - Electrophoretic Transfer of Proteins from Sodium Dodecyl Sulfate-Polyacrylamide Gels to Unmodified Nitrocellulose and Radiographic Detection with Antibody and Radioiodinated Protein-A. Analytical Biochemistry1981;112(2):195-203.
6. Alwine JC, Kemp DJ, Stark GR. Method for Detection of Specific Rnas in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes. Proceedings of the National Academy of Sciences of the United States of America1977;74(12):5350-4.
7. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-Directed, Spatially Addressable Parallel Chemical Synthesis. Science1991 Feb 15;251(4995):767-73.
8. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proceedings of the National Academy of Sciences of the United States of America1996 Oct 1;93(20):10614-9.
9. Feng YJ, Yang JH, Huang HD, Kennedy SP, Turi TG, Thompson JF, et al. Transcriptional profile of mechanically induced genes in human vascular smooth muscle cells. Circulation Research1999 Dec 3;85(12):1118-23.

10. Rivas LA, Garcia-Villadangos M, Moreno-Paz M, Cruz-Gil P, Gomez-Elvira J, Parro V. A 200-antibody microarray biochip for environmental monitoring: searching for universal microbial biomarkers through immunoprofiling. Anal Chem2008 Nov 1;80(21):7970-9.
11. Li W, Ruan K. MicroRNA detection by microarray. Anal Bioanal Chem2009 Jun;394(4):1117-24.
12. Trevino V, Falciani F, Barrera-Saldana HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Molecular Medicine2007 Sep-Oct;13(9-10):527-41.
13. Ma C, Lyons-Weiler M, Liang W, LaFramboise W, Gilbertson JR, Becich MJ, et al. In vitro transcription amplification and labeling methods contribute to the variability of gene expression profiling with DNA microarrays. J Mol Diagn2006 May;8(2):183-92.
14. Mehlmann M, Townsend MB, Stears RL, Kuchta RD, Rowlen KL. Optimization of fragmentation conditions for microarray analysis of viral RNA. Analytical Biochemistry2005 Dec 15;347(2):316-23.
15. Suarez-Farinas M, Haider A, Wittkowski KM. "Harshlighting" small blemishes on microarrays. BMC Bioinformatics2005 Mar 22;6:-.
16. Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association2004 Dec;99(468):909-17.
17. Dai MH, Wang PL, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Research2005;33(20):-.
18. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics2003 Jan 22;19(2):185-93.
19. Tukey JW. Exploratory Data Analysis. Reading, Mass.: Addison-Wesley; 1977.
20. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. Bioinformatics2002 Dec;18(12):1585-92.
21. Robinson MD, Speed TP. A comparison of Affymetrix gene expression arrays. BMC Bioinformatics2007 Nov 15;8:-.
22. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG, et al. A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res2009 Jan;38(3):e17.
23. Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao CF, et al. Decoding randomly ordered DNA arrays. Genome Research2004 May;14(5):870-7.
24. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics2003 Apr;4(2):249-64.
25. Schadt EE, Li C, Su C, Wong WH. Analyzing high-density oligonucleotide gene expression array data. J Cell Biochem2000 Oct 20;80(2):192-202.
26. Moore JH, Asselbergs FW, Williams SM. Bioinformatics challenges for genome-wide association studies. Bioinformatics2010 Feb 15;26(4):445-55.
27. Kennedy GC, Matsuzaki H, Dong SL, Liu WM, Huang J, Liu GY, et al. Large-scale genotyping of complex DNA. Nature Biotechnology2003 Oct;21(10):1233-7.
28. Shen R, Fan JB, Campbell D, Chang WH, Chen J, Doucet D, et al. High-throughput SNP genotyping on universal bead arrays. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis2005 Jun 3;573(1-2):70-82.
29. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature2009 Aug 6;460(7256):748-52.

30. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. Stat Sci2009 Nov 1;24(4):561-73.
31. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science2007 Jun 1;316(5829):1341-5.
32. Katta S, Kaur I, Chakrabarti S. The molecular genetic basis of age-related macular degeneration: an overview. J Genet2009 Dec;88(4):425-49.
33. Riley B, Thiselton D, Maher BS, Bigdeli T, Wormley B, McMichael GO, et al. Replication of association between schizophrenia and ZNF804A in the Irish Case-Control Study of Schizophrenia sample. Molecular Psychiatry2010 Jan;15(1):29-37.
34. Moskvina V, Craddock N, Holmans P, Nikolov I, Pahwa JS, Green E, et al. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. Molecular Psychiatry2009 Mar;14(3):252-60.
35. Soronen P, Ollila HM, Antila M, Silander K, Palo OM, Kieseppa T, et al. Replication of GWAS of bipolar disorder: association of SNPs near CDH7 with bipolar disorder and visual processing. Molecular Psychiatry2010 Jan;15(1):4-6.
36. O'Donovan MC, Craddock NJ, Owen MJ. Genetics of psychosis; insights from views across the genome. Human Genetics2009 Jul;126(1):3-12.
37. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet2010 Jul;42(7):565-9.
38. Fan X, Shi L, Fang H, Cheng Y, Perkins R, Tong W. DNA microarrays are predictive of cancer prognosis: a re-evaluation. Clin Cancer Res2010 Jan 15;16(2):629-36.
39. Jonsson G, Staaf J, Vallon-Christersson J, Ringner M, Holm K, Hegardt C, et al. Genomic subtypes of breast cancer identified by array comparative genomic hybridization display distinct molecular and clinical characteristics. Breast Cancer Res2010 Jun 24;12(3):R42.
40. Andersson A, Eden P, Olofsson T, Fioretos T. Gene expression signatures in childhood acute leukemias are largely unique and distinct from those of normal tissues and other malignancies. BMC Med Genomics2010;3:6.
41. Bahn S, Augood SJ, Ryan M, Standaert DG, Starkey M, Emson PC. Gene expression profiling in the post-mortem human brain--no cause for dismay. J Chem Neuroanat2001 Jul;22(1-2):79-94.
42. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, et al. A survey of genetic human cortical gene expression. Nat Genet2007 Dec;39(12):1494-9.
43. Waring SC, Rosenberg RN. Genome-wide association studies in Alzheimer disease. Arch Neurol2008 Mar;65(3):329-34.
44. Akiskal HS, Bourgeois ML, Angst J, Post R, Moller HJ, Hirschfeld R. Re-evaluating the prevalence of and diagnostic composition within the broad clinical spectrum of bipolar disorders. Journal of Affective Disorders2000 Sep;59:S5-S30.
45. Edvardsen J, Torgersen S, Roysamb E, Lygren S, Skre I, Onstad S, et al. Heritability of bipolar spectrum disorders. Unity or heterogeneity? Journal of Affective Disorders2008 Mar;106(3):229-40.
46. van Os J, Kapur S. Schizophrenia. Lancet2009 Aug 22;374(9690):635-45.
47. Cardno AG, Gottesman, II. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. Am J Med Genet2000 Spring;97(1):12-7.
48. Woo TUW, Crowell AL. Targeting synapses and myelin in the prevention of schizophrenia. Schizophrenia Research2005 Mar 1;73(2-3):193-207.
49. Fatemi SH, Folsom TD. The neurodevelopmental hypothesis of schizophrenia, revisited. Schizophr Bull2009 May;35(3):528-48.
50. Owen MJ, O'Donovan MC, Harrison PJ. Schizophrenia: a genetic disorder of the synapse? BMJ2005 Jan 22;330(7484):158-9.
51. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, et al. Common variants conferring risk of schizophrenia. Nature2009 Aug 6;460(7256):744-7.

52. McMahon FJ, Akula N, Schulze TG, Muglia P, Tozzi F, Detera-Wadleigh SD, et al. Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. Nat Genet2010 Feb;42(2):128-31.
53. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet2008 Sep;40(9):1056-8.
54. Williams HJ, Norton N, Dwyer S, Moskvina V, Nikolov I, Carroll L, et al. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. Mol Psychiatry2010 Apr 6.
55. Williams HJ, Owen MJ, O'Donovan MC. Schizophrenia genetics: new insights from new approaches. Br Med Bull2009;91:61-74.
56. Le-Niculescu H, Balaraman Y, Patel S, Tan J, Sidhu K, Jerome RE, et al. Towards understanding the schizophrenia code: an expanded convergent functional genomics approach. Am J Med Genet B Neuropsychiatr Genet2007 Mar 5;144B(2):129-58.
57. Dettling M, Gabrielson E, Parmigiani G. Searching for differentially expressed gene combinations. Genome Biology2005;6(10):R88.
58. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics2005 September 14;6(227).
59. Allison DB, Cui X, Page GP, Sabripour M. Microarray Data Analysis: from disarray to consolidation to consensus. Nature Reviews Genetics2006;7:55-65.
60. Prelić A, Bleuler S, Zimmerman P, Wille A, Bühlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics2006;22(9):1122-9.
61. Riva A, Carpentier A-S, Torrésani B, Hénaut A. Comments on selected fundamental aspects of microarray analysis. Computational Biology and Chemistry2005(29):319-36.
62. Fang Z, Liu L, Yang J, Luo Q-M, Li Y-X. Comparisons of Graph-structure Clustering Methods for Gene Expression Data. Acta Biochimica et Biophysica Sinica2006;38(6):379-84.
63. Stansberg C, Vik-Mo AO, Holdhus R, Breilid H, Srebro B, Petersen K, et al. Gene expression profiles in rat brain disclose CNS signature genes and regional patterns of functional specialisation. BMC Genomics2007;8(94).
64. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, et al. Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America2002;99(7).
65. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics2006;22(19):2405-12.
66. de Hoon MJL, Imoto S, Nolan J, Miyano S. Open Source Clustering Software. Bioinformatics2004;20(9):1453-4.
67. Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics2006;22(16):1988-97.
68. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review E, Statistical, nonlinear and soft matter physics2003;67(3 pt 1).
69. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. Nature Genetics2002;31:370-7.
70. Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB. Reproducible Clusters from Microarray Research: Whither? BMC Bioinformatics2005(6 (Suppl 2)):S10.
71. Kloster M, Tang C, Wingreen NS. Finding regulatory modules through large-scale gene expression analysis. Bioinformatics2005;21(7):1172-9.
72. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A1998 Dec 8;95(25):14863-8.

73. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis: John Wiley and Sons, Inc., New York; 1990.
74. Fraley C, Raftery AE. MCLUST: Software for model-based cluster analysis. Journal of Classification1999;16(2):297-306.
75. Medvedovic M, Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics2002 Sep;18(9):1194-206.
76. Wall ME, Dyck PA, Brettin TS. SVDMAN--singular value decomposition analysis of microarray data. Bioinformatics2001 Jun;17(6):566-8.
77. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol2000;1(2):RESEARCH0003.
78. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics2007;8:22.
79. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, et al. EXPANDER--an integrative program suite for microarray data analysis. BMC Bioinformatics2005;6:232.
80. Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics2006;7:280.
81. Higgs BW, Elashoff M, Richman S, Barci B. An online database for brain disease research. BMC Genomics2006;7(70).
82. Benes FM, Walsh J, Ennulat DJ. National Brain Databank: Brain Tissue Gene Expression Repository. Available from: http://national_databank.mclean.harvard.edu/brainbank/Main.
83. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudney D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update Nucleic Acids Research2006;35(Database issue):D760-D5.
84. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research2002 Jan 1;30(1):207-10.
85. Development Core Team R. R: A language and environment for statistical computing. 2006; Available from: http://www.R-project.org.
86. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of Gene Chip data. Nucleic Acids Research2005;33(20):e175.
87. Bolstad BM, Irizarray RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics2003;19(2):185-93.
88. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics2005;21(18):3587-95.
89. Beißbarth T, Speed TP. GOstat: Find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics2004;20(9):1464-5.
90. Lester A. WWW-Mechanize. [Perl module] 2007; 1.34:[Available from: http://search.cpan.org/dist/WWW-Mechanize/.
91. Hartigan JA, Wong MA. A K-Means Clustering Algorithm. Applied Statistics1979;28(1):100-8.
92. Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. Bioinformatics2003;19(8):973-80.
93. Tseng GC. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. Bioinformatics2007;23:2247-55.
94. Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Stevens MHH. vegan: Community Ecology Package, R Package. 2008; Available from: http://vegan.r-forge.r-project.org/.
95. Thain D, Tannenbaum T, Livny M. Distributed computing in practice: the Condor experience. Concurrency and Computation: Practice and Experience2004;17((2-4)):323-56.

96. Tseng GC, Wong WH. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. Biometrics2005 Mar;61(1):10-6.
97. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet Jan 8;86(1):6-22.
98. Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. Science2001 Feb 16;291(5507):1224-9.
99. Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. Cis-acting variation in the expression of a high proportion of genes in human brain. Hum Genet2003 Jul;113(2):149-53.
100. Lander ES. The new genomics: global views of biology. Science1996 Oct 25;274(5287):536-9.
101. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature2009 Oct 8;461(7265):747-53.
102. Maher B. Personal genomes: The case of the missing heritability. Nature2008 Nov 6;456(7218):18-21.
103. O'Donovan MC, Craddock NJ, Owen MJ. Genetics of psychosis; insights from views across the genome. Hum Genet2009 Jun 12.
104. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature2009 Aug 6;460(7256):753-7.
105. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A2009 Jun 9;106(23):9362-7.
106. Peirce TR, Bray NJ, Williams NM, Norton N, Moskvina V, Preece A, et al. Convergent evidence for 2',3'-cyclic nucleotide 3'-phosphodiesterase as a possible susceptibility gene for schizophrenia. Arch Gen Psychiatry2006 Jan;63(1):18-24.
107. Bray NJ, Preece A, Williams NM, Moskvina V, Buckland PR, Owen MJ, et al. Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression. Hum Mol Genet2005 Jul 15;14(14):1947-54.
108. Law AJ, Lipska BK, Weickert CS, Hyde TM, Straub RE, Hashimoto R, et al. Neuregulin 1 transcripts are differentially expressed in schizophrenia and regulated by 5' SNPs associated with the disease. Proc Natl Acad Sci U S A2006 Apr 25;103(17):6747-52.
109. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet2009 Mar;10(3):184-94.
110. Le-Niculescu H, McFarland MJ, Mamidipalli S, Ogden CA, Kuczenski R, Kurian SM, et al. Convergent Functional Genomics of bipolar disorder: from animal model pharmacogenomics to human genetics and biomarkers. Neurosci Biobehav Rev2007;31(6):897-903.
111. Rapoport JL, Addington AM, Frangou S, Psych MR. The neurodevelopmental model of schizophrenia: update 2005. Mol Psychiatry2005 May;10(5):434-49.
112. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet2008 Aug;24(8):408-15.
113. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, et al. Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet2009 Apr;84(4):445-58.
114. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant quantitative trait Loci exist for DNA methylation and gene expression in human brain. PLoS Genet2010;6(5):e1000952.
115. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet2007 Oct;39(10):1217-24.
116. Higgs BW, Elashoff M, Richman S, Barci B. An online database for brain disease research. BMC Genomics2006;7:70.
117. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics1996;5(3):299-314.

118. Marangos PJ, Schmechel DE. Neuron specific enolase, a clinically useful marker for neurons and neuroendocrine cells. Annu Rev Neurosci1987;10:269-95.
119. Teepker M, Munk K, Mylius V, Haag A, Moller JC, Oertel WH, et al. Serum concentrations of s100b and NSE in migraine. Headache2009 Feb;49(2):245-52.
120. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet2007(81).
121. Purcell S. PLINK v1.05.
122. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol2008 May 6;6(5):e107.
123. Balding DJ. Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol2003 May;63(3):221-30.
124. Moskvina V, Smith M, Ivanov D, Blackwood D, St Clair D, Hultman C, et al. Genetic Differences between Five European Populations. Hum Hered2010;70(2):141-9.
125. Nagelkerke NJD. A Note on a General Definition of the Coefficient of Determination. Biometrika1991 Sep;78(3):691-2.
126. Bray NJ, Buckland PR, Williams NM, Williams HJ, Norton N, Owen MJ, et al. A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain. Am J Hum Genet2003 Jul;73(1):152-61.
127. Rollins B, Martin MV, Morgan L, Vawter MP. Analysis of whole genome biomarker expression in blood and brain. Am J Med Genet B Neuropsychiatr Genet2010 Jun 5;153B(4):919-36.
128. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet2010;6(4):e1000888.
129. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature2007 Oct 18;449(7164):851-61.
130. Southan C. Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics2004 Jun;4(6):1712-26.
131. Bebbington P, Ramana R. The epidemiology of bipolar affective disorder. Soc Psychiatry Psychiatr Epidemiol1995 Nov;30(6):279-92.
132. Kendrick T, Sibbald B, Burns T, Freeling P. Role of general practitioners in care of long term mentally ill patients. BMJ1991 Mar 2;302(6775):508-10.
133. McGuffin P, Perroud N, Uher R, Butler A, Aitchison KJ, Craig I, et al. The genetics of affective disorder and suicide. Eur Psychiatry2010 Jun;25(5):275-7.
134. Howes OD, Kapur S. The dopamine hypothesis of schizophrenia: version III--the final common pathway. Schizophr Bull2009 May;35(3):549-62.
135. Karoutzou G, Emrich HM, Dietrich DE. The myelin-pathogenesis puzzle in schizophrenia: a literature review. Mol Psychiatry2008 Mar;13(3):245-60.
136. Sodhi M, Wood KH, Meador-Woodruff J. Role of glutamate in schizophrenia: integrating excitatory avenues of research. Expert Rev Neurother2008 Sep;8(9):1389-406.
137. Olsen L, Hansen T, Jakobsen KD, Djurovic S, Melle I, Agartz I, et al. The estrogen hypothesis of schizophrenia implicates glucose metabolism: association study in three independent samples. BMC Med Genet2008;9:39.
138. Wassef A, Baker J, Kochan LD. GABA and schizophrenia: a review of basic science and clinical studies. J Clin Psychopharmacol2003 Dec;23(6):601-40.
139. Shastry BS. Bipolar disorder: an update. Neurochem Int2005 Mar;46(4):273-9.
140. Craddock N, O'Donovan MC, Owen MJ. Psychosis genetics: modeling the relationship between schizophrenia, bipolar disorder, and mixed (or "schizoaffective") psychoses. Schizophr Bull2009 May;35(3):482-90.
141. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res2009 Jan;37(1):1-13.

142. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A2005 Oct 25;102(43):15545-50.
143. Stansberg C, Vik-Mo AO, Holdhus R, Breilid H, Srebro B, Petersen K, et al. Gene expression profiles in rat brain disclose CNS signature genes and regional patterns of functional specialisation. BMC Genomics2007;8:94.
144. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, et al. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A2002 Apr 2;99(7):4465-70.
145. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet2000 May;25(1):25-9.
146. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res1999 Jan 1;27(1):29-34.
147. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res2003 Sep;13(9):2129-41.
148. Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. Nucleic Acids Res2006 Jul 1;34(Web Server issue):W645-50.
149. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological1995;57(1):289-300.
150. Beissbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics2004 Jun 12;20(9):1464-5.
151. Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biology2003;4(10):-.
152. MetaCore [database on the Internet]. GeneGO, Inc. Available from: http://www.genego.com.
153. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nat Genet2008 Sep;40(9):1053-5.
154. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics2008 Nov;92(5):265-72.
155. Winnenburg R, Wachter T, Plake C, Doms A, Schroeder M. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies. Briefings in Bioinformatics2008 Nov;9(6):466-78.
156. Lipscomb CE. Medical Subject Headings (MeSH). Bull Med Libr Assoc2000 Jul;88(3):265-6.
157. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol2005;4:Article17.
158. Harel A, Inger A, Stelzer G, Strichman-Almashanu L, Dalah I, Safran M, et al. GIFtS: annotation landscape analysis with GeneCards. BMC Bioinformatics2009;10:348.
159. Hosack DA, Dennis Jr. G, Sherman BT, Lane HC, Lempicki RA. Identifying Biological Themes within Lists of Genes with EASE Genome Biology2003;4(6):4.
160. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature2007 Jun 7;447(7145):661-78.
161. Liu CC, Lin CC, Chen WSE, Chen HY, Chang PC, Chen JJW, et al. CRSD: a comprehensive web server for composite regulatory signature discovery. Nucleic Acids Research2006;34:W571-W7.
162. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, et al. A transcriptome database for astrocytes, neurons and oligodendrocytes: a new resource for understanding brain development and function. Journal of Neuroscience2008;28(1):264-78.
163. Hong MG, Pawitan Y, Magnusson PK, Prince JA. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum Genet2009 Aug;126(2):289-301.

164. Kyosseva SV. Differential expression of mitogen-activated protein kinases and immediate early genes fos and jun in thalamus in schizophrenia. Progress in Neuro-Psychopharmacology & Biological Psychiatry2004;28:997-1006.
165. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. Science2008;320(539-543).
166. Benes FM. Relationship of GAD(67) regulation to cell cycle and DNA repair in GABA neurons in the adult hippocampus: bipolar disorder versus schizophrenia. Cell Cycle2010 Feb;9(4):625-7.
167. Cherlyn SY, Woon PS, Liu JJ, Ong WY, Tsai GC, Sim K. Genetic association studies of glutamate, GABA and related genes in schizophrenia and bipolar disorder: a decade of advance. Neurosci Biobehav Rev May;34(6):958-77.
168. Gutierrez B, Rosa A, Papiol S, Arrufat FJ, Catalan R, Salgado P, et al. Identification of two risk haplotypes for schizophrenia and bipolar disorder in the synaptic vesicle monoamine transporter gene (SVMT). Am J Med Genet B Neuropsychiatr Genet2007 Jun 5;144B(4):502-7.
169. O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, Morris D, et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol Psychiatry2010 Feb 16.
170. Mudge J, Miller NA, Khrebtukova I, Lindquist IE, May GD, Huntley JJ, et al. Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. PLoS One2008;3(11):e3625.
171. Charych EI, Liu F, Moss SJ, Brandon NJ. GABA(A) receptors and their associated proteins: implications in the etiology and treatment of schizophrenia and related disorders. Neuropharmacology2009 Oct-Nov;57(5-6):481-95.
172. Dickman DK, Davis GW. The schizophrenia susceptibility gene dysbindin controls synaptic homeostasis. Science2009 Nov 20;326(5956):1127-30.
173. Mead CL, Kuzyk MA, Moradian A, Wilson GM, Holt RA, Morin GB. Cytosolic Protein Interactions of the Schizophrenia Susceptibility Gene Dysbindin. J Neurochem2010 Mar 17.
174. Bradshaw NJ, Ogawa F, Antolin-Fontes B, Chubb JE, Carlyle BC, Christie S, et al. DISC1, PDE4B, and NDE1 at the centrosome and synapse. Biochemical and Biophysical Research Communications2008 Dec 26;377(4):1091-6.
175. DeRosse P, Lencz T, Burdick KE, Siris SG, Kane JM, Malhotra AK. The genetics of symptom-based phenotypes: toward a molecular classification of schizophrenia. Schizophr Bull2008 Nov;34(6):1047-53.
176. Trevino V, Falciani F, Barrera-Saldana HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol Med2007 Sep-Oct;13(9-10):527-41.
177. Iwamoto K, Kato T. Gene expression profiling in schizophrenia and related mental disorders. Neuroscientist2006 Aug;12(4):349-61.
178. Tan PK, Downey TJ, Spitznagel EL, Xu P, Fu D, Dimitrov DS, et al. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Research2003 Oct 1;31(19):5676-84.
179. Jolin EM, Weller RA, Weller EB. Psychosis in children with velocardiofacial syndrome (22q11.2 deletion syndrome). Curr Psychiatry Rep2009 Apr;11(2):99-105.
180. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics2005 Sep 14;6:-.
181. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E Stat Nonlin Soft Matter Phys2003 Mar;67(3 Pt 1):031902.
182. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics2007 Jan 24;8:-.

183. Nicolae DL, Gamazon E, Zhang W, Duan SW, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. Plos Genetics2010 Apr;6(4):-
.
184. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nature Genetics2008 Sep;40(9):1053-5.
185. Faratian D, Bown JL, Smith VA, Langdon SP, Harrison DJ. Cancer systems biology. Methods Mol Biol2010;662:245-63.
186. Diez D, Wheelock AM, Goto S, Haeggstrom JZ, Paulsson-Berne G, Hansson GK, et al. The use of network analyses for elucidating mechanisms in cardiovascular disease. Molecular Biosystems2010;6(2):289-304.
187. Ramsey SA, Gold ES, Aderem A. A systems biology approach to understanding atherosclerosis. EMBO Mol Med2010 Mar;2(3):79-89.
188. Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. J Clin Oncol2006 Nov 1;24(31):5079-90.

Appendix A – additional files

Appendix file 1 – ISAscripts.zip

ZIP archive containing the Perl and R scripts needed to run the version of ISA and memISA described here. Includes Instructions.txt, a step-by-step guide to using them, and the Perl scripts needed to make them work with CONDOR.

Appendix file 2 – AllOverlaps.xls

Spreadsheet showing inter-method overlap for clusters from all methods, in all datasets. Overlap is defined as the percentage of genes present in the smaller cluster that are also found in the larger cluster.

Appendix file 3 – SizeDistribution.xls

Spreadsheet showing number of genes present (cluster size) in each cluster for each method across all datasets. Also shows mean cluster size and standard deviation of cluster sizes.

Appendix file 4 – Subcluster_Functional_Enrichment.zip

ZIP archive containing text files with significantly enriched functional categories for both parent clusters and all 24 subclusters. Functional category types are: MetaCore maps, MetaCore networks, GO biological process categories, GO molecular function categories, GO localisation categories. Where more than 50 functional categories are enriched, the top 50 are presented. 'NS' indicates that this was the top hit, but it did not reach significance.

Appendix file 5 – Final_Thesis.doc

An electronic copy of this thesis.

Appendix file 6 – Clustering_Comparison_Paper.pdf

An electronic copy of the peer-reviewed paper based upon Chapter 2.

Appendix file 7 – eQTL_Polygenic_Score_Paper.docx

An electronic copy of the final draft of the paper based upon Chapter 3.

Appendix B Tables showing regression of affected status on risk allele score for Chapter 3 secondary analyses

Table S1: Regression of affected status on risk allele score, secondary analyses (*trans/cis* context)

Row	Expression dataset	Training dataset	Target dataset	Top / bottom eQTL percentage	Nagelkerke pseudo-R2	Regression p-value	Case/control risk allele score difference	SNP count
1	Myers *et al*	Split ISC	Split ISC	Top 50%	1.48	7.58E-07	2.97E-05	27061
2	Myers *et al*	Split ISC	Split ISC	Top 5%	0.46	6.24E-04	4.59E-05	4115
3	Myers *et al*	Split ISC	Split ISC	Bottom 50%	1.92	1.37E-09	3.70E-05	26275
4	Myers *et al*	Split ISC	Split ISC	Bottom 5%	0.40	7.74E-05	3.05E-05	4584
5	Myers *et al*	MGS	ISC	Top 50%	0.70	5.61E-13	9.06E-06	15414
6	Myers *et al*	MGS	ISC	Top 5%	0.13	1.03E-03	1.30E-05	2369
7	Myers *et al*	MGS	ISC	Bottom 50%	0.68	1.25E-12	7.62E-06	14929
8	Myers *et al*	MGS	ISC	Bottom 5%	0.30	1.63E-06	1.86E-05	2551

Table S2. Regression of affected status on risk allele score, secondary analyses using GeneVar expression dataset (*cis* context with 100kb *cis* window)

Row	Expression dataset	Gene subgroup	Training dataset	Target dataset	Top or bottom eQTL SNP list	Regression coefficient	Regression p-value	Nagelkerke pseudo-R^2	SNP count
1	GeneVar	All genes	MGS	ISC	Bottom 50%	165.2	3.55E-08	0.00402	10643
2	GeneVar	All genes	MGS	ISC	Top 5%	23.23	0.0353	0.000471	1273
3	GeneVar	All genes	MGS	ISC	Top 50%	150.5	2.56E-07	0.00350	10130
4	GeneVar	All genes	Split ISC	Split ISC	Bottom 50%	283.8	2.16E-12	0.0132	18954
5	GeneVar	All genes	Split ISC	Split ISC	Top 5%	45.01	0.00297	0.00215	2238
6	GeneVar	All genes	Split ISC	Split ISC	Top 50%	246.9	5.80E-10	0.0102	18022

Table S3. Regression of affected status on risk allele score, secondary analyses using Gibbs *et al* expression dataset (*cis* context with 100kb *cis* window)

Row	Expression dataset	Training dataset	Target dataset	Top / bottom eQTL percentage	Nagelkerke pseudo-R2	Regression p-value	Case/control risk allele score difference	SNP count
1	Gibbs *et al*	Split ISC	Split ISC	Top 50%	1.68	2.63E-15	4.77E-05	14448
2	Gibbs *et al*	Split ISC	Split ISC	Top 5%	0.40	8.56E-05	7.50E-05	1664
3	Gibbs *et al*	Split ISC	Split ISC	Bottom 50%	1.31	2.82E-12	3.85E-05	14900
4	Gibbs *et al*	Split ISC	Split ISC	Bottom 5%	0.32	0.000405	4.82E-05	2497
5	Gibbs *et al*	MGS	ISC	Top 50%	0.60	2.08E-11	1.64E-05	8363
6	Gibbs *et al*	MGS	ISC	Top 5%	0.31	1.16E-06	6.94E-05	989
7	Gibbs *et al*	MGS	ISC	Bottom 50%	0.02	0.081694	-1.00E-05	8205
8	Gibbs *et al*	MGS	ISC	Bottom 5%	-5.49E-05	0.438684	-1.43E-05	1465

Table S4. Regression of affected status on risk allele score, secondary analyses based upon all genes (*cis* results with variant *cis* windows and results based upon the full set of SNPs in the dataset of Myers *et al*)

Row	Expression dataset	Gene subgroup	Training dataset	Target dataset	*cis* window	Top or bottom eQTL SNP list	Regression coefficient	Regression p-value	Nagelkerke pseudo-R^2	SNP count
1	Myers *et al*	All genes	MGS	ISC	150kb	Bottom 50%	136.2	1.60E-08	0.00424	7001
2	Myers *et al*	All genes	MGS	ISC	150kb	Top 5%	5.348	0.528	-8.25E-05	819
3	Myers *et al*	All genes	MGS	ISC	150kb	Top 50%	86.73	0.000335	0.00163	7014
4	Myers *et al*	All genes	MGS	ISC	50kb	Bottom 50%	99.03	1.15E-06	0.00311	4901
5	Myers *et al*	All genes	MGS	ISC	50kb	Top 5%	10.79	0.134	0.000171	556
6	Myers *et al*	All genes	MGS	ISC	50kb	Top 50%	84.61	1.01E-13	0.00941	4170
7	Myers *et al*	All genes	Split ISC	Split ISC	150kb	Bottom 50%	173.7	2.64E-07	0.00699	12582
8	Myers *et al*	All genes	Split ISC	Split ISC	150kb	Top 5%	53.31	1.11E-05	0.00503	1495
9	Myers *et al*	All genes	Split ISC	Split ISC	150kb	Top 50%	255.7	1.01E-14	0.0161	12398
10	Myers *et al*	All genes	Split ISC	Split ISC	50kb	Bottom 50%	127.6	5.35E-06	0.00541	8865
11	Myers *et al*	All genes	Split ISC	Split ISC	50kb	Top 5%	42.59	2.97E-05	0.00451	1047
12	Myers *et al*	All genes	Split ISC	Split ISC	50kb	Top 50%	188.0	2.02E-11	0.012	8830

Table S5. Regression of affected status on risk allele score, secondary analyses with eQTLs based upon expression cluster genes (*cis* context with 100kb *cis* window)

Row	Expression dataset	Gene subgroup	Training dataset	Target dataset	Top or bottom eQTL SNP list	Regression coefficient	Regression p-value	Nagelkerke pseudo-R^2	SNP count
1	Myers *et al*	Dobrin 3093 cluster	MGS	ISC	Bottom 50%	8.494	0.476	-6.75E-05	1685
2	Myers *et al*	Dobrin 3093 cluster	MGS	ISC	Top 5%	-8.574	0.0390	0.000448	198
3	Myers *et al*	Dobrin 3093 cluster	MGS	ISC	Top 50%	9.390	0.448	-5.82E-05	1742
4	Myers *et al*	Dobrin 3093 cluster	Split ISC	Split ISC	Bottom 50%	74.11	2.57E-05	0.00459	3103
5	Myers *et al*	Dobrin 3093 cluster	Split ISC	Split ISC	Top 5%	11.00	0.0878	0.000526	358
6	Myers *et al*	Dobrin 3093 cluster	Split ISC	Split ISC	Top 50%	81.35	1.95E-06	0.00594	3124

Table S6. Regression of affected status on risk allele score, secondary analyses with eQTLs based upon genes differentially expressed in schizophrenia or bipolar disorder (*cis* context with *cis* window of 100kb)

Row	Expression dataset	Gene subgroup	Training dataset	Target dataset	Top or bottom eQTL SNP list	Regression coefficient	Regression p-value	Nagelkerke pseudo-R^2	SNP count
1	Myers *et al*	Schizophrenia differential expression	MGS	ISC	Bottom 50%	12.56	0.027	0.000534	335
2	Myers *et al*	Schizophrenia differential expression	MGS	ISC	Top 5%	0.2266	0.902	-0.000135	46
3	Myers *et al*	Schizophrenia differential expression	MGS	ISC	Top 50%	6.581	0.215	7.38E-05	335
4	Myers *et al*	Schizophrenia differential expression	Split ISC	Split ISC	Bottom 50%	8.136	0.286	3.79E-05	591
5	Myers *et al*	Schizophrenia differential expression	Split ISC	Split ISC	Top 5%	2.182	0.406	-8.48E-05	73
6	Myers *et al*	Schizophrenia differential expression	Split ISC	Split ISC	Top 50%	20.94	0.00535	0.00186	613
7	Myers *et al*	Bipolar disorder differential expression	MGS	ISC	Bottom 50%	14.06	0.044	0.00042	574
8	Myers *et al*	Bipolar disorder differential expression	MGS	ISC	Top 5%	-1.969	0.413	-4.51E-05	66
9	Myers *et al*	Bipolar disorder differential expression	MGS	ISC	Top 50%	9.334	0.16	0.000134	549
10	Myers *et al*	Bipolar disorder differential expression	Split ISC	Split ISC	Bottom 50%	16.06	0.103	0.000458	982
11	Myers *et al*	Bipolar disorder differential expression	Split ISC	Split ISC	Top 5%	5.81	0.082	0.000557	117
12	Myers *et al*	Bipolar disorder differential expression	Split ISC	Split ISC	Top 50%	13.99	0.148	0.0003	976

Appendix C Paper based upon chapter 2: A comparison of four clustering methods for brain expression microarray data

Published in BMC Bioinformatics 9:490 on November 25th 2008.

BMC Bioinformatics

BioMed Central

Research article

Open Access

A comparison of four clustering methods for brain expression microarray data

Alexander L Richards*, Peter Holmans, Michael C O'Donovan, Michael J Owen and Lesley Jones

Address: Department of Psychological Medicine, School of Medicine, University Hospital Wales, Heath Park, Cardiff, Wales, UK, CF14 4XN

Email: Alexander L Richards* - richardsal1@cardiff.ac.uk; Peter Holmans - holmanspa@cf.ac.uk; Michael C O'Donovan - odonovanmc@Cardiff.ac.uk; Michael J Owen - owenmj@cf.ac.uk; Lesley Jones - jonesl1@cf.ac.uk

* Corresponding author

Published: 25 November 2008

BMC Bioinformatics 2008, 9:490 doi:10.1186/1471-2105-9-490

Received: 20 August 2008

Accepted: 25 November 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/490

Abstract

Background: DNA microarrays, which determine the expression levels of tens of thousands of genes from a sample, are an important research tool. However, the volume of data they produce can be an obstacle to interpretation of the results. Clustering the genes on the basis of similarity of their expression profiles can simplify the data, and potentially provides an important source of biological inference, but these methods have not been tested systematically on datasets from complex human tissues. In this paper, four clustering methods, CRC, k-means, ISA and memISA, are used upon three brain expression datasets. The results are compared on speed, gene coverage and GO enrichment. The effects of combining the clusters produced by each method are also assessed.

Results: k-means outperforms the other methods, with 100% gene coverage and GO enrichments only slightly exceeded by memISA and ISA. Those two methods produce greater GO enrichments on the datasets used, but at the cost of much lower gene coverage, fewer clusters produced, and speed. The clusters they find are largely different to those produced by k-means. Combining clusters produced by k-means and memISA or ISA leads to increased GO enrichment and number of clusters produced (compared to k-means alone), without negatively impacting gene coverage. memISA can also find potentially disease-related clusters. In two independent dorsolateral prefrontal cortex datasets, it finds three overlapping clusters that are either enriched for genes associated with schizophrenia, genes differentially expressed in schizophrenia, or both. Two of these clusters are enriched for genes of the MAP kinase pathway, suggesting a possible role for this pathway in the aetiology of schizophrenia.

Conclusion: Considered alone, k-means clustering is the most effective of the four methods on typical microarray brain expression datasets. However, memISA and ISA can add extra high-quality clusters to the set produced by k-means, so combining these three methods is the method of choice.

Background

Clustering genes according to their expression profiles is an important step in interpreting data from microarray studies. Clustering can help summarise datasets, reducing tens of thousands of genes to a much smaller number of clusters. It can aid understanding of systemic effects; looking for a small change in expression between disease states across many genes in a cluster could be a better strategy for finding the causes of complex, polygenic disorders than looking for large changes in single genes[1]. Clustering can also help predict gene function, as coexpressed genes are more likely to have similar functions than non-coexpressed genes[2].

There are many clustering methods for microarray expression data currently available[3]. However, there are few comparisons of these methods, making it hard for researchers to make a rational choice between them. The majority of papers comparing multiple clustering methods use simulated data or data from simple organisms such as bacteria and yeast [4-6], which may limit the applicability of their findings to data from more complex sources such as human tissues which express more genes. Thus, to investigate human disease, it would be useful to test the methods upon expression data derived from complex human tissues, among which brain tissue is particularly complex since it expresses a higher proportion of the genome transcribed than other tissues[7,8]. Thalamuthu *et al* [9] have previously looked at a wide range of datasets, including some human expression datasets. However, since they restricted their analysis to functionally defined subsets of genes, that analysis did not fully reflect the complexity of human expression, particularly for disorders where there is insufficient knowledge of their aetiology to focus on specific subsets of genes.

We have examined four methods, k-means clustering[10], Chinese Restaurant Clustering (CRC)[11], the Iterative Signature Algorithm (ISA)[12,13] and a new, progressive variant of ISA called memISA. memISA was loosely based upon another method called PISA, for which there was no suitable implementation[14]. These were chosen after a literature survey of the available methods (see table in Additional Files 1). All four are unsupervised methods that derive the clusters from the input data, rather than supervised methods which classify genes into user-specified clusters.

Many of the available comparative clustering studies focus exclusively on older methods [5,15], or restrict their analysis to a single class of clustering methods [4,6]. In our study, the methods were chosen on the basis of variety. ISA and memISA are examples of biclustering methods, CRC is a mixture model based method, while k-means clustering is a simple, well-understood algorithm. They were reported as performing well by their authors and/or other studies [4,5].

The methods were also chosen partly on the basis of novelty. Apart from k-means clustering, they are too recent to have been included in many previous surveys of clustering methods, and so are particularly in need of testing.

We compared the performance of these three methods by examining the results for biologically meaningful clustering by looking for gene ontology (GO) enrichments within the resulting clusters. We also generated and compared a modified variation of ISA, memISA, which weighted against genes that were already members of a cluster to prevent bias of clusters detected from the strongest genes within them.

Methods

Datasets

Three datasets were used for testing, the Dobrin [16], McLean 66 [17] (MC66) and Perrone-Bizzozero (PB – GEO dataset GSE4036) [18] datasets (Table 1). They were downloaded in CEL format from the Stanley Medical Research Online Genomics database[16], the Harvard National Brain Databank database[17] and GEO[19], respectively. They were then processed using R[20], with custom CDF files to map the probes to genes[21]. Box plots were used to examine the quality of the data, and several outlier samples were removed. Three versions of each dataset were produced. One was normalised by the RMA median polish method, for use in CRC and k-means[22]. The other two were normalised to produce a gene-normalised and sample-normalised dataset for running ISA[12].

Gene coverage

Gene coverage, the percentage of genes on the chip that are put into at least one cluster, was assessed for the cluster set produced by each method.

Speed

The methods were also assessed by speed. As ISA and memISA are dependent on parallelisation to run at a reasonable speed, this is taken as real-world time taken to run, rather than computer run-time used. For k-means and penalised k-means, this includes the time taken to estimate k.

GO enrichment

GO enrichment is a method that assesses the percentage of clusters that are significantly enriched (compared to all annotated genes on the microarray) with genes from one or more Gene Ontology categories (from the goa_human database) at different significance levels, using Fisher's exact test and the Benjamini false discovery rate multiple

Table 1: Datasets used to test clustering methods

	Pre quality control number of samples			Post quality control number of samples			Tissue	Chip	Number of genes
	Control	SCZ	BP	Control	SCZ	BP			
Dobrin	25	26	27	20	22	22	Brodmann Area 46	Affymetrix 133 plus 2.0	20292
McLean 66	27	18	19	27	15	19	Dorsolateral Prefrontal Cortex	Affymetrix 133A	12757
Perrone-Bizzozero cerebellum	14	14	0	14	14	0	Cerebellum	Affymetrix 133 plus 2.0	20292

Quality control consisted of box plotting the samples and removing outliers.

testing correction[23]. Clusters were tested for enrichment (using Fisher's exact test) for all GO biological process terms 3 or more levels deep into the hierarchical tree of GO terms, at several different levels of significance. At least 3 genes from the input cluster had to match a GO category for the cluster to be counted as enriched for that category, to ensure that chance appearance of 1 or 2 genes from a GO category with few members could not affect the results. The percentage of clusters matching this criterion gives a measure of the biological, functional relevance of the clusters.

GO enrichment was determined with the web-based service, GOstat[24]. This accepts multiple kinds of gene name or ID as input, allowing approximately 85% of genes within the input clusters to be included. This was automated using WWW-Mechanize, a Perl module[25].

To compare the results of GO enrichments for the various clustering algorithms, we also examined several random cluster sets using GO enrichment. Four sets of clusters with the same distribution of cluster sizes as those made by k-means (at the value of k recommended by cascadeKM), CRC, ISA and memISA (both after removal of overlapping clusters) were produced. The cluster sets made from the Dobrin, MC66 and PB datasets were combined when determining the distribution of sizes. The new cluster sets had genes chosen at random from all those available on the Affymetrix 133P chip.

k-means

k-means clustering is a standard clustering method that has been in use for several decades [26]. It requires that the user specify the number of clusters to sort the genes into (k). k-means clustering is a single cluster membership method - each gene can belong to only one cluster and it also assigns every gene to a cluster. Essentially, it distributes k centroids (quasi-data points representing cluster centres) throughout the data. Data points are then assigned to their nearest cluster, and the centroids are moved to minimise the distance between them and their assigned data points. This is repeated until the centroids stop moving. A number of distance measures can be used to define distance between data point and centroid, with Euclidean distance being one of the most commonly used and simplest. The procedure is summarised in Fig. 1.

There are numerous variants of k-means clustering [27,28]. Here, two are tried - standard k-means clustering, as above, and penalised k-means clustering. Penalised k-means clustering uses a threshold parameter (λ) to allow some of the genes to be treated as noise, and not clustered.

Initially, an estimate for the value of k was found for all three datasets using the cascadeKM function in R. Values of k between 2 and 35 were assessed, with 25 iterations per value, and the k values that minimised the Calinski criterion were chosen [29]. The recommended values of k were 6 for Dobrin, 7 for MC66 and 8 for PB cerebellum. k-means and penalised k-means were then performed on all four datasets at 200 iterations and these values of k. The recommended value of 0.1 was used for λ in penalised k-means.

These small values of k will only partition the data into several large clusters, which may be too general a grouping to provide biologically relevant inferences. To examine the performance of k-means when producing smaller, more specific clusters, and also for a more direct comparison to CRC, k-means and penalised k-means clustering were also performed with values of k equal to the numbers of clusters produced by CRC on that dataset (23 in all cases).

k-means was performed using Cluster 3.0 [10]. Penalised k-means clustering was performed using PWKmeans [28]. Both were performed on a Windows desktop PC with 2 GB RAM, using a 2.66 Ghz processor.

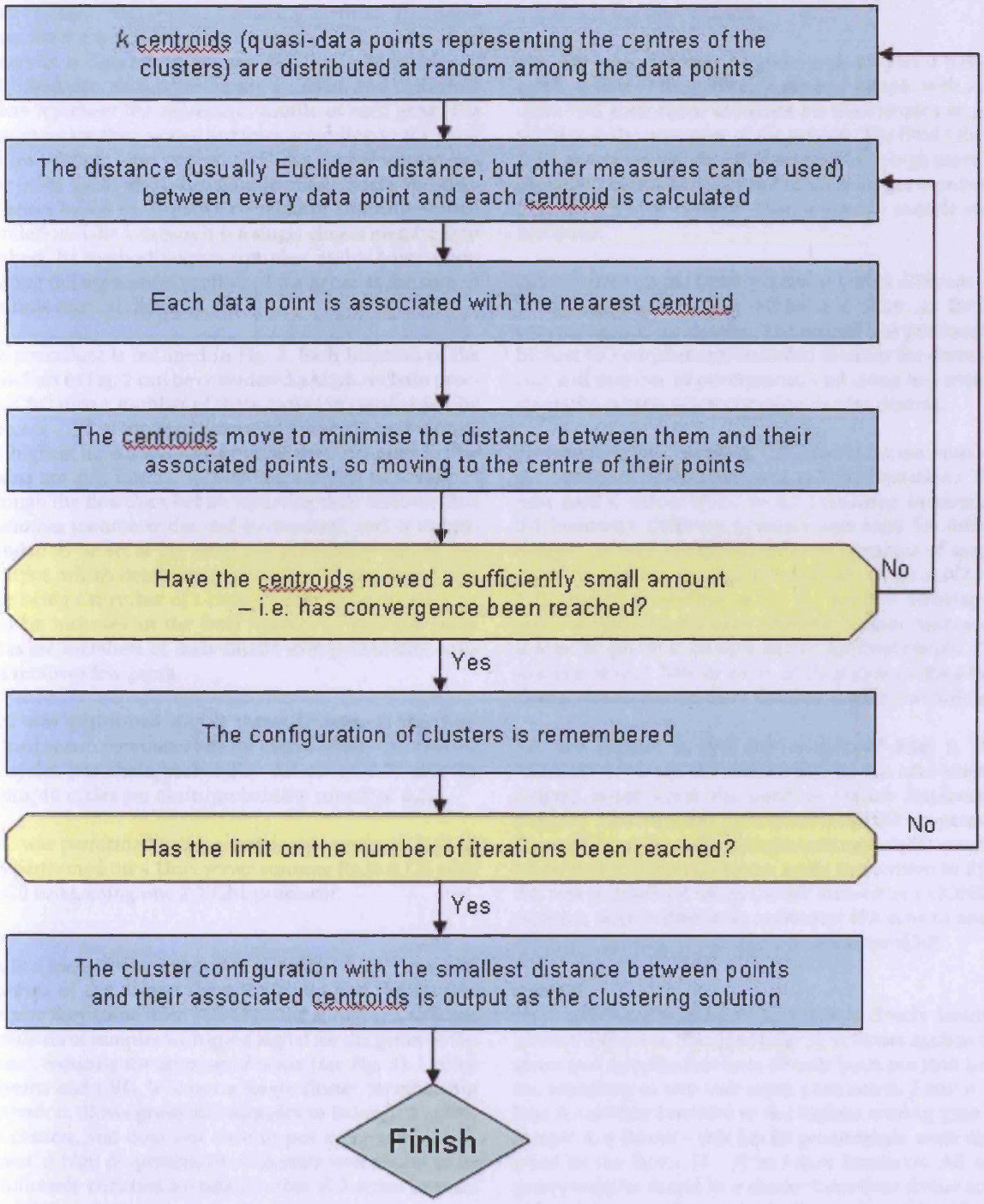


Figure 1
Flowchart summarising the method used by k-means clustering. k is a user-defined input parameter which sets the number of clusters k-means clustering will find.

CRC

CRC[11] is a model-based clustering method. The name arises from a metaphor where genes are regarded as customers in a Chinese restaurant with unlimited tables of unlimited size, each representing a cluster, and their food orders represent the expression profile of each gene. The customers are then seated at tables according to the similarities of their food orders. CRC has several advantages over other methods. It can handle missing data and cluster genes based on negative correlation and time-shifted correlation. Like k-means it is a single cluster membership method. Its methodology is complex, and is based upon treating the expression profiles of the genes as the sum of multiple normal distributions.

The procedure is outlined in Fig. 2. Each iteration of the flowchart in Fig. 2 can be considered a Markov chain process. CRC runs a number of these chains in parallel (set by the user – 10 is the recommended amount), and reports the highest likelihood cluster set as the final output. The chains are also limited to a certain number of iterations through the flowchart before reporting their clusters. This is another parameter decided by the user, and is recommended to be set at 20. Finally, a probability cut-off can be input, which determines how high the likelihood of a gene being a member of a cluster needs to be in order for it to be included in the final output. In practice, most genes are members of their cluster with probability 1, so this removes few genes.

CRC was performed on all three datasets. It was performed at two parameter sets for each dataset – 10 chains/20 cycles per chain/probability cut-off of 0.7, and 20 chains/40 cycles per chain/probability cut-off of 0.9.

CRC was performed using a standalone program [11]. It was performed on a Unix server running Redhat OS with 32 GB RAM, using one 2.2 Ghz processor.

ISA

ISA is a biclustering method – it clusters both rows and columns of the dataset, here the genes and the specific samples they come from [12,13]. This allows ISA to focus on subsets of samples with good signal for the genes of the cluster, reducing the amount of noise (see Fig. 3). Unlike k-means and CRC, it is not a single-cluster membership method: it allows genes and samples to belong to multiple clusters, and does not have to put every gene into a cluster. A high proportion of its clusters were found to be significantly enriched for one or more GO terms in yeast data[4].

ISA produces tens of thousands of clusters. In postprocessing, to reduce this set to a manageable size, duplicate clusters are removed, similar clusters are merged, and clusters can be reiterated through ISA. The nature of postprocessing affects the final clusters.

ISA also assigns 'scores' to genes and samples it has clustered, as part of its method. A gene or sample with a high score will have more influence on the samples or genes selected at the next stage of the process. The final values of these scores are reported in ISA's output. A high score here indicates that the gene or sample has had greater influence over the clusters' contents than a gene or sample with a low score.

ISA was used on the Dobrin datasets with 8 different postprocessing regimes (see Additional Files 2, ISAPostprocessing.doc, for details). The regime that produced the highest GO enrichments included filtering the clusters by size and number of occurrences, and using less stringent similarity criteria when combing similar clusters.

To compare with memISA, CRC and k-means, runs were performed on all three datasets, at 20000 iterations. These runs used t_C values of 1.0 to 4.2 (inclusive, increasing in 0.1 intervals). Different t_C values were used for different datasets, as each contained different numbers of samples – Dobrin was run at t_C 0.2, 0.5 and 1.0, MC66 at 0.25 and 1.25, and PB cerebellum at 0.1, 0.4 and 0.7. Filtering was used – a cluster had to have appeared 3 times, and contain at least 40 genes, to be included in the final output. Clusters that shared 70% or more of their genes with a larger cluster were removed from the final results (see below).

ISA was written in Perl (see Additional Files 3, ISAScripts.zip for a zip file containing all ISA and memISA scripts), based upon the previous Matlab implementation[13]. This implementation has all of the properties of the Matlab version. The postprocessing scripts were written in Perl. The normalisation script was written in R[20]. ISA was parallelised using Cardiff University's CONDOR network, which distributes individual ISA runs to unused Windows desktop computers across campus[30].

memISA

The underlying method of memISA is closely based on ISA and similar to PISA[14] (Fig. 3). It biases against both genes and samples that have already been put into a cluster, according to two user input parameters, f and n. The bias is calculated relative to the highest scoring gene and sample in a cluster – this has its gene/sample score multiplied by the factor $(1 - f)$ in future iterations. All other genes/samples found in a cluster have their future scores reduced by a smaller amount. This is determined by the proportion of their score and the highest gene/sample score – a gene with a quarter of the score of the highest gene will have its future scores multiplied by $1 - (f * 0.25)$. The intent of this is to bias against the highest scoring

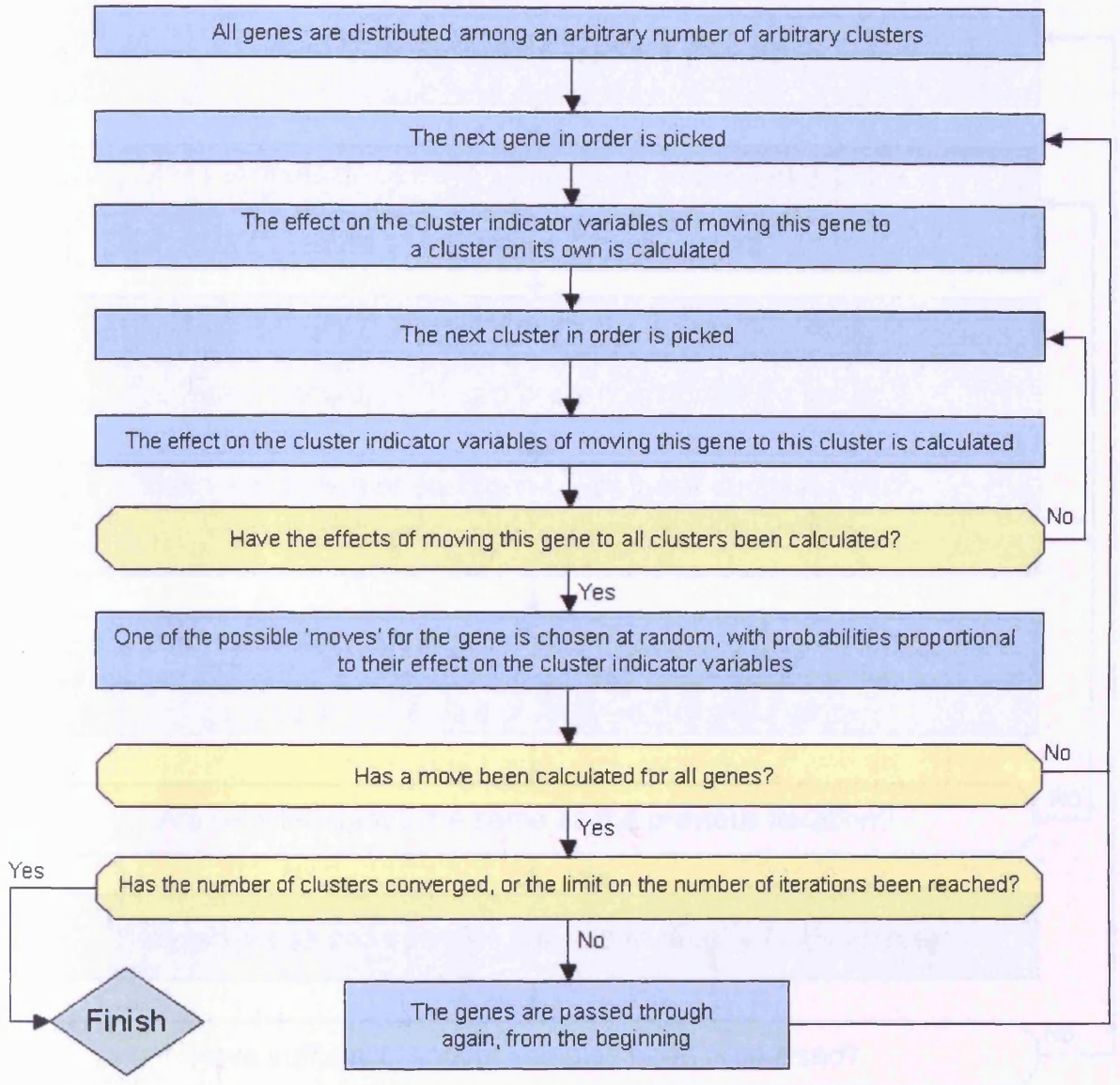


Figure 2
Flowchart summarising the method used by CRC. One run through this flowchart equates to a single chain in CRC, with several chains being run in parallel. The number of parallel chains and the maximum number of iterations are user-defined parameters.

genes of a cluster while allowing lower scoring genes to be relatively unaffected and still be included in subsequent clusters (the highest scoring genes typically have scores 10 times greater than the majority of genes in a cluster). If a gene/sample is included in a subsequent cluster, the biases are multiplied together – a gene which is the strongest gene in two successive clusters would have its score multiplied by $(1 - f)^2$ in following iterations.

These biases are only remembered for a certain number of iterations (n). Every n iterations, the slate is wiped clean. This is to ensure that memISA does not begin returning

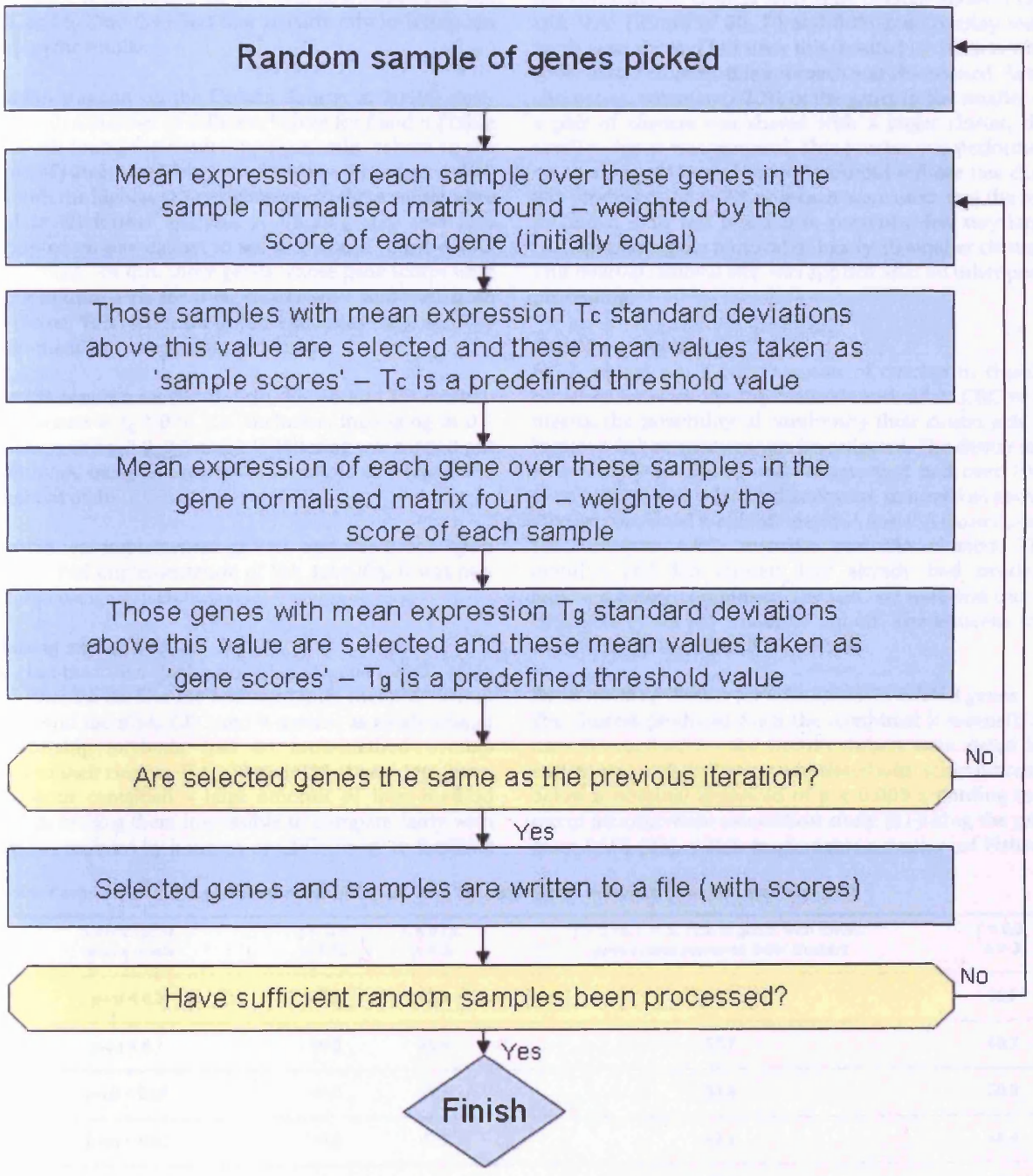


Figure 3
Flowchart summarising the method used by ISA. t_G and t_C are user-defined threshold parameters. They determine how great the level of expression for a gene or sample (defined in standard deviations from the weighted mean of all genes over those samples, or all samples over those genes) needs to be for selection in the cluster. Higher values lead to more, smaller clusters, lower values to fewer, larger clusters. A preliminary run at a low number of iterations, with a wide range of values for t_G and t_C, is used to determine a sensible range of t_G and t_C values for use in the main run.

noise once it has found all the available clusters in the data, and to limit the effect that an early misclustering can have on the results.

memISA was run on the Dobrin dataset at 20000 iterations with a number of different values for f and n (Table 2). It was found the results were generally robust to the values of f and n, and that $f = 0.7$ and $n = 5$ produced clusters with the highest GO enrichment, so these values were used in all further analysis. A filtering step was also attempted on one dataset to see if it would improve GO enrichment. For this, those genes whose gene scores were in the bottom 10% for their cluster were removed from the cluster. This step reduced both gene coverage and GO enrichment and so was not used further.

memISA was run on the Dobrin, MC66 and PB cerebellum datasets at t_G 1.0 to 4.2 (inclusive, increasing in 0.1 intervals) and t_G 0.2, 0.5 and 1.0. Filtering was carried out as with ISA, using an occurrence threshold of 3 and a size threshold of 40.

memISA was implemented in Perl, and was based upon the new Perl implementation of ISA. Like ISA, it was parallelised using CONDOR.

Assessing overlap between clusters

We examined inter-method overlap in gene membership of clusters for the four methods and intra-method overlap of ISA and memISA. CRC and k-means, as single-cluster membership methods, had no intra-method overlap between their clusters. ISA and memISA cluster sets, however, both contained a large amount of intra-method overlap, making them impossible to compare fairly with clusters produced by k-means or CRC. To try to facilitate fair comparison, clusters with gene overlap above a certain level (values of 60, 70 and 80% gene overlap were tried) were merged but since this resulted in datasets with fewer than 3 clusters, this approach was abandoned. As an alternative, where over 70% of the genes in the smaller of a pair of clusters was shared with a larger cluster, the smaller cluster was removed. This process was performed on a subset of ISA and memISA output – those raw clusters produced at $t_G = 2.1$ or greater were used, and the rest discarded. This was in order to prevent a few very large clusters causing the removal of nearly all smaller clusters. This overlap removal step was applied after all other post-processing.

Combining methods

As there was not a large amount of overlap in clusters obtained between the ISA methods and either CRC or k-means, the possibility of combining their cluster sets to improve GO enrichment was investigated. The cluster sets were simply combined and clusters that had over 70% gene overlap with a larger cluster were removed as above. One set contained k-means, memISA and ISA clusters, one set contained CRC, memISA and ISA clusters. The memISA and ISA clusters had already had overlaps removed before combining. The CRC set used was the 10 chains/20 cycles per chain/0.7 cut-off. The k-means sets used were the k = 23 and k = 22 sets.

Enrichment of clusters for schizophrenia related genes

The clusters produced from the combined k-means/ISA/memISA method on the Dobrin dataset were tested for enrichment with 607 genes associated with schizophrenia below a nominal threshold of $p < 0.005$ according to a recent genome-wide association study [31]using the program EASE [32], which implements a version of Fisher's

Table 2: Comparison of GO enrichments for different memISA parameters in Dobrin (overlaps not removed)

% enriched at varying p-vals	$f = 0.5$, $n = 10$	$f = 0.75$, $n = 5$	$f = 0.75$, $n = 5$, 10% of genes with lowest gene scores removed from clusters	$f = 0.5$, $n = 3$
p-val < 0.3	62.5	92.3	88.5	85.7
p-val < 0.1	50.0	65.4	57.7	60.7
p-val < 0.05	50.0	57.7	53.8	50.0
p-val < 0.01	43.8	42.3	42.3	46.4
p-val < 0.001	37.5	38.5	38.5	42.9
p-val < 0.0001	31.3	34.6	26.9	28.6
Gene coverage	61.1	78.8	74.7	74.7
Number of clusters found	16	26	26	28

Exact Test. Enriched clusters were also tested for enrichment for 352 genes differentially expressed between schizophrenics and controls in the analysis of the Stanley Medical Research Institute Online Genomics Database[16] at an uncorrected p-value of 0.02 or lower.

Clusters from combined k-means/ISA/memISA in the independent MC66 dataset that shared over 45% of their genes with any enriched cluster from the Dobrin dataset were then identified. Their enrichment for schizophrenia-associated genes and genes differentially expressed in schizophrenia was then determined with EASE. A permutation-based method of enrichment determination was also used. This allows the enrichment p-value for the MC66 clusters to be determined independently of the Dobrin cluster. 4000 pairs of clusters were constructed at random from the genes present on the Affymetrix 133A chip.

The random clusters were constructed in pairs, as follows. Firstly, the number of genes shared between the three clusters was calculated (see Fig. 4). These figures were then used to create randomised MC66 clusters with the same level of overlap with the Dobrin cluster and each other.

165 genes from the Dobrin 3093-gene cluster were selected at random, and placed in both the 2546-gene and 436-gene MC66 randomised clusters. From the remaining Dobrin cluster genes, 1068 and 24 genes were selected at random, the former placed in the 2546-gene randomised cluster, the latter placed in the 436-gene randomised cluster. Then, 102 genes from the genes on the chip not present in the Dobrin 3093-gene cluster were selected at random, and placed in both the 2546-gene and 436-gene randomised clusters. From the remaining genes on the chip not present in the Dobrin 3093-gene cluster, 983 and 90 genes were selected at random, the former placed in the 2546-gene randomised cluster, the latter in the 436-gene randomised cluster. This was repeated 4000 times to produce a population of 8000 random clusters. These clusters were then processed with EASE in the same way as the original cluster, allowing the original results to be compared to them.

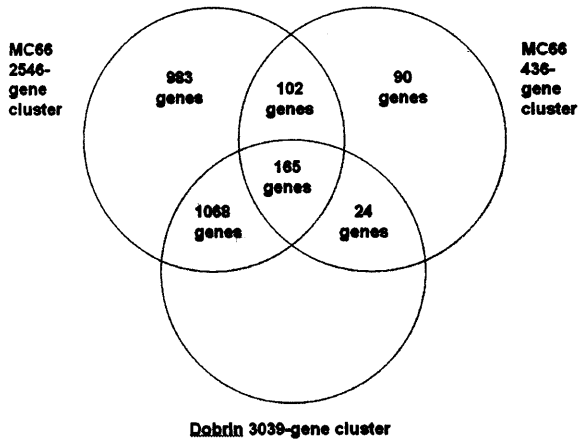


Figure 4
Overlap between putative schizophrenia-related clusters produced from Dobrin and MC66 datasets. Venn diagram showing the amount of overlap between the clusters enriched for schizophrenia-related genes, in order to construct randomised clusters for permutation.

These clusters were also examined for enrichment in KEGG and BioCarta pathways, using the Composite Regulatory Signature Database [33] (http://140.120.213.10:8080/crsd/main/home.jsp), and for enrichment in GO biological process categories using GOstat.

EASE was also used to test these clusters for enrichment with genes found to be ten-fold or more upregulated in specific cell types within brain tissue according to Cahoy *et al* [34]-specifically, neurons, oligodendrocytes and astrocytes.

Results and discussion

All four methods performed better than the random cluster sets when examined using GO enrichment to represent known biological relationships (Figs. 5, 6, 7). This implies that all the clustering methods result in groupings of biological significance. Of the three random cluster sets, those with the same size distribution as ISA had slightly lower GO enrichment than those with the same size distribution as memISA or CRC. This may suggest that GO enrichment has a small bias against ISA due to the sizes of clusters it produces. However, at $p < 0.05$ the difference dropped to under 1% GO enrichment, suggesting that any such bias is extremely slight and may well be due to chance.

k-means and penalised k-means

k-means and penalised k-means produced clusters with high GO enrichments, especially at the lower k values recommended by cascadeKM. In these low-k cluster sets, k-means obtained higher GO enrichments than penalised k-means. In the k = 22 and k = 23 cluster sets, they produced cluster sets with similar GO enrichment (Figs. 5, 6, 7). As k-means gave similar GO enrichment to penalised k-means and by definition clustered more genes it was used in comparisons with the other methods.

Effect of CRC parameters on GO enrichment

The different parameter sets used for CRC made little difference to the GO enrichments of its clusters. (Figs. 5, 6,

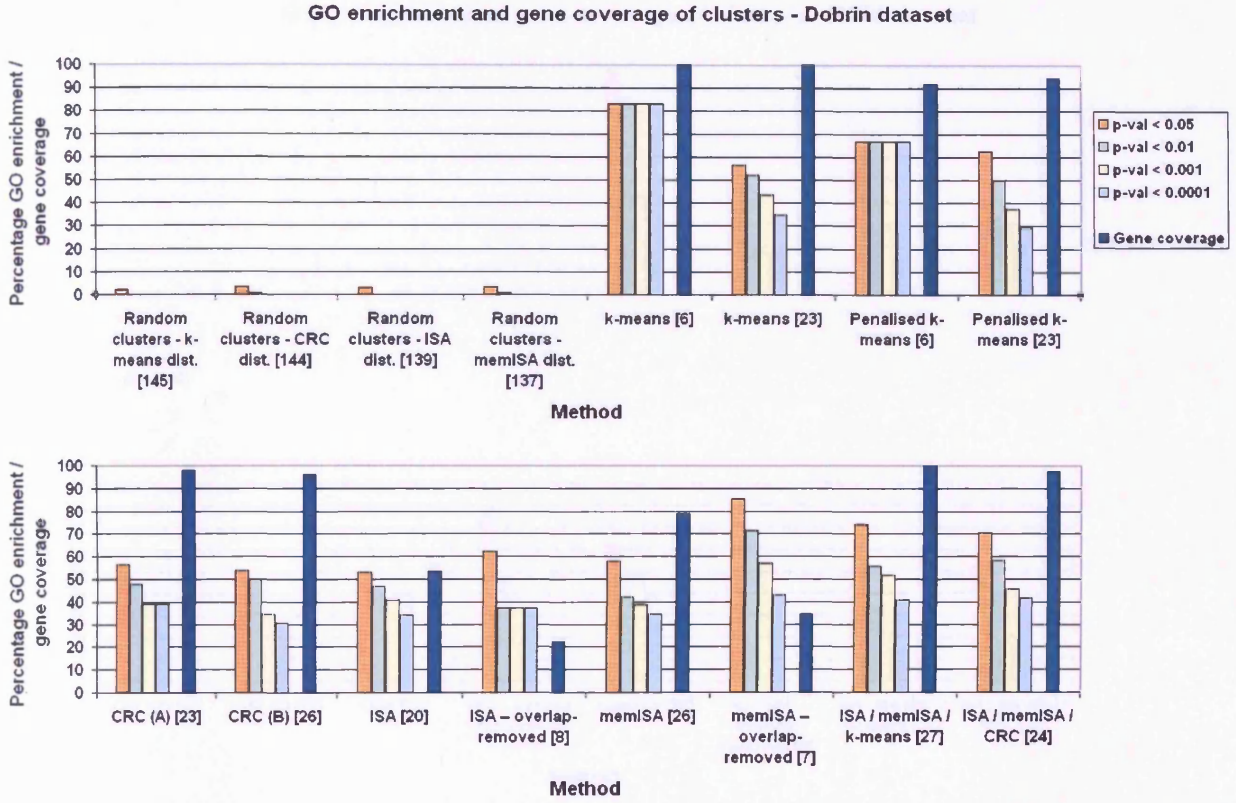


Figure 5

GO enrichment and gene coverage of clusters for all methods – Dobrin dataset. Orange, green, yellow and light blue bars are the percentage of clusters that are significantly enriched for one or more GO categories at $p < 0.05$, 0.01, 0.001 and 0.0001 respectively. Dark blue bar is gene coverage, the percentage of genes available on the chip that are assigned to at least one cluster. Numbers in square brackets are the number of clusters produced by that method. 'Dist.' = distribution of sizes. Parameter set A for CRC is 10 chains and 20 iterations per chain. Parameter set B for CRC is 20 chains and 40 iterations per chain.

7). Increasing the numbers of iterations or cycles or increasing the probability cut off had little effect which suggests that altering these parameters is unnecessary, and that the default values of 10 cycles and 20 iterations per cycle should be used for most datasets, with parameters only being increased on very large datasets. One problem noted with CRC was that analysing more than 202 samples caused the program to crash. This occurred on both Windows and Linux versions of the program, so was presumed to be an inherent problem with the program.

Effect of ISA parameters on GO enrichment

In contrast to CRC, changing the parameters of ISA can have unpredictable effects on the GO enrichment of its clusters, particularly after overlaps have been removed (see Figs. 5, 6, 7). The different values of t_C used in memISA and ISA for the PB cerebellum and MC66 datasets may help explain some unexpected results – in particular, the very large number of clusters produced by memISA prior to removing the overlaps in PB cerebellum, and the unexpectedly poor performance of memISA on the MC66 dataset. However, these may also be due to chance differences in the selection of random starting clusters, or to inherent qualities of the methods.

Effect of memISA parameters on GO enrichment

memISA is robust to the choice of f and n, as all of the combinations tried produced reasonable GO enrichments (see Table 2). $f = 0.7$ and $n = 5$ were chosen because they produced clusters with slightly better GO enrichments than other parameter sets.

Comparison of clusters detected

There was a large amount of overlap between the clusters produced using penalised k-means and k-means at k = 23, with the majority of clusters (from all three datasets) hav-

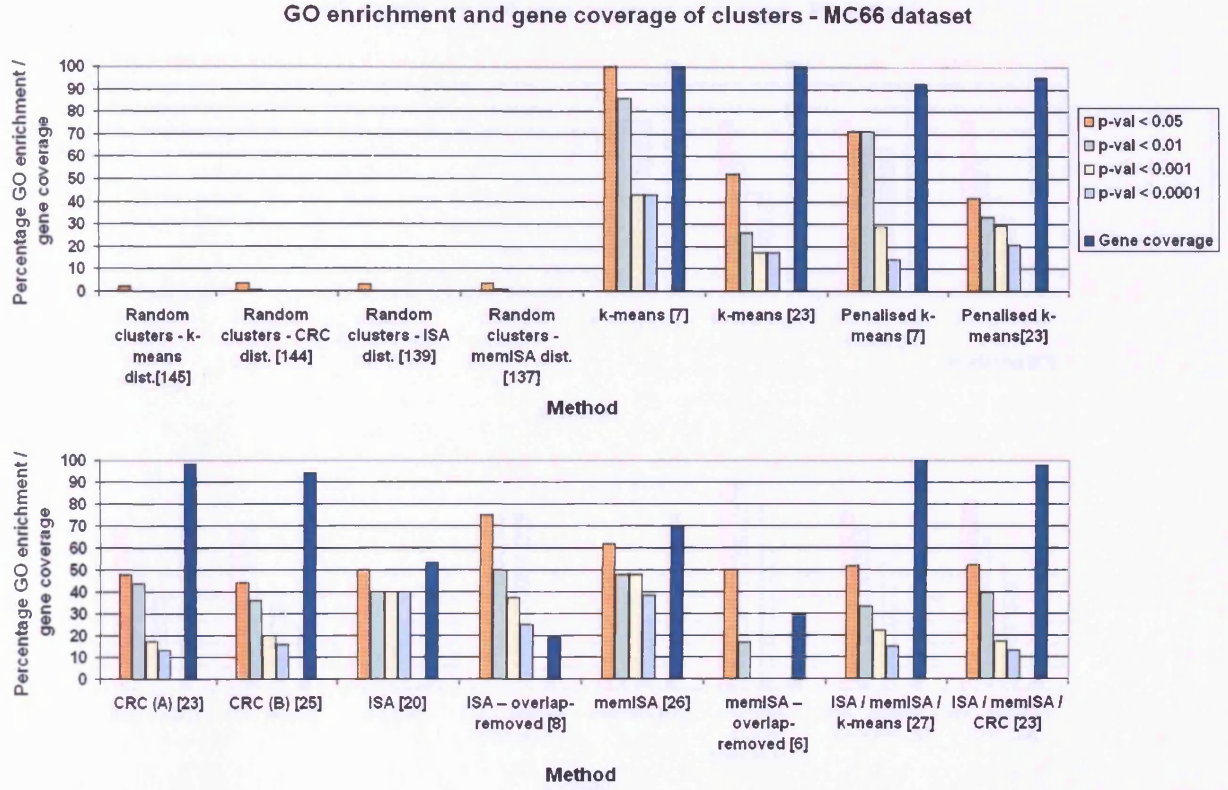


Figure 6
GO enrichment and gene coverage of clusters for all methods – MC66 dataset. Orange, green, yellow and light blue bars are the percentage of clusters that are significantly enriched for one or more GO categories at $p < 0.05$, 0.01, 0.001 and 0.0001 respectively. Dark blue bar is gene coverage, the percentage of genes available on the chip that are assigned to at least one cluster. Numbers in square brackets are the number of clusters produced by that method. 'Dist.' = distribution of sizes. Parameter set A for CRC is 10 chains and 20 iterations per chain. Parameter set B for CRC is 20 chains and 40 iterations per chain.

ing over 70% overlap with a cluster from the other method, and all others having over 40% overlap (see Table 3 and Additional Files 4 – AllOverlaps.xls for more detail). Since these methods found similar clusters, further analysis was focused on standard k-means clustering, as it had 100% gene coverage.

There was considerable overlap in the results obtained between k-means and CRC across all three datasets. This suggests that k-means and CRC find similar patterns within the datasets. Conversely, there was little overlap between either k-means or CRC and either memISA or ISA clusters. In the case of ISA, there were a few overlaps at 70% or above for each dataset. In the case of memISA, there was a large cluster that overlapped with several of the smaller clusters produced by CRC at 70% or more, plus one other 70% plus overlap between more similarly sized clusters, in all three datasets.

Removing clusters with over 70% intra-method gene overlap from the ISA and memISA cluster sets reduced the number of clusters considerably. These sets contained only 4–10 clusters and were much smaller than the original ones. However, their GO enrichments were generally considerably higher (see Figs. 5, 6, 7) but at the cost of a considerable drop in gene coverage.

Nearly all ISA clusters had over 70% overlap with a memISA cluster across all three datasets. However, less than half of the memISA clusters had over 70% overlap with a cluster from ISA, as many of the ISA clusters overlap with the same memISA cluster. This level of overlap is surprisingly high, considering that their post-processing

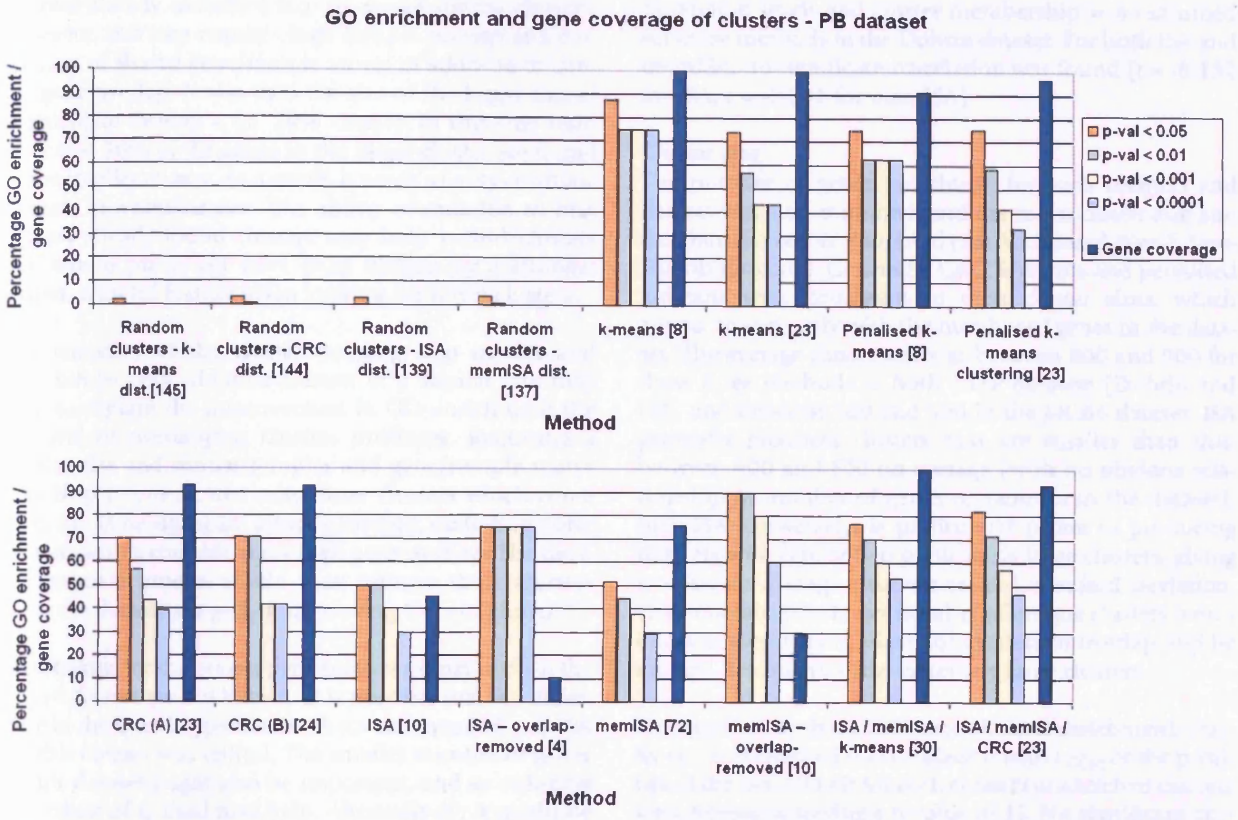


Figure 7

GO enrichment and gene coverage of clusters for all methods – PB dataset. Orange, green, yellow and light blue bars are the percentage of clusters that are significantly enriched for one or more GO categories at $p < 0.05$, 0.01, 0.001 and 0.0001 respectively. Dark blue bar is gene coverage, the percentage of genes available on the chip that are assigned to at least one cluster. Numbers in square brackets are the number of clusters produced by that method. 'Dist.' = distribution of sizes. Parameter set A for CRC is 10 chains and 20 iterations per chain. Parameter set B for CRC is 20 chains and 40 iterations per chain.

Table 3: Percentage overlap between clusters produced by different methods

	k-means	Penalised k-means	CRC	ISA	memISA
k-means	100	62.3	52.2	8.7	8.7
Penalised k-means	63.8	100	57.5	4.3	4.3
CRC	52.2	54.5	100	7.2	27.5
ISA	25	25	23.1	100	95.2
memISA	26.1	26.1	26.1	56.5	100

Values in table indicate the percentage of clusters produced by the method in the left margin that have over 70% gene overlap with one or more clusters produced by the method in the top margin.

regimes already include a step to merge similar clusters. However, this step requires high sample overlap and correlation of shared gene/sample scores in addition to simple gene overlap. It also uses the size of the larger cluster to calculate overlap – i.e. 50% overlap in this step indicates that 50% of the genes in the larger cluster are found in the smaller cluster. As a result, it tends to only combine clusters of a similar size. The ability of memISA to bias against already-found clusters may help it find clusters that would previously have been hidden by a stronger cluster, a useful feature when looking for novel clusters.

The tendency of the cluster merging step in ISA and memISA to only combine clusters of a similar size may help to explain the improvement in GO enrichment the removal of overlapping clusters produces. Requiring a similar size and similar samples and gene/sample scores may help to ensure that only those clusters which come from the same signal are actually merged, excluding noise clusters with a coincidentally high gene overlap. The overlap removal process would then remove these clusters from the dataset altogether, improving GO enrichment.

The reasons for the poorer performance of memISA on the MC66 dataset are not known. It is possible that the difference in the t_C and t_G parameters between memISA and ISA for this dataset was critical. The smaller number of genes in this dataset might also be important, and so reducing the values of t_G used may help. Alternatively, it might be that chance played a role. memISA may be inherently more prone to chance variation than ISA or CRC.

Combining methods

The cluster sets produced by combining the methods had similar gene coverage to those produced by CRC/k-means alone (see Figs. 5, 6, 7). They generally had a higher number of clusters. For the CRC/ISA/memISA combined set, the GO enrichment of these clusters was higher in the Dobrin and PB cerebellum datasets,. In the k-means/ISA/memISA combined sets, the gains in GO enrichment relative to k-means alone were generally smaller: under 5% at most levels of p. There were a few small losses in GO enrichment in some datasets and at some levels of p, but generally the impact on GO enrichment was still positive.

Gene coverage

Before highly overlapping clusters were removed from the clusters produced by ISA, k-means had the highest gene coverage (100% by definition), followed by CRC, and then by memISA and lastly ISA. However, these cluster sets are not directly comparable on number of clusters or on GO enrichment, as the cluster sets produced by ISA and memISA contain a large amount of redundancy.

As memISA and ISA had much lower gene coverage than k-means or CRC, the relationship between mean gene expression levels and cluster membership was examined for these methods in the Dobrin dataset. For both ISA and memISA, no significant correlation was found ($r = -0.132$ for ISA, $r = -0.081$ for memISA).

Cluster size

The number of genes per cluster for each method and dataset was also examined, and the mean cluster size and standard deviation computed (see Additional Files 5, SizeDistribution.xls). Generally, CRC, k-means and penalised k-means were consistent in their cluster sizes, which appear to vary only with the number of genes in the dataset. The average cluster size was between 800 and 900 for these three methods in both 133P datasets (Dobrin and PB), and between 500 and 600 in the MC66 dataset. ISA generally produces clusters that are smaller than this, between 400 and 600 on average (with no obvious relationship to number of genes or samples in the dataset). memISA, conversely, is particularly prone to producing datasets with one or two particularly large clusters, giving it a higher average cluster size and standard deviation. This is because the larger number of unique clusters it produces makes it more likely for clusters to overlap and be merged, leading to these extremely large clusters.

To examine whether cluster size affected enrichment, cluster size was checked for correlation with $\log_{10}$ of the p-values of the best GO hit for each cluster (unenriched clusters were treated as having a p-value of 1). No significant correlation was found for any of the methods.

Speed

The three datasets were used to evaluate approximate runtimes for the four methods (see Table 4). CRC and k-means are very fast methods, with a runtime of a few hours on current computer technology. ISA and memISA, meanwhile, are much slower, taking up to a month without parallelisation. Even with parallelisation using CONDOR, ISA and memISA can take over 24 hours for a full parameter set when post-processing is included. Restricting the parameters to t_G 2.1 and above, as in the non-overlapping cluster set before, reduces these times by up to half.

Enrichment of clusters for schizophrenia related genes

The clusters produced from the combined k-means/ISA/memISA method on the Dobrin dataset were tested for enrichment with 607 genes associated with schizophrenia according to a recent genome-wide association study[31], using the program EASE [32]. These 607 genes each contained at least one SNP associated with schizophrenia at an Armitage p-value of 0.005 or under. One cluster, containing 3093 genes and originally found by memISA, was enriched ($p = 0.0104$ after Bonferroni correction for 26 clusters).

Table 4: Comparison of method runtimes

Runtime on different datasets	ISA (using CONDOR)	memISA (using CONDOR)	CRC – 10/20	CRC – 20/40
Dobrin	23 h 6 min	37 h 22 min	2 h 12 min	7 h 53 min
MC66	17 h 23 min	28 h 55 min	1 h 15 min	4 h 33 min
PB cerebellum	15 h 11 min	24 h 13 min	1 h 7 min	3 h 53 min

Table showing the real-world time taken for the methods to run on each dataset.

This cluster was also tested for enrichment with 352 genes found to be differentially expressed between schizophrenics and controls in the analysis of the Stanley Medical Research Institute Online Genomics Database[16] at an uncorrected p-value of 0.02 or lower. The cluster was slightly enriched, at a p-value of 0.09.

Clusters from combined k-means/ISA/memISA in the independent MC66 dataset that shared over 45% of their genes with this enriched cluster were then identified. Two clusters were found (containing 2546 and 436 genes respectively), both of which were nominally enriched for both schizophrenia-associated genes (2546-gene cluster at p = 0.0127, 436-gene cluster at p = 0.0117) and genes differentially expressed in schizophrenia (2546-gene cluster at p = 0.0064, 436-gene cluster at p = 0.00047 – see Additional Files 6, Clusters.xls, for the gene symbols of the genes in these clusters). However, since these clusters have some overlap with the 3093-gene Dobrin cluster, this cannot be considered independent replication of the original cluster.

To avoid this confounding effect, their enrichment for schizophrenia-associated genes and genes differentially expressed in schizophrenia was determined using a permutation-based method. The 436-gene cluster remained significantly enriched for the schizophrenia associated genes, while the 2546-gene cluster showed some enrichment, but this was insufficient to be significant (permutation p = 0.169 for the 2546-gene cluster, permutation p = 0.0255 for the 436-gene cluster). However, both clusters were significantly enriched for genes differentially expressed in schizophrenia (permutation p = 0.0053 for the 2546-gene cluster, permutation p = 0.0005 for the 436-gene cluster).

These clusters were also examined for enrichment in KEGG and BioCarta pathways, using the Composite Regulatory Signature Database [33] (http://140.120.213.10:8080/crsd/main/home.jsp). The top hit for the Dobrin cluster and the 2546-gene MC66 cluster was the KEGG entry for the MAPK signalling pathway (p = $1.12e^{-7}$, FDR q = 0.00024 in Dobrin, p = $6.95e^{-10}$, FDR q = $1.46e^{-6}$ in MC66). The only significant hit for the MC66 436-gene cluster was from the BioCarta Synaptic Junction pathway (p = $3.88e^{-5}$, FDR q = $2.71e^{-2}$).

The MC66 436-gene cluster was also examined using GOstat, where the best hit was for GO:0007399 (nervous system development) GO category (p = 0.044 after FDR correction).

The three clusters were also tested for enrichment with genes found to be ten-fold or more upregulated in specific cell types within brain tissue according to Cahoy *et al* [34]- specifically, neurons, oligodendrocytes and astrocytes. All three clusters were found to be highly significantly enriched with genes upregulated in neurons (p = $2.5e^{-21}$ in Dobrin, p = $1.55e^{-16}$ in MC66, Bonferroni corrected). There was also enrichment for genes upregulated in oligodendrocytes (Dobrin p = 0.06, MC66 p = $2.4e^{-4}$, Bonferroni corrected) and astrocytes (Dobrin p = $5.13e^{-22}$, MC66 p = $2.26e^{-10}$, Bonferroni corrected).

Three overlapping clusters, enriched to varying degrees for either schizophrenia-associated genes or genes differentially expressed in schizophrenia, were found from the two independent dorsolateral prefrontal cortex datasets. The apparent excess of schizophrenia-associated genes in the 2546-gene MC66 cluster could be explained by its overlap with the Dobrin cluster. Thus, this cluster does not constitute independent evidence for schizophrenia-associated genes clustering together with respect to their expression levels. However, the 436-gene MC66 cluster remained significantly enriched when assessed by the permutation method. Both MC66 clusters did show significant over-representation for genes differentially expressed in schizophrenia, even after correction for the overlap with the Dobrin cluster. This demonstrates the ability of the methods to find potentially disease-related gene clusters that are replicable in multiple datasets.

The large size of two of the clusters makes inferences about individual genes difficult. However, both the larger clusters are enriched for genes present in the KEGG MAP kinase pathway, suggesting that this pathway may relate to the aetiology of schizophrenia. Members of this pathway have also been found to be differentially expressed

between controls and schizophrenics in other brain regions [35]. In addition, when structural variants such as microdeletions occur in the genomes of schizophrenics, they are particularly likely to occur in the genes of the MAP kinase pathway [36].

The smaller cluster was also found to be near-significantly enriched for serine/threonine kinase genes (the class of kinases which MAP kinases belong to), and also for synaptic junction and neurological development genes. As this cluster is enriched for both schizophrenia associated genes and genes differentially expressed in schizophrenia, further investigation of the role of these pathways in schizophrenia aetiology may be useful.

However, the MAP kinase-related genes present in the two large clusters do not overlap with the schizophrenia associated gene set or the differentially expressed in schizophrenia gene set (they share no genes at all in either the MC66 or Dobrin cluster). This might suggest the MAP kinase function of the clusters may be incidental to their roles in schizophrenia aetiology. Further investigation with other functional analysis tools may allow more biological inferences from these clusters.

Comparisons with other clustering method surveys

Our findings broadly agree with several other surveys of clustering methods (Figs. 5, 6, 7). Like Prelić *et al*, we find that ISA is an effective method that produces clusters with high GO enrichment [4], but our cluster sets generally do not have as high a proportion of GO enriched clusters as theirs. This is likely to be a consequence of the greater complexity of the input data.

Garge *et al* found k-means clustering effective [15] on a wide range of input datasets. This is echoed by the k-means cluster sets reported here, which have high GO enrichment and gene coverage scores. These scores were generally higher than CRC, the mixture modelling method examined here. This contrasts with the findings of Thalamuthu *et al*, who found that modelling methods were superior to k-means clustering [9]. This difference is again likely to be due to the datasets used; in particular the datasets used here were much larger in size.

Conclusion

k-means clustering, CRC, ISA and memISA are all potentially useful methods. Considered alone, k-means clustering is probably the most useful of the four, as it is fast, does not require parallelisation, and produces clusters with slightly higher levels of GO enrichment than CRC when producing similar numbers of clusters. When used to find smaller numbers of clusters more in line with the estimation of k, the GO enrichments are higher still, reaching 100% at some levels of p. It also assigns a cluster to every gene (100% gene coverage), unlike overlap-removed ISA and memISA (under 30% gene coverage). Although this must lead to some false positives, this does not seem to have affected the GO enrichment scores unduly, and is an advantage in exploratory studies where as wide a view as possible is desired. Furthermore, k-means is a relatively simple and very well understood method. This simplicity may be the reason for its good performance here, as it may allow it to cope with a wide variety of input data. CRC, conversely, has many more parameters and so may have had scope to become optimised for the smaller yeast and bacterial datasets it was built for and tested upon.

However, for the fullest picture of clusters available in a dataset, combining memISA, ISA and k-means is the best option, as it offers higher GO enrichment than k-means alone in two out of the three test datasets while retaining 100% gene coverage (see Figs. 5, 6, 7). Even in the MC66 dataset, it added additional clusters not found by k-means without reducing GO enrichment. One of these memISA clusters (found in both dorsolateral prefrontal cortex datasets) was found to be significantly enriched for schizophrenia-associated genes and genes differentially expressed in schizophrenia, further emphasising the utility of combining methods. If time allows, this combined method should be the method of choice for clustering microarray brain expression data.

Authors' contributions

ALR wrote all programs, performed analyses and wrote the paper. PH designed the permutation method used to assess cluster enrichment. LJ, PH and MCO constructively evaluated and edited the paper, and advised additional analyses. MJO provided access to the schizophrenia association data.

Additional material

Additional file 1

Survey of microarray expression data clustering methods. *Table showing the methods surveyed in the literature to decide which ones to investigate more closely.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-490-S1.xls]

Additional file 2

Details of ISA postprocessing regimes. *Table and description of ISA postprocessing regimes tried on the Dobrin dataset.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-490-S2.doc]

Additional file 3

Scripts to run ISA and memISA. ZIP archive containing the Perl and R scripts needed to run the version of ISA and memISA described here. Includes Instructions.txt, a step-by-step guide to using them.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-490-S3.zip]

Additional file 4

Inter-method gene overlap. Spreadsheet showing inter-method gene overlap for clusters from all methods, in all datasets. Overlap is defined as the percentage of genes present in the smaller cluster that are also found in the larger cluster.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-490-S4.xls]

Additional file 5

Distribution of cluster sizes. Spreadsheet showing number of genes present (cluster size) in each cluster for each method across all datasets. Also shows mean cluster size and standard deviation of cluster sizes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-490-S5.xls]

Additional file 6

Clusters enriched for schizophrenia-related genes. Spreadsheet showing the three clusters described in the paper. The 3093-gene cluster was made from the Dobrin dataset by memISA, and the 2546-gene and 436-gene clusters were made from the MC66 dataset by memISA.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-490-S6.xls]

Acknowledgements

ALR is supported by the MRC through a bursary. We would also like to thank Seth Dobrin for access to his data, and the two anonymous reviewers for their suggestions and constructive criticism.

References

1. Dettling M, Gabrielson E, Parmigiani G: **Searching for differentially expressed gene combinations.** *Genome Biology* 2005, **6(10)**:R88.
2. Wolfe CJ, Kohane IS, Butte AJ: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** *BMC Bioinformatics* 2005, **6(227)**:.
3. Allison DB, Cui X, Page GP, Sabripour M: **Microarray Data Analysis: from disarray to consolidation to consensus.** *Nature Reviews Genetics* 2006, **7**:55-65.
4. Prelić A, Bleuler S, Zimmerman P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzer E: **A systematic comparison and evaluation of biclustering methods for gene expression data.** *Bioinformatics* 2006, **22(9)**:1122-1129.
5. Riva A, Carpentier A-S, Torrésani B, Hénaut A: **Comments on selected fundamental aspects of microarray analysis.** *Comput Biol Chem* 2005, **29(5)**:319-336.
6. Fang Z, Liu L, Yang J, Luo Q-M, Li Y-X: **Comparisons of Graph-structure Clustering Methods for Gene Expression Data.** *Acta Biochimica et Biophysica Sinica* 2006, **38(6)**:379-384.
7. Stansberg C, Vik-Mo AO, Holdhus R, Breilid H, Srebro B, Petersen K, Jørgensen HA, Jonassen I, Steen VM: **Gene expression profiles in rat brain disclose CNS signature genes and regional patterns of functional specialisation.** *BMC Genomics* 2007, **8(94)**:.
8. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99(7)**:4465-70.
9. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC: **Evaluation and comparison of gene clustering methods in microarray analysis.** *Bioinformatics* 2006, **22(19)**:2405-2412.
10. de Hoon MJL, Imoto S, Nolan J, Miyano S: **Open Source Clustering Software.** *Bioinformatics* 2004, **20(9)**:1453-1454.
11. Qin ZS: **Clustering microarray gene expression data using weighted Chinese restaurant process.** *Bioinformatics* 2006, **22(16)**:1988-1997.
12. Bergmann S, Ihmels J, Barkai N: **Iterative signature algorithm for the analysis of large-scale gene expression data.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67(3 pt 1)**:031902.
13. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network.** *Nature Genetics* 2002, **31**:370-377.
14. Kloster M, Tang C, Wingreen NS: **Finding regulatory modules through large-scale gene expression analysis.** *Bioinformatics* 2005, **21(7)**:1172-1179.
15. Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB: **Reproducible Clusters from Microarray Research: Whither?** *BMC Bioinformatics* 2005, **6(Suppl 2)**:S10.
16. Higgs BW, Elashoff M, Richman S, Barci B: **An online database for brain disease research.** *BMC Genomics* 2006, **7(70)**:.
17. National Brain Databank: **Brain Tissue Gene Expression Repository.** [http://national_databank.mclean.harvard.edu/brainbank/Main].
18. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudney D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles – database and tools update.** *Nucleic Acids Research* 2006:D760-D765.
19. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Research* 2002, **30(1)**:207-210.
20. **The R Project for Statistical Computing** [http://www.R-project.org]. R: A language and environment for statistical computing
21. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of Gene Chip data.** *Nucleic Acids Research* 2005, **33(20)**:e175.
22. Bolstad BM, Irizarray RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on bias and variance.** *Bioinformatics* 2003, **19(2)**:185-193.
23. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21(18)**:3587-3595.
24. Beißbarth T, Speed TP: **GOstat: Find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20(9)**:1464-1465.
25. **WWW-Mechanize** [http://search.cpan.org/dist/WWW-Mechanize/]
26. Hartigan JA, Wong MA: **A K-Means Clustering Algorithm.** *Applied Statistics* 1979, **28(1)**:100-108.
27. Dembélé D, Kastner P: **Fuzzy C-means method for clustering microarray data.** *Bioinformatics* 2003, **19(8)**:973-980.
28. Tseng GC: **Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data.** *Bioinformatics* 2007, **23**:2247-2255.
29. **vegan: Community Ecology Package, R Package** [http://vegan.r-forge.r-project.org/]
30. Thain D, Tannenbaum T, Livny M: **Distributed computing in practice: the Condor experience.** *Concurrency and Computation: Practice and Experience* 2004, **17(2–4)**:323-356 [http://www.cs.wisc.edu/condor/doc/condor-practice.pdf].
31. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, Hamshere M, Carroll L, Georgieva L, Dwyer S, Holmans P, Marchini JL, Spencer CCA, Howie B, Leung H-T, Hartmann AM, Moller H-J, Morris DW, Shi Y, Feng G, Hoffmann P, Propping P, Vasilescu C, Maier W, Rietschel M, Zammit S, Schumacher J, Quinn EM, Schulze TG, Williams NM, Giegling I, Iwata N, Ikeda M, Darvasi A, Shifman S, He L, Duan J, Sanders AR, Levinson DF, Gejman PV, Cichon S, Nothen MM, Gill M, Corvin A, Rujescu D, Kirov G,

Owen MJ: **Identification of loci associated with schizophrenia by genome-wide association and follow-up.** *Nat Genet* 2008, **40(9)**:1053-1055.

32. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA: **Identifying Biological Themes within Lists of Genes with EASE.** *Genome Biology* 2003, **4(6)**:4.
33. Liu CC, Lin CC, Chen WSE, Chen HY, Chang PC, Chen JJW, Yang PC: **CRSD: a comprehensive web server for composite regulatory signature discovery.** *Nucleic Acids Research* 2006, **34**:W571-W577.
34. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA, Thompson WJ, Barres BA: **A transcriptome database for astrocytes, neurons and oligodendrocytes: a new resource for understanding brain development and function.** *Journal of Neuroscience* 2008, **28(1)**:264-278.
35. Kyosseva SV: **Differential expression of mitogen-activated protein kinases and immediate early genes fos and jun in thalamus in schizophrenia.** *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 2004, **28**:997-1006.
36. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J: **Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia.** *Science* 2008, **320(5875)**:539-543.

Appendix D Paper based upon chapter 3: Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain

Currently under review at Molecular Psychiatry.

Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain

Alexander L Richards[1], Lesley Jones[1], Valentina Moskvina[1], George Kirov[1], Pablo V Gejman[2], Douglas F Levinson[3], Alan R Sanders[2], Molecular Genetics of Schizophrenia Collaboration (MGS)[4], International Schizophrenia Consortium (ISC)[4], Shaun Purcell[5,6,7,8], Peter M Visscher[9], Nick Craddock[1], Michael J Owen[1], Peter Holmans[1], *Michael C O'Donovan[1]

[1] MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[2]Center for Psychiatric Genetics, Department of Psychiatry and Behavioral Sciences, Northshore University Health System Research Institute, 1001 University Place, Evanston, IL 60201, USA

[3]Department of Psychiatry, Stanford University, Stanford, CA, USA

[4] Full author details and affiliations are given in acknowledgements section.

[5] Psychiatric and Neurodevelopmental Genetics Unit, and [6]Center for Human Genetic Research, Massachusetts General Hospital, Massachusetts 02114, USA

[7] Stanley Center for Psychiatric Research, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA

[8] The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA

[9] Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, 300 Herston Road, Brisbane 4006, Australia

* corresponding author; Michael C O'Donovan, MRC Centre for Neuropsychiatric Genetics and Genomics, Henry Wellcome Building, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK. Telephone: 44 (0)2920687066 Fax: 44 (0)2920687068 email: odonovanmc@Cardiff.ac.uk

Running title: eQTLs in schizophrenia

Abstract

It is widely thought that alleles that influence susceptibility to common diseases, including schizophrenia, will frequently do so through effects on gene expression. Since only a small proportion of the genetic variance for schizophrenia has been attributed to specific loci, this remains an unproven hypothesis. The International Schizophrenia Consortium (ISC) recently reported a substantial polygenic contribution to that disorder, and that schizophrenia risk alleles are enriched among SNPs selected for marginal evidence for association ($p<0.5$) from genome wide association studies (GWAS). It follows that if schizophrenia susceptibility alleles are enriched for those that affect gene expression, those marginally associated SNPs which are also eQTLs should carry more true association signals compared with SNPs which are not. To test this, we identified marginally associated ($p<0.5$) SNPs from two of the largest available schizophrenia GWAS datasets. We assigned eQTL status to those SNPs based upon an eQTL dataset derived from adult human brain. Using the polygenic score method of analysis reported by the ISC, we observed and replicated the observation that higher probability *cis*-eQTLs predicted schizophrenia better than those with a lower probability for being a *cis*-eQTL. Our data support the hypothesis that alleles conferring risk of schizophrenia are enriched among those that affect gene expression. Moreover, our data show that notwithstanding the likely developmental origin of schizophrenia, studies of adult brain tissue can in principle allow relevant susceptibility eQTLs to be identified.

Introduction

A high proportion of mutations for simple (Mendelian) genetic disorders exert their pathogenic effects by altering the structure of the encoded protein but this does not appear to be the case for the majority of susceptibility alleles for common phenotypes identified through genome-wide association studies (GWAS) (1). This is compatible with the hypothesis that inherited variation that impacts upon mRNA expression plays an important part in susceptibility to complex traits (2-4). Only a small proportion of the genetic variance for risk to common diseases has been attributed to specific loci (5, 6) including schizophrenia (7-10). Therefore, while it has been argued that gene expression analysis is a key component of understanding the pathogenesis of schizophrenia (11,12), the hypothesis of the involvement in that disorder of alleles that influence gene expression is unproven.

From the perspective of identifying risk alleles, the hypothesis that susceptibility variants for schizophrenia will be enriched for variants that influence mRNA expression is not merely of academic interest. We (11, 12) and others (13) have reported associations between gene expression and genetic variants whose associations with schizophrenia are controversial, the idea being that association with expression lends credibility to association with disease status. Others have used this principle in non-psychiatric disorders to localise the likely susceptibility genes or functional variants within regions of association (14). Since the effect sizes of common alleles are small (7), and most are unlikely to be reliably separated from chance findings in the full genome context in the near future (9), the ability to assign an enhanced prior probability to variants associated with gene expression may be of value in identifying novel disease associations.

Although the convergent use of expression and genetic data for informing pathophysiological theory seems intuitively reasonable (15), the validity of this approach for informing genetic studies depends on the assumption that true associations are enriched among variants that impact upon gene expression. Moreover, in the case of schizophrenia, attempts to relate disorder-associated variants to gene expression are generally based upon mRNA studies of adult brain, peripheral tissues, or cell lines. Whether such studies are justified for disorders like schizophrenia, whose origins are thought to be developmental, is unclear. Interestingly, however,

in a recent study (16), SNPs that affected expression in lymphoblasts were enriched among the top 10,000 GWAS associations for a number of disorders including associations from a bipolar GWAS.

Here, we tested the hypothesis that polymorphisms that are associated with schizophrenia are enriched among those that show evidence for association to gene expression in adult brain. Loci that exert an effect on gene expression are often called expression quantitative trait loci (eQTLs) (17). In the present study, to identify putative eQTLs, we used the dataset originally reported by Myers and colleagues (18, 19), currently the largest expression dataset derived from human brain available to us that also contains genotype data for each sample.

To identify sets of variants enriched for schizophrenia susceptibility alleles, we exploited the approach of the International Schizophrenia Consortium (ISC) (7) who recently demonstrated the existence of large numbers of risk alleles for schizophrenia. They also showed that these are enriched among large sets of SNPs surpassing very liberal significance thresholds of association (e.g. $P<0.5$). The ISC defined sets of putative schizophrenia risk alleles in a training GWAS dataset as those that were more common in cases than controls at loci meeting the relaxed thresholds. Individuals in independent test GWAS datasets were assigned a 'polygenic score' based upon the number of putative risk alleles carried by that individual, and then the scores for cases and controls in those datasets were compared. In independent datasets, these 'polygenic scores' were significantly higher in cases than in controls, with the most significant distinction between groups occurring when the threshold for association in the training GWAS was set at $p<0.5$. Modeling suggested that the most plausible explanation for this finding was that there is a substantial polygenic component to schizophrenia comprising thousands of risk alleles, and that this contributes at least 30% of the overall variance in risk of the disorder at the population level.

Here, we used this general approach to test whether eQTLs are enriched among schizophrenia associated alleles. We defined schizophrenia 'risk' alleles according to the method reported by the ISC (7) in a subset of the ISC data and also in the European American subset of the Molecular Genetics of Schizophrenia study (10). Using the dataset of Myers and colleagues (18), these SNPs were then classified as 'top eQTL' and 'bottom eQTL' sets based upon their p-

value for association with expression levels of transcripts, and these sets were then tested for differences in their polygenic scores in cases and controls independent of the training sets.

Method

The eQTL dataset (18, 19) contains genotypes (Affymetrix GeneChip Human Mapping 500K Array) from 380157 SNPs, and expression (Illumina v1 Human RefSeq-8 BeadChip) data on 8650 transcripts, meeting the quality control criteria described in (22). There were 176 Alzheimer's disease cases and 188 controls in the dataset, however our analysis was restricted to controls to exclude the impact of neurodegeneration on gene expression measures. We selected this option rather than allowing for affected status in the analysis as a crude categorical adjustment will not allow for a number of variables within the affected group that can be expected to have major effects, including aetiological heterogeneity, duration of illness, and rate of disease progression.

Beginning with the rank-invariant normalised expression data (18), samples with over 10% missing data were removed, as were probes with over 25% missing data in the remaining individuals. Where multiple probes mapped to the same gene, we retained only the probe with the lowest proportion of missing data (arbitrarily retaining the first to appear in the dataset file in the case of a tie). To minimize the impact of different brain regions in the dataset, we included only samples from the two most common regions represented in the study (frontal cortex and temporal cortex). Overall, we retained 163 samples and 8361 probes for analysis.

As in the primary publication, the data were log transformed to minimise the effect of departures from normality (using the statistical package R (20)). The log-transformed expression values were adjusted for a number of non-genetic covariates using linear regression. These covariates were gender, post mortem interval, brain area, age at death, institute and hybridisation date, and the expression value for *Enolase 2* (ENO2). The residuals of this regression were used as covariate-adjusted expression values in all further analyses. ENO2 is a neuronal marker. Our intention in making this correction was to reduce expression variance arising from varying proportions of neurons in the samples (21, 22). We were unable to adjust our analyses for pH as

those data were not available. However, we note that failure to adjust for this, or for other important variables that might lead to classification errors (false positive or negative) will bias our study towards the null. This is because false calls will blur any true differences between top and bottom eQTL groups, including differences in the extent to which they are enriched for schizophrenia susceptibility alleles.

For the Myers genotype data, we used the same quality control metrics as the original publication (18). All SNPs were required to have minor allele frequency of at least 1%, a call rate of at least 90%, and an exact Hardy-Weinberg equilibrium p-value > 0.05.

The ISC (7) and MGS (10) GWAS datasets were used for the study as these are currently the largest GWAS datasets available to us. We essentially followed the study design of the ISC. The ISC dataset was divided to create training and test subsets by assigning alternate cases and alternate controls to the training and test datasets; these we call the 'Split ISC' datasets. To derive a set of putative risk alleles independent of the ISC, we used the p-values from the MGS European American dataset (10) and tested these in the full ISC dataset. Full descriptions of those datasets are given in the primary publications (7, 10).

eQTL determination

Linear regression of the expression values for each gene (correcting for covariates) on SNP genotypes (coded as the number of minor alleles: 0, 1 or 2) was performed using PLINK v1.05 (23, 24). This gave p-values for association between each SNP and mRNA expression as measured by each probe-set. To test our hypothesis, we based our analysis upon *cis*-eQTL p-values. *Cis*-eQTLs are variants that are in chromosomal proximity to the transcripts they putatively regulate, and have a higher prior probability for being true eQTLs than *trans*-eQTLs (17), the latter being defined on the basis of association with transcripts with which they are not co-located. Moreover, *trans*-eQTL analysis involves a much greater degree of multiple testing (all SNPs against all probesets) than *cis*-eQTL analysis. These considerations suggest that sets of 'top *cis*-eQTLs' will be more greatly enriched for true eQTLs than sets of top *trans*-eQTLs, so restriction to *cis*-eQTLs should enhance the power of our analysis. *cis*-eQTLs were ranked by p-value with respect to any transcript within

100kb of the SNP locus. The criterion of 100kb is to an extent arbitrary, but was based upon a previous study suggesting that *cis*-eQTLs are enriched within this boundary (25). If a SNP was within range of multiple transcripts, the lowest p-value for any transcript was taken as the eQTL p-value.

Given the presumed lower probability for any *trans*-eQTL representing a true association, we expected that even if our primary hypothesis was correct, SNPs selected on this basis of *trans*-eQTL status would be less effective at distinguishing between cases and controls. Nevertheless, as a secondary analysis, we explored the relative ability of top and bottom eQTLs after ranking those loci by the most significant p-value for association to any transcript in the dataset.

We did not specifically exclude probes corresponding to target sequences that contain SNPs, some of which might influence the efficiency of probe hybridisation. Where this occurs, expression of the target transcript could appear correlated with the SNP in the probe sequence, which could then be falsely classified as an eQTL, and the same is true for any SNPs in high linkage disequilibrium (LD) with that SNP. Conversely, where there is a true eQTLs that is in weak or low LD with a second SNP under a probe that influences hybridisation efficiency, the impact of that second SNP is likely to be to reduce the estimated correlation between the eQTL and gene expression, the result being a tendency to false negative eQTL classification. As argued above, eQTL misclassifications will bias this study towards towards the null. Nevertheless, for information, we present some summary information about the occurrence of known SNPs within probe target sequences.

Of 1372 probes representing the transcripts associated with the top 5% of QTLs, only 56 (4%) contain a SNP called at high quality (less than 5% missing genotypes) with a minor allele frequency >1% in the HapMap CEU sample (HapMap Phase 2 version 23). Only a single SNP out of the 2580 SNPs that comprised our pruned list of top 5% of eQTLs was either within a probe, or in strong LD (R2>0.8) with a variant known to be within a probe. This appears to contrast with an earlier study (21) in which about 13% of significant eQTLs were to probes targeting polymorphic transcript sequences. However, that earlier study was concerned with highly significantly associated eQTLs which might be particularly enriched for this particular artefact. Also, the dataset

we used (22) was much more stringently filtered (reducing transcripts from 14,078 to 8650) than the other study (21), and we additionally further reduced this by removing probes with >25% missing data. Probes binding to sequences with common SNPs that influence hybridisation may have relatively high data-failure rates, and therefore we speculate this process would remove some of the affected transcripts. Finally, we aggressively LD prune our SNP data, which reduces the probability of including a SNP in even moderately high LD with a SNP in a probe sequence.

Post hoc analysis confirmed that our conclusions remain the same whether or not we exclude probes corresponding to sequences with known SNPs. Since the average impact of variants under target probes on misclassification is uncertain (22) but in the context of this study, it is likely to be a trivial source of misclassification compared with chance (see above), and since any bias is conservative (i.e. towards the null) we present the analysis of all probes in this manuscript.

Risk allele counts

The SNPs available in the training datasets were placed into the following categories according to eQTL p-value: top 5% eQTLs (corresponding to $p<0.02$), top 50% eQTLs (corresponding to $p<0.38$), bottom 50% eQTLs, and bottom 5% eQTLs. As in the ISC study, the SNPs in all sets were LD pruned (PLINK's --indep-pairwise option; window size=200, step=5, r^2 threshold=0.25).

In the randomly split ISC training datasets, as in the ISC paper (7), allelic p-values and odds ratios for association were calculated by a Cochrane-Mantel-Haenszel test conditioned by country of origin using the QC-cleaned datasets provided by that group. Training on the MGS European American Sample was based upon the association results that formed the basis of the primary publication (10). SNPs that had association $p<0.5$ in training sets were carried through for polygenic score analysis. Alleles that were more common in cases were defined as risk alleles. PLINK (using the --score option) was then used to perform a count of the number of risk alleles for each sample in the target dataset, weighted by the odds ratio at each SNP. PLINK gives the mean risk allele score for each individual, that is, the risk allele score is divided by the number of SNPs for which there are data in that individual.

Controlling for minor allele frequency and population stratification

For each pruned *cis*-eQTL SNP list, to test if ranking the SNPs by their most significant eQTL p-value introduced systematic differences in allele frequency between high and low eQTL SNP sets we calculated the mean and standard deviation of MAF and then compared them using t-tests.

To examine whether our results might be influenced by population stratification, we obtained from a previous study (26) F_{ST} values derived from the ISC sample for each SNP. F_{ST} is a measure of population stratification and is based upon the sequence similarity of members of a subpopulation, compared to their similarity with the population as a whole (27). In a stratified population, members of the subpopulations will be more similar to each other than to the whole population, leading to a high F_{ST} score.

SNPs with as close a F_{ST} value as possible to each SNP in the smaller of the two SNP lists (top or bottom eQTL) were extracted without replacement from the larger of the two SNP lists (top or bottom) to create eQTL sets matched for F_{ST}. A small number of SNPs could not be matched (those where the closest match differed by an F_{ST} >0.0005) and were removed from the analysis. This created pairs of SNP lists with the same number of SNPs and extremely similar means and standard deviations of F_{ST} (Supplementary Table 1).

Logistic regression

For each individual in the test ISC datasets, we calculated the difference between the polygenic score derived from the top eQTLs (5% or 50%) and that derived from the bottom eQTLs (5% or 50%), the null hypothesis being that these differences should be equal in cases and controls. We performed logistic regression of case/control status on risk allele score difference and also ISC sample country of origin to evaluate the significance of this difference. A significant positive regression coefficient indicates that the difference in risk scores between cases and controls is significantly greater for the top eQTL set.

Logistic regression of disease status on risk allele score was also calculated to determine how well each individual SNP list predicted disease status. We calculated the Nagelkerke pseudo-R^2 (28), which is a measure of how well the risk allele score predicts schizophrenia by subtracting

the R^2 of the regression without the risk allele score term included from the R^2 of the regression with the risk allele score term included.

Results

When we defined risk alleles using half of the ISC sample as the training set (Table 1, Split ISC analyses), the difference in the scores between the top and bottom *cis*-eQTLs was significantly greater in the cases than in the controls for all analyses. This is consistent with the hypothesis that schizophrenia susceptibility alleles are enriched among *cis*-eQTLs. Similar findings were observed when the risk alleles were defined from the MGS European dataset (entirely independent of the ISC dataset), with significant replication being obtained for two of the tests, even corrected for three replication tests (Table 1). Supplementary table 2 lists the pruned set of SNPs comprising those that were associated in the MGS training set at $P<0.5$ that were both within the top 5% of eQTLs and for which the allele designated as the 'risk' allele in the MGS sample was associated in the ISC sample at a nominally significant level ($P<0.05$). We should stress that for the reasons discussed already in this manuscript, the existence of potential sources of misclassification means the confidence that any one of these variants is either a genuine eQTL or that it is associated with the disorder is low, our study being designed to test a general hypothesis using global datasets and a methodology that can tolerate low signal to noise ratios rather than to identify individual findings of high significance. We also note the information driving our analysis comes not just from those alleles that are associated at nominally significant levels; rather it comes from the cumulative scores from all variants included in the analysis, however weakly associated they are.

There were no significant differences between the scores from the top and bottom *trans*-eQTLs between cases and controls (data not shown) in any analysis.

Minor allele frequency and population stratification

Of the 5 tests in which the top *cis*-eQTLs were significantly better at discriminating case-control status, the mean MAF was slightly but significantly higher in 2 of the top *cis*-eQTL sets, whereas

for the other three tests, any trends were for a lower MAF in the top *cis*-eQTLs set (Table 1). This suggests that our findings are unlikely to be due to differences in MAF between the sets.

However, for each analysis, the top *cis*-eQTL set had significantly higher mean F_{ST} than the bottom eQTL SNP lists indicating that our analysis might be confounded by enhanced stratification in the top *cis*-eQTL set. For this to bias to our results, the MGS and ISC samples would have to be ascertained such that the same alleles are similarly biased towards overrepresentation in cases in each dataset. Although we do not consider this likely (7), to evaluate whether this does influence our results, we repeated all analyses using F_{ST} matched SNP sets. After matching, there were no significant differences in mean F_{ST} between pairs of comparator groups (Supplementary Table 1). Nevertheless, for two of the three analyses in the split ISC datasets, the top *cis*-eQTLs significantly discriminated better between cases and controls than the bottom *cis*-eQTLs, both of which replicated when the MGS sample was used as the training set. Moreover, in the F_{ST} adjusted data, for two of the significant runs, the top *cis*-eQTL sets had lower MAF and for two of the runs, the top group had higher MAF. We therefore conclude that our findings are not driven by systematic biases in these variables.

Discussion

To date, only a minuscule proportion of genetic susceptibility to schizophrenia, or indeed any psychiatric disorder, has been explained by robustly associated DNA variants. Moreover, in no case has the functional effect of a DNA variant responsible for a robust schizophrenia association been determined. It follows that the basic mechanisms by which genetic variation contribute to this disorder are unknown. One leading hypothesis is that a substantial amount of genetic risk is conferred by common alleles that influence gene expression, that is, common *cis*-eQTLs. However, while the existence of many common schizophrenia risk alleles has been demonstrated (7), there is no evidence to support the hypothesis that any of these influence gene expression. In the light of a recent rekindling of interest in the hypothesis that genetic risk for the disorder is mainly attributable to rare variants of major effect (29), which by analogy with Mendelian disorders

are likely to be dominated by mutations that change the protein coding sequences of genes, the demonstration of a contribution from *cis*-eQTLs is of practical importance for several reasons.

The search for functional variants underpinning disease associations observed in GWAS studies is in general proving to be far from a trivial endeavour. Although it is relatively simple to scan the exonic sequences of individual genes for common non-synonymous variants, the process of scanning the full genomic context of a gene for potential *cis*-eQTLs, and then demonstrating that those variants impact on expression in a disease relevant manner remains arduous. To justify those endeavours, it is important to demonstrate that effects on gene expression are relevant mechanisms underpinning the influence of common susceptibility variants. Second, as discussed above, the use of gene expression data to support genetic associations or to assign higher prior probability to particular variants requires evidence that *cis*-eQTLs do in fact have a higher probability of being associated with disease. Finally, even if risk variants are enriched for common *cis*-eQTLs, it cannot be taken for granted that adult brain tissues, far less other sources of mRNA, are suitable substrates for generating eQTLs for disorders like schizophrenia whose presumed origins are developmental.

Using two independent training datasets we now demonstrate that among the variants selected for marginal association to schizophrenia, those that additionally show evidence for being *cis*-eQTLs predict affection status better than those variants showing no evidence for being *cis*-eQTLs. Thus, we show for the first time that schizophrenia risk alleles are indeed enriched for eQTLs. As expected from the ISC study, no set of SNPs explained more than a small fraction of the variance in disease risk (Table 2), although more comprehensive genome coverage in more powerful larger samples is likely to explain a much higher proportion (7).

In contrast to the findings with *cis*-eQTLs, SNPs, classified on the basis of potential *trans* effects were not superior at predicting schizophrenia affection status. This may be because the much greater multiple testing burden inherent to *trans*-eQTL analysis means a smaller proportion of the top rated *trans*-eQTLs are true positives.

While top sets of *cis*-eQTLs perform better than bottom sets in predicting disease risk, it is evident (Table 2) that even the latter significantly predict affected status. Moreover, after training in

the MGS dataset, the top 5% of eQTLs were only 1.3 times more likely than the bottom 5% of eQTLs to achieve a nominal significance level of $p<0.05$ in the ISC dataset. This might be because a substantial part of the true association signal is not related to variants that alter gene expression. Alternatively, it may be that virtually all true association signals are eQTLs, but that many of these were incorrectly classified. We note the sample from which we derived eQTL status is relatively small in GWAS terms, and therefore has limited power to identify weak eQTLs. Moreover, the already limited power will be further constrained by variance introduced by the many well known confounders that plague the use of *post mortem* expression datasets (30). Both factors are likely to result in eQTL classification errors.

Potentially pointing to an important impact of eQTL misclassification, comparisons of the most extreme *cis*-eQTL categories (top and bottom 5% sets) revealed considerable differences in the ability of those groups to discriminate case and control status (Table 2). Thus, the risk allele score differences between cases and controls were about 10 times greater for the top 5% of *cis*-eQTLs and were 3-4 orders of magnitude more significant than they were for the bottom 5% of *cis*-eQTLs. The former also had better predictive power as indicated by a larger Nagelkerke R^2, despite greater numbers of SNPs in the bottom 5% group. Indeed the bottom 5% of *cis*- eQTLs were either not significant predictors at all (trained in ISC) or the statistical significance of prediction was relatively modest (trained in the MGS). Assuming the extreme top and bottom *cis*-eQTL groups contain SNPs that are least likely to be misclassified, we postulate that the proportion of the polygenic signal captured by eQTLs will be enhanced by more precise delineation of eQTL status. Better eQTL classification could be relatively simply achieved by 1) using larger human brain expression and SNP datasets 2) increasing the transcriptome coverage; the present analysis only incorporates 8361 probes representing only 25-30% of the protein encoding genes in the human genome (31) and 3) using expression datasets derived from different brain regions rather than simply cortical structures as we have done here, and from different stages of human development, as functional variants may have variable temporal and spatial influences.

In summary, we have undertaken the first large scale analysis of the hypothesis that schizophrenia risk is mediated in part by common DNA variants that influence gene expression.

Our results support this hypothesis. In doing so, we provide the first systematic demonstration that gene expression studies in human adult brain are informative for genetic investigations of schizophrenia. Larger eQTL datasets with the power to achieve lower eQTL misclassification rates, representing different brain regions and developmental stages, will be required to exploit the enhanced prior probability for *cis*-eQTLs to identify specific susceptibility loci.

Acknowledgements

This work was supported by grants from the MRC and the Wellcome Trust. AR was initially supported by a MRC PhD studentship, and subsequently by NIMH (USA) CONTE Award: 2 P50 MH066392-05A1.

The following authors are included under:

Molecular Genetics of Schizophrenia Collaboration

PV Gejman (Evanston Northwestern Healthcare and Northwestern University, IL, USA), AR Sanders (Evanston Northwestern Healthcare and Northwestern University, IL, USA), J Duan (Evanston Northwestern Healthcare and Northwestern University, IL, USA), DF Levinson (Stanford University, CA, USA), NG Buccola (Louisiana State University Health Sciences Center, LA, USA), BJ Mowry (Queensland Centre for Mental Health Research, and Queensland Institute for Medical Research, Queensland, Australia), R Freedman (University of Colorado Denver, Colorado, USA), F Amin (Atlanta Veterans Affairs Medical Center and Emory University, Atlanta, USA), DW Black (University of Iowa Carver College of Medicine, IA, USA), JM Silverman (Mount Sinai School of Medicine, New York, USA), WJ Byerley (University of California at San Francisco, California, USA), CR Cloninger (Washington University, Missouri, USA).

International Schizophrenia Consortium (ISC)

Michael C. O'Donovan (Cardiff University, Cardiff, UK), George K. Kirov (Cardiff University, Cardiff, UK), Nick J. Craddock (Cardiff University, Cardiff, UK), Peter A. Holmans (Cardiff University,

Cardiff, UK), Nigel M. Williams (Cardiff University, Cardiff, UK), Lyudmila Georgieva (Cardiff University, Cardiff, UK), Ivan Nikolov (Cardiff University, Cardiff, UK), N. Norton (Cardiff University, Cardiff, UK), H. Williams (Cardiff University, Cardiff, UK), Draga Toncheva (University Hospital Maichin Dom, Sofia, Bulgaria), Vihra Milanova (Alexander University Hospital, Sofia, Bulgaria), Michael J. Owen (Cardiff University, Cardiff, UK), Christina M. Hultman (Karolinska Institutet, Stockholm, Sweden and Uppsala University, Uppsala, Sweden), Paul Lichtenstein (Karolinska Institutet, Stockholm, Sweden), Emma F. Thelander (Karolinska Institutet, Stockholm, Sweden), PatrickSullivan (University of North Carolina at Chapel Hill, North Carolina, USA), Derek W. Morris (Trinity College Dublin, Dublin, Ireland), Colm T. O'Dushlaine (Trinity College Dublin, Dublin, Ireland), Elaine Kenny (Trinity College Dublin, Dublin, Ireland), Emma M. Quinn (Trinity College Dublin, Dublin, Ireland), Michael Gill (Trinity College Dublin, Dublin, Ireland), Aiden Corvin (Trinity College Dublin, Dublin, Ireland), Andrew McQuillin (University College London, London, UK), Khalid Choudhury (University College London, London, UK), Susmita Datta (University College London, London, UK), Jonathan Pimm (University College London, London, UK), Srinivasa Thirumalai (West Berkshire NHS Trust, Reading, UK), Vinay Puri (University College London, London, UK), Robert Krasucki (University College London, London, UK), Jacob Lawrence (University College London, London, UK), Digby Quested (University of Oxford, Oxford, UK), Nicholas Bass (University College London, London, UK), Hugh Gurling (University College London, London, UK), Caroline Crombie (University of Aberdeen, Aberdeen, UK), GillianFraser (University of Aberdeen, Aberdeen, UK), Soh Leh Kuan (University of Aberdeen, Aberdeen, UK), Nicholas Walker (Ravenscraig Hospital, Greenock, UK), David St Clair (University of Aberdeen, Aberdeen, UK), Douglas H. R. Blackwood (University of Edinburgh, Edinburgh, UK), Walter J. Muir (University of Edinburgh, Edinburgh, UK), Kevin A. McGhee (University of Edinburgh, Edinburgh, UK), Ben Pickard (University of Edinburgh, Edinburgh, UK), PatMalloy (University of Edinburgh, Edinburgh, UK), Alan W. Maclean (University of Edinburgh, Edinburgh, UK), Margaret Van Beck (University of Edinburgh, Edinburgh, UK), Naomi R. Wray (Queensland Institute of Medical Research, Queensland, Australia), Stuart Macgregor (Queensland Institute of Medical Research, Queensland, Australia), Peter M. Visscher (Queensland Institute of Medical Research,

Queensland, Australia), Michele T. Pato (University of Southern California, California, USA), Helena Medeiros (University of Southern California, California, USA), Frank Middleton (Upstate Medical University, New York, USA), Celia Carvalho (University of Southern California, California, USA), Christopher Morley (Upstate Medical University, New York, USA), Ayman Fanous (University of Southern California, California, USA and Washington VA Medical Center, Washington, USA and Georgetown University School of Medicine, Washington DC, USA and Virginia Commonwealth University, Virginia, USA), David Conti (University of Southern California, California, USA), James A.Knowles (University of Southern California, California, USA), Carlos Paz Ferreira (Department of Psychiatry, Azores, Portugal), Antonio Macedo (University of Coimbra, Coimbra, Portugal), M. Helena Azevedo (University of Coimbra, Coimbra, Portugal), Carlos N.Pato (University of Southern California, California, USA); Massachusetts General Hospital Jennifer L. Stone (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), Andrew N. Kirby (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), Manuel A. R. Ferreira (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), Mark J. Daly (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), ShaunM. Purcell (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), Jennifer L. Stone (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), Kimberly Chambert (The Broad Institute of Harvard and MIT, Massachusetts, USA), Douglas M. Ruderfer (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA), Finny Kuruvilla (The Broad Institute of Harvard and MIT, Massachusetts, USA), Stacey B. Gabriel (The Broad Institute of Harvard and MIT, Massachusetts, USA), Kristin Ardlie (The Broad Institute of Harvard and MIT, Massachusetts, USA), Jennifer L. Moran (The Broad Institute of Harvard and MIT, Massachusetts, USA), Edward M. Scolnick (The Broad Institute of Harvard and MIT, Massachusetts, USA), Pamela Sklar (Massachusetts General Hospital, Massachusetts, USA and The Broad Institute of Harvard and MIT, Massachusetts, USA).

Conflicts of interest

The authors declare no competing interests.

References

1. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am J Hum Genet* Jan 8; **86**(1): 6-22.

2. Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* 2001 Feb 16; **291**(5507): 1224-1229.

3. Bray NJ, Buckland PR, Owen MJ, O'Donovan MC. Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet* 2003 Jul; **113**(2): 149-153.

4. Lander ES. The new genomics: global views of biology. *Science* 1996 Oct 25; **274**(5287): 536-539.

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al.* Finding the missing heritability of complex diseases. *Nature* 2009 Oct 8; **461**(7265): 747-753.

6. Maher B. Personal genomes: The case of the missing heritability. *Nature* 2008 Nov 6; **456**(7218): 18-21.

7. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009 Aug 6; **460**(7256): 748-752.

8. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D *et al.* Common variants conferring risk of schizophrenia. *Nature* 2009 Aug 6; **460**(7256): 744-747.

9. O'Donovan MC, Craddock NJ, Owen MJ. Genetics of psychosis; insights from views across the genome. *Hum Genet* 2009 Jun 12.

10. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009 Aug 6; **460**(7256): 753-757.

11. Peirce TR, Bray NJ, Williams NM, Norton N, Moskvina V, Preece A *et al.* Convergent evidence for 2',3'-cyclic nucleotide 3'-phosphodiesterase as a possible susceptibility gene for schizophrenia. *Arch Gen Psychiatry* 2006 Jan; **63**(1): 18-24.

12. Bray NJ, Preece A, Williams NM, Moskvina V, Buckland PR, Owen MJ *et al.* Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression. *Hum Mol Genet* 2005 Jul 15; **14**(14): 1947-1954.

13. Law AJ, Lipska BK, Weickert CS, Hyde TM, Straub RE, Hashimoto R *et al.* Neuregulin 1 transcripts are differentially expressed in schizophrenia and regulated by 5' SNPs associated with the disease. *Proc Natl Acad Sci U S A* 2006 Apr 25; **103**(17): 6747-6752.

14. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009 Mar; **10**(3): 184-194.

15. Le-Niculescu H, McFarland MJ, Mamidipalli S, Ogden CA, Kuczenski R, Kurian SM *et al.* Convergent Functional Genomics of bipolar disorder: from animal model pharmacogenomics to human genetics and biomarkers. *Neurosci Biobehav Rev* 2007; **31**(6): 897-903.

16. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010; **6**(4): e1000888.

17. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 2008 Aug; **24**(8): 408-415.

18. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L *et al.* A survey of genetic human cortical gene expression. *Nat Genet* 2007 Dec; **39**(12): 1494-1499.

19. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P *et al.* Genetic control of human brain transcript expression in Alzheimer disease. *Am J Hum Genet* 2009 Apr; **84**(4): 445-458.

20. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 1996; **5**(3): 299-314.

21. Marangos PJ, Schmechel DE. Neuron specific enolase, a clinically useful marker for neurons and neuroendocrine cells. *Annu Rev Neurosci* 1987; **10**: 269-295.

22. Teepker M, Munk K, Mylius V, Haag A, Moller JC, Oertel WH *et al.* Serum concentrations of s100b and NSE in migraine. *Headache* 2009 Feb; **49**(2): 245-252.

23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007; (81).

24. Purcell S. PLINK v1.05.

25. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 2008 May 6; **6**(5): e107.

26. Moskvina V, Ivanov D, Blackwood D, St Clair D, Smith AV, Hultman C *et al.* Genetic differences between four European populations. *Hum Hered* 2010; (In Press).

27. Balding DJ. Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol* 2003 May; **63**(3): 221-230.

28. Nagelkerke NJD. A Note on a General Definition of the Coefficient of Determination. *Biometrika* 1991 Sep; **78**(3): 691-692.

29. Mitchell KJ, Porteous DJ. Rethinking the genetic architecture of schizophrenia. *Psychol Med* 2010 Apr 12: 1-14.

30. Bray NJ, Buckland PR, Williams NM, Williams HJ, Norton N, Owen MJ *et al.* A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain. *Am J Hum Genet* 2003 Jul; **73**(1): 152-161.

31. Southan C. Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics* 2004 Jun; **4**(6): 1712-1726.

Tables

Table 1. Regression of affected status on difference in risk allele score derived from top and bottom eQTL sets.

Trained in	Targeted in	eQTL comparison	Difference in risk allele score	Regression p-value	Mean MAF of top eQTLs	Mean MAF of bottom eQTLs	T-test significance of difference in MAF
Split ISC	Split ISC	Top 50% versus bottom 50%	2.56E-05	0.014	0.227	0.228	0.948
Split ISC	Split ISC	Top 5% versus bottom 50%	8.15E-05	0.014	0.241	0.228	0.001
Split ISC	Split ISC	Top 5% versus bottom 5%	9.63E-05	0.012	0.241	0.246	0.268
MGS	ISC	Top 50% versus bottom 50%	1.63E-05	0.298	0.227	0.228	0.594
MGS	ISC	Top 5% versus bottom 50%	9.27E-05	0.002	0.238	0.228	0.047
MGS	ISC	Top 5% versus bottom 5%	8.57E-05	0.003	0.238	0.249	0.054

Abbreviations: MAF – minor allele frequency. Regression of affected status on difference in risk allele score derived from top and bottom eQTL sets. A positive score in the 'Difference in risk allele score' column indicates that the difference between the top eQTL and bottom eQTL sets is greater in cases than controls.

Table 2. Regression of affected status on risk allele score.

Training dataset	Target dataset	Top / bottom eQTL set	SNP count	Nagelkerke pseudo-R2	Regression p-value	Case/control risk allele score difference
Split ISC	Split ISC	Top 50%	10805	1.59	1.43E-14	5.36E-05
Split ISC	Split ISC	Top 5%	1285	0.47	2.09E-05	1.10E-04
Split ISC	Split ISC	Bottom 50%	10967	0.63	9.22E-07	2.86E-05
Split ISC	Split ISC	Bottom 5%	2033	0.04	0.1122	1.38E-05
MGS	ISC	Top 50%	3903	0.50	8.78E-10	3.08E-05
MGS	ISC	Top 5%	435	0.30	1.47E-06	1.07E-04
MGS	ISC	Bottom 50%	4037	0.30	1.63E-06	1.45E-05
MGS	ISC	Bottom 5%	1154	0.11	0.0027	2.15E-05

Regression of affected status on risk allele score for individual *cis*-eQTL SNP lists. Population of origin was used as a covariate in this regression. Nagelkerke R^2 is a measure of variance in disease state that is explained by the risk score. SNP count is the number of SNPs in the set.

Supplementary table 1 (see Supplementary_Table_1.xls)

Regression of affected status on difference in risk allele score derived from top and bottom eQTL sets matched for F_{ST}. A positive score in the 'Difference in risk allele score' column indicates that the difference between the top eQTL and bottom eQTL sets is greater in cases than controls.

Supplementary table 2 (see Supplementary_Table_2.xls)

Pruned set of SNPs comprising those that were associated in the MGS training set at $P<0.5$ that were both within the top 5% of eQTLs and for which the allele designated as the risk allele in the MGS sample was associated in the ISC sample at a nominally significant level ($P<0.05$). We also provide the estimated OR as it applies to the allele designated allele 1.