Methods for brain disease genetics using gene expression data of the healthy brain

Huisman, Sjoerd

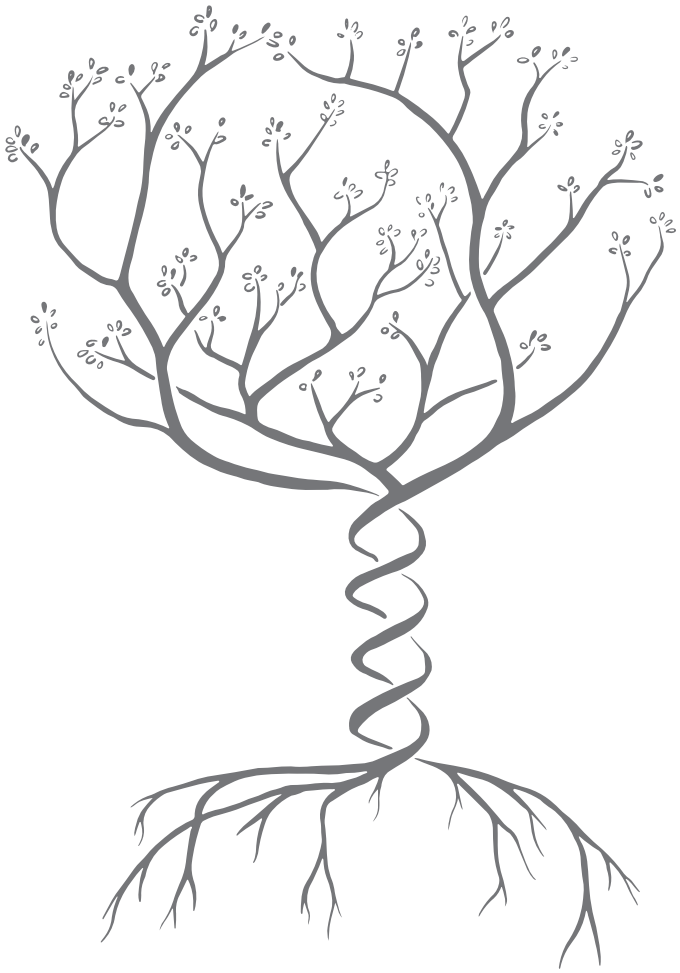
DOI
10.4233/uuid:ace78c36-d0a3-40d7-bb50-505bce956042

Publication date
2020

Document Version
Final published version

Citation (APA)
Huisman, S. (2020). *Methods for brain disease genetics using gene expression data of the healthy brain*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ace78c36-d0a3-40d7-bb50-505bce956042

Methods for brain disease genetics using gene expression data of the healthy brain



Sjoerd Huisman

Methods for brain disease genetics using gene expression data of the healthy brain

Sjoerd Huisman

Methods for brain disease genetics using gene expression data of the healthy brain

Dissertation
for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus
prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Wednesday 3 June 2020 at 12:30 o'clock

by

Sjoerd Maarten Helena HUISMAN

Master of Science in Mathematics, Leiden University, the Netherlands
born in Bergen (L), the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr.ir. M.J.T. Reinders	Delft University of Technology, promotor
Prof.dr.ir. B.P.F. Lelieveldt	Delft University of Technology Leiden University, promotor

Independent members:

Prof.dr. P.J.F. Groenen	Erasmus University Rotterdam
Prof.dr. P.-B.A.C. 't Hoen	Radboud University
Prof.dr. D. Posthuma	Vrije Universiteit Amsterdam
Prof.dr. J.J. Goeman	Leiden University
Prof.dr. L.F.A. Wessels	Delft University of Technology
Prof.dr. R.C.H.J. van Ham	Delft University of Technology, reserve member

The research described in this thesis was financed by the Dutch Technology Foundation STW, as part of the STW project 12721 ("Genes in Space") under the IMAGENE perspective program.

Huisman, Sjoerd M.H.
Methods for brain disease genetics using gene expression data of the healthy brain

Cover drawing by Anneke Kikkert
Printed by ProefschriftMaken || www.proefschriftmaken.nl

ISBN 978-94-6380-844-6

Contents

1 General introduction **7**
1.1 Data types 9
1.2 Methods 10

2 Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas **15**
2.1 Introduction 16
2.2 Results 19
2.3 Discussion 27
2.4 Materials and Methods 31
2.5 Conflict of interest 34
2.6 Funding 35

3 BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome **37**
3.1 Introduction 38
3.2 Materials and methods 42
3.3 Results 44
3.4 Discussion 55
3.5 Acknowledgements 57

4 A structural equation model for imaging genetics using spatial transcriptomics **59**
4.1 Introduction 60
4.2 Materials and methods 62

4.3 Results 71
4.4 Conclusion 76

5 Imaging genetics for schizophrenia using transcriptome informed clustering 79
5.1 Introduction 79
5.2 Methods 81
5.3 Results 87
5.4 Discussion 100
5.5 Supplement 103

6 General discussion 109
6.1 Generating and testing hypotheses 110
6.2 The brain 114
6.3 The genetics 115
6.4 The future 116

References 119

Summary 133

Samenvatting 135

Acknowledgements 139

List of publications 141

— 1 —

General introduction

Understanding how the human brain works is one of the biggest challenges of the 21st century. This quest not only involves researchers in many scientific fields, but also vast numbers of measurements on the brain. These measurements have and will be done in animal models and in humans, both in a healthy state and when things go wrong. At the moment, much research focuses on the morphology and electrophysiology of brain cells, to get a grasp of the cellular interactions that lead to a working mind. In this thesis, however, we focus on the molecular level of genes and their activation patterns in the brain.

The activities of genes have been measured on post-mortem brains using micro-arrays and RNA sequencing. Tissues for these studies have been obtained from general or brain specific biobanks in the United States (Lonsdale et al., 2013), Sweden (Sjöstedt et al., 2015) and the United Kingdom (Trabzuni et al., 2011). However, some of the most comprehensive databases of gene expression in the brain have been made by the Allen Institute for Brain Science in Seattle. In 2003 the Allen Institute set out to provide the world with molecular data of the brain. They produced gene expression atlases: overviews of anatomically and spatially labelled gene expression in the healthy brain. The data is presented in a public data portal (www.brain-map.org), and can be downloaded freely. The portal contains the genome-wide adult mouse atlas, obtained with a high-resolution in situ hybridisation technique (Lein et al., 2007). For a subset of 2000 genes, the same technique was applied to prenatal and postnatal mouse brains in seven stages of development (Thompson et al., 2014). Although these data have been used to study human brain function and disease (Vied et al., 2014), mouse brain anatomy strongly

differs from that of humans. A more closely related model organism is the rhesus macaque. The non-human primate atlas includes in situ hybridisation data for subsets of genes and brain regions in the developing macaque brain and it includes genome-wide expression measurements of dissected prefrontal cortex, visual cortex, hippocampus, amygdala and striatum samples in the same developmental stages (Bakken et al., 2016). Samples of human brains are more difficult to obtain than those of mouse and macaque. Nevertheless, the Allen Institute provides two human brain gene expression atlases. The developmental human atlas contains measurements of 16 dissected regions in 42 brains ranging in age from 8 weeks post conception to 40 years after birth (Miller et al., 2014). Finally, for the adult human brain the Allen Institute provides genome-wide gene expression with a much higher spatial resolution (Hawrylycz et al., 2012). This atlas contains data from 6 carefully screened healthy brains from which in total 3702 samples were dissected. The exact sampling differs per brain, but 105 anatomical regions were sampled at least once in each of the brains.

The Allen brain atlases have been used to study transcriptional activities in the healthy brain, for which they provide several types of information. In the first place, they show where each gene is expressed, which can tell us about the importance of individual genes for specific brain functions. Secondly, the atlases provide spatial co-expression information. Genes that are expressed together (are active in the same brain regions) may be involved in the same processes. We will see in Chapter 3 of this thesis that groups of co-expressed genes are often involved together in known biological pathways, have similar molecular functions, or can be linked to specific cellular components or cell types. In addition to the information about genes of interest, the gene expression atlases contain information about anatomical brain regions. Which genes are active in a region is strongly indicative of the locations samples were taken from, and anatomical similarities are reflected in transcriptional similarities (Mahfouz et al., 2015). These observations can be explained at least partly by the cell-type composition of brain regions (Grange et al., 2014), since gene expression is highly cell-type specific (Darmanis et al., 2015).

The samples in the Allen atlases and many of the other databases were obtained from healthy brains. However, that does not mean the data is of no use for disease studies. The information about gene-gene similarities and region-region similarities can be used to inform models on brain-disease data. Brain

associated diseases such as Alzheimer's disease (Gaiteri, Mostafavi, Honey, De Jager & Bennett, 2016), migraine (Schu & Lrp, 2012), and several psychiatric disorders (Consortium et al., 2013) have a high heritability. This means their occurrence in human populations can be linked to genetic variation. The genetic variants associated with complex diseases are often located in regulatory elements that impact gene expression (Roadmap Epigenomics Consortium et al., 2015). Both gene expression (Lonsdale et al., 2013) and regulatory elements can be highly tissue-specific (Roadmap Epigenomics Consortium et al., 2015). As a result, we pose that gene-gene similarities and region-region similarities obtained from brain specific gene expression data provides valuable information for genetic studies in brain specific diseases. So we arrive at an over-arching topic of this thesis: *how to improve brain disease studies with prior knowledge about gene expression in the healthy brain.*

1.1 Data types

A common goal in disease studies is to find genetic variations that have an effect on some outcome (a phenotype). These studies started in the 19th century, when the general principals of inheritance were described by Gregor Mendel (Abbott & Fairbanks, 2016). Ronald Fisher and his contemporaries combined the ideas about inheritance with statistical models to start the field of population genetics (R. A. Fisher, 1932), in which genetic variation is mathematically linked to variation in phenotypes on a population level. Thomas Hunt Morgan's research on linkage led to the first studies that linked phenotypic variation to specific parts of chromosomes (Morgan, Sturtevant, Muller & Bridges, 1915), and after the discovery of the structure of DNA it was understood how changes in DNA can impact cells through transcription into RNA and translation into protein. After the advent of sequencing in the 1970's and the completion of the human genome in 2004 (International Human Genome Sequencing Consortium, 2004), phenotypic variation could be linked to variation in specific nucleotides in the chromosome. In genome wide association studies (GWAS) variants across the full genome are screened for these associations. However, observing the association between genetic variation and a phenotype is not enough to understand the molecular causes of this phenotype. To get some understanding, variations are often linked to genes, and these genes to functions. Now, if we return to our topic of brain

disease, we can see how spatial gene expression data can help us in that last step. Genome wide expression in the brain can inform us about gene function in the brain.

We have not yet addressed the nature of the phenotypes. These can be a simple label (disease vs. healthy) as in Chapter 2, but also a quantitative measurement like body height. A relevant type of measurement in brain disease is obtained by brain imaging techniques such as magnetic resonance imaging (MRI). A structural MR scan produces a three-dimensional image reflecting mainly the ratio between fat and water content across the brain tissue. By anatomically labelling structures in these images (segmentation) we can derive relevant measurements, for instance the volumes of specific brain regions. Linking such imaging features to genetic variation is the topic of the field of imaging genetics. In many ways, imaging genetics is no different from other genetic research, but it presents the issue of a high-dimensional outcome (instead of a single measurement). These studies have to deal not only with many genetic features, but also a large number of imaging features. In Chapters 4 and 5, we will deal with this challenge. To do this, we again make use of the gene expression data in the healthy brain, specifically the region-region similarity information that this data provides.

1.2 Methods

Now we have considered the type of data we have at out disposal, we will focus on methodological considerations. Modern molecular technology can provide genetic variation data for the whole genome, with several million variants measured in a regular GWAS. In a common association testing scheme, each of these variants is tested for association with a phenotype. As is often the case in empirical research, we perform our analyses on a sample of individuals and generalise our results to the population. Due to the randomness this introduces in the results of the analyses, performing a large number of tests can yield a large number of false positive test outcomes. This should be prevented using a (multiple testing) correction procedure. Commonly used corrections (Goeman & Solari, 2014) reduce the number of false positives, but they make it hard to find variants of interest.

One way to deal with the challenges that come with multiple testing correction, is to increase the sample size. However, this can be expensive or

practically impossible, as it would be for very rare diseases. A statistical solution to the problem is to use prior knowledge. In the most simple way, we can restrict the number of variations we consider to a candidate set. In Chapter 2 we use the spatial gene expression data of the brain to identify these candidate genes (in local co-expression networks), and in Chapter 4 we use external association study results to narrow down the set of variants that are considered. An alternative approach to reduce the testing burden, is to group variants first, and then look for groups of interest. In a second method of Chapter 2 we split up the genome in modules of genes with a similar expression in the healthy brain and try to find modules that are enriched in phenotype associated variants. Chapter 5 is similar in that respect, but here the modules of genes are not considered fixed, but are sampled from a distribution. In these examples, prior knowledge about gene expression in the brain is incorporated in the methodology.

Most of the data analyses presented in this thesis combine data of different types, such as gene expression measurements and genetic variations. This means our methods are examples of data integration techniques. There are several ways to integrate variation and expression data, and to relate both to phenotypic traits (Gusev et al., 2016). In many applications the different types of data can be integrated either by a simple concatenation, by first performing some transformation and then concatenating, or combing them directly in a model (Kim, 2015). However, for these approaches to work, the measurements have to be obtained from the same individuals. We use gene expression measurements of a small number of healthy individuals to inform models on a disease study population, as illustrated in Chapters 2, 4 and 5. The approaches in these chapters therefore use the expression data on the healthy individuals to calculate gene-gene and region-region similarities, which are used as an input to the models fitted on the disease population data.

The explosion in data availability in molecular biology has led to an increase in data driven research (Wang, Zhang & Chen, 2018). Data driven research is exploratory, and can be contrasted with more traditional hypothesis driven confirmatory research. A hypothesis driven study starts with a specific hypothesis about a process, and is followed by data collection to test it. But molecular techniques now allow for collecting data on variations in the whole genome, or on the activity of all genes. A GWAS is somewhat exploratory by definition, since it tries to answer the question what could cause

a phenotype. It could be considered a search for a suitable hypothesis or, alternatively, a test of a very large set of hypotheses simultaneously. If we consider it a search for a hypothesis, this means a (hypothesis driven) validation experiment is needed. If we consider a GWAS to be a test of a large set of hypotheses, we return to the issue of the multiple testing problem. If we report all statistically significant results without any correction, many of those will be false positives. A wide range of correction methods is used to solve this (Goeman & Solari, 2014).

Both for exploratory and more hypothesis driven research two distinct statistical paradigms are available. The most common one uses frequentist inference. In frequentist inference, data is commonly considered to be a result of a random process, while this underlying process is considered fixed. A frequentist test makes use of a precisely specified (parametrised) model for this process. Now, if the observed data is not in line with the proposed hypothesis, this hypothesis can be rejected. The somewhat philosophical basis for this is the idea that a hypothesis is either false or true, and the process that it describes will in the long run give data that follows the probability distribution of this process. Chapters 2, 3 and 4 contain frequentist tests. The second paradigm is based on Bayesian inference. Bayesian inference results from a different view of probability, which is often termed subjective. The idea is that we usually do not want to make statements about a frequency with which something occurs, but rather about our belief in a hypothesis. We have ideas about the truth and use data to adjust those beliefs. As a result, we can talk about probabilities for hypotheses, by attaching probability distributions to model parameters. To calculate these probabilities in the data analysis, one needs to define what they were before the data was observed: the prior probabilities. Often prior distributions for the parameters are picked to be uninformative (Gelman, 2008). However, we can also make use of this property of Bayesian inference by incorporating information from external data into our models. In Chapter 5 we use the gene expression data of the healthy human brain to define a prior distribution for an imaging genetics model for schizophrenia.

Regardless of the type of research and statistical testing, to understand data we need to visualise it. In the simplest form, this can be done by making scatter plots of continuous measurements or boxplots of variables measured in groups of samples. Each of the studies that form a part of this thesis makes

use of these simple data visualisations, but Chapter 3 is all about visualisation. The gene expression data of the Allen Brain Atlas contains a wealth of information that is used in each chapter of the thesis. We saw a need to get a visual overview of this data. Our visualisation has helped us to quickly answer small questions in our research and will hopefully spawn new hypotheses for other researchers.

Throughout the chapters of this thesis, we look at brain disease data, and try to go beyond the initial data analysis. By using measurements of gene expression across the healthy human brain, we extend the data interpretation to coherent groups of genes and brain regions of interest. And, in this process, we explore methodology for data integration and face the challenges in multiple testing correction.

— 2 —

Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas

Abstract

Migraine is a common disabling neurovascular brain disorder typically characterised by attacks of severe headache and associated with autonomic and neurological symptoms. Migraine is caused by an interplay of genetic and environmental factors. Genome-wide association studies (GWAS) have identified over a dozen genetic loci associated with migraine. Here, we integrated migraine GWAS data with high-resolution spatial gene expression data of normal adult brains from the Allen Human Brain Atlas to identify specific brain regions and molecular pathways that are possibly involved in migraine pathophysiology. To this end, we used two complementary methods. In GWAS data from 23,285 migraine cases and 95,425 controls, we first studied modules of co-expressed genes that were calculated based on human brain expression data for enrichment of genes that showed association with migraine. Enrichment of a migraine GWAS signal was found for five modules that suggest involvement in migraine pathophysiology of: i) neurotransmission, protein catabolism and mitochondria in the cortex; ii) transcription regulation in the cortex and cerebellum; and iii) oligodendrocytes and mitochondria in

This chapter has been published as: Eising, E., Huisman, S. M. H., Mahfouz, A., Vijfhuizen, L. S., Anttila, V., Winsvold, B. S., ... Reinders, M. J. T. (2016). Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas. *Human Genetics, 135*(4), 425–439. https://doi.org/10.1007/s00439-016-1638-x
All supplemental materials can be found in the online publication.

subcortical areas. Second, we used the high-confidence genes from the migraine GWAS as a basis to construct local migraine-related co-expression gene networks. Signatures of all brain regions and pathways that were prominent in the first method also surfaced in the second method, thus providing support that these brain regions and pathways are indeed involved in migraine pathophysiology.

2.1 Introduction

Migraine is a common neurovascular brain disorder characterised by attacks of severe, unilateral headache, often accompanied by nausea and phono- and photophobia (Headache Classification Committee of the International Headache Society (IHS), 2013). Two main migraine types are distinguished based on the presence or absence of an aura, which consists of transient neurologic symptoms including visual and sensory disturbances that can precede attacks in up to one-third of patients. Migraine is a complex genetic disorder with an estimated heritability of approximately 50% (Mulder et al., 2003) and thought to be caused by an interplay of multiple genetic variants, each with a small effect size, and environmental factors. Numerous candidate gene association studies have been performed for migraine, however, their value turned out rather low as none could be replicated in a large genome-wide marker dataset of thousands of migraine patients and controls (de Vries et al., 2016). Genome-wide association studies (GWAS) investigating the common forms of migraine have identified 13 disease susceptibility loci (Anttila et al., 2010, 2013; Chasman et al., 2011; Freilinger et al., 2012). These loci identified genes that are involved in glutamatergic neurotransmission (*MTDH, LRP1, MEF2D*), neuron and synapse development (MEF2D, *ASTN2*, *PRDM16*, *FHL5*, *PHACTR1*, *TGFBR2* and *MMP16*), brain vasculature (*PHACTR1*, *TGFBR2*, *C7orf10*), extracellular matrix (*MMP16*, *TSPAN2*, *AJAP1*), and pain-sensing (TRPM8). These findings support knowledge that came from investigating disease mechanisms in monogenic migraine-related disorders including familial hemiplegic migraine (*FHM*), a monogenic subtype of migraine with aura (Ferrari, Klever, Terwindt, Ayata & van den Maagdenberg, 2015; Tolner et al., 2015). Notably, transgenic knock-in (KI) mouse models that express human pathogenic FHM1 (A. M. van den Maagdenberg et al., 2004; A. M. J. M. van den Maagdenberg et al., 2010) or FHM2 (Leo et al., 2011) mutations revealed increased suscept-

ibility for experimentally induced cortical spreading depression (CSD), the electrophysiological correlate of the migraine aura (Lauritzen, 1994), which could be directly linked to increased cortical glutamatergic neurotransmission in FHM1 KI mice (Tottene et al., 2009). Other monogenic disorders in which migraine is prevalent are cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) and retinal vasculopathy with cerebral leukodystrophy (RVCL) that indicate a role for dysfunction of the brain vasculature in migraine (Tolner et al., 2015). Migraine genes identified by GWAS are primarily identified based on their location near top hits, so true causality of (at least some of) them remains uncertain, which is not different from other disorders. Furthermore, current GWAS top hits explain only a small part of the disease heritability, and, therefore, genes identified in this way reflect only a fraction of the pathways conferring genetic disease risk. Hence, pathway analysis methods that harvest a larger portion of the GWAS data (i.e. not only loci with significant P-values) may give more valuable insight into disease genetics, as has been tried for other diseases (Atias, Istrail & Sharan, 2013; Sun, 2012).

Commonly used tools to explore disease-associated pathways in GWAS data make use of functional enrichments (MAGENTA Gene Set Enrichment Analysis (Segrè, Groop, Mootha, Daly & Altshuler, 2010)), protein interactions (DAPPLE (Rossin et al., 2011)) or text-mining (GRAIL (Raychaudhuri et al., 2009)), but did not successfully identify overrepresented molecular pathways involved in migraine (Anttila et al., 2013). One explanation why it may be difficult to confidently identify disease pathways from GWAS data is that loci often contain multiple genes, of which only (one or) a subset might influence the trait of interest. Moreover, each of these genes can be expressed in multiple cell types and may have different functions in each of them. We envisaged that gene expression data can be used to preselect genes for functional analysis based on their expression in disease-relevant tissues, thereby increasing the chance of identifying disease-relevant genes and pathways. In addition, gene co-expression analysis can be used to identify genes with a similar expression patterns. Previous studies have shown that gene co-expression can infer a wide range of meaningful biological information, e.g. shared gene functions, biological pathways or cell type-specific expression (Grange et al., 2014; Hawrylycz et al., 2012; Kang et al., 2011).

Gene co-expression analysis has been applied successfully to identify disease

mechanisms from GWAS or other genomics data for other disorders, including allergic rhinitis and autism spectrum disorder (Ben-David & Shifman, 2012; Bunyavanich et al., 2014; Parikshak et al., 2013; Willsey et al., 2013). Admittedly, these studies benefited from having available gene expression data obtained under disease-specific conditions (Bunyavanich et al., 2014) or the use of causal genetic variants with large effect sizes (Ben-David & Shifman, 2012; Parikshak et al., 2013; Willsey et al., 2013). For migraine, no gene expression data from disease-conditions are available. A few gene expression profiling studies have been carried out for migraine, i.e. in whole blood of episodic and chronic migraine patients (A. D. Hershey et al., 2004) and menstrual migraine patients (A. Hershey, Horn, Kabbouche, O'Brien & Powers, 2012), in immortalized cell lines of migraine with aura patients (Nagata et al., 2009), and in brain material of transgenic KI FHM1 mice (de Vries et al., 2014), but no overlapping deregulated genes or pathways have been identified. Nor is there a large set of causal genes, except for three genes (*CACNA1A*, *ATP1A2* and *SCN1A*) (De Fusco et al., 2003; Dichgans et al., 2005; Ophoff et al., 1996) that have been identified for FHM, that can guide gene identification efforts in the common forms of migraine. Therefore, we focused our analyses on gene expression data from the normal human brain.

Here we used two complementary methods to connect gene expression data from adult human brain, the most relevant tissue for migraine, with GWAS data in order to identify migraine-related pathways. To this end, spatially-mapped gene expression data of the adult human brain, obtained from the Allen Human Brain Atlas (Hawrylycz et al., 2012), was used to calculate brain-specific co-expression levels between genes. We used GWAS data, available through the International Headache Genetics Consortium, of 23,285 migraine cases and 95,425 population-matched controls (Anttila et al., 2013) to calculate gene-based associations with migraine. This enabled the inclusion of below-threshold association signals that did not reach genome-wide significance (P-value $< 5 \cdot 10^{-8}$) due to lack of power (Gibson, 2012; Mooney, Nigg, McWeeney & Wilmot, 2014). For our first method, we grouped all genes into co-expression modules and studied the enrichment of genes with nominally significant gene-based associations with migraine in the different modules. For our second method, we constructed local co-expression networks around 'high-confidence genes' (i.e. those genes with gene-based P-values that survived multiple testing correction) that we combined into a local migraine-related

co-expression gene network. By studying the modules enriched for migraine-associated genes (method 1) and the local migraine-related co-expression gene networks (method 2), we identified multiple brain regions, cell types and pathways overlapping between the two methods that are possibly involved in migraine pathophysiology.

2.2 Results

2.2.1 Spatial co-expression network of the adult human brain

To identify brain regions and pathways involved in migraine pathophysiology, we performed co-expression network analysis using spatial gene expression information of the Allen Human Brain Atlas (Hawrylycz et al., 2012). We focused on the adult human brain transcriptome, since migraine is a brain-related disorder that affects mostly the adult population. Microarray data were available from six healthy adult human brains; five males and one female, aged 24 to 57 with a mean age of 42 years, each dissected into 363 to 946 samples (3,702 in total) from well-defined brain regions. We used the gene expression data of 29,374 microarray probes that could be mapped unambiguously to 19,972 genes. Gene co-expression levels were calculated separately for each brain (across the samples), and subsequently averaged (per gene) to obtain a single spatial co-expression network not affected by individual brain differences (see Material and Methods). Note that these levels, therefore, reflect brain-wide spatial co-expression. Differences in expression values between the female brain and five male brains were not more pronounced than the differences between any of the male brains and all other brains (see Supplemental Materials and Methods; Figure S1), justifying the unbalanced gender composition of the Allen Brain Atlas for our analyses. In fact a recent publication by (Hawrylycz, Sunkin & Ng, 2015) showed that functionally relevant genes seem to have a stable expression across the six donors. Using hierarchical clustering analysis, we identified 18 modules in the spatial brain-wide co-expression network, with module sizes varying from 179 to 2,007 genes (Figure 2.1). Each module thus contains genes that have similar expression patterns across the different brain samples. Clustering the gene expression data can be done in various ways (see Supplemental Material and Methods). The final clustering

tree showed strongest enrichment for migraine genes. Modules enriched for migraine genes are further investigated for these spatial patterns across brain regions and for functional enrichments of the migraine genes.

2.2.2 Genes associated with migraine

We used summary statistics data from the GWAS meta-analysis for migraine (Anttila et al., 2013) performed by the International Headache Genetics Consortium to calculate gene-based P-values for the association with migraine. The 2,116 genes with nominal gene-based P-values below 0.05 were considered to have a potential link to migraine and are therefore referred to as migraine 'candidate genes'. The 14 genome-wide significant genes, with multiple testing corrected gene-based P-values below 0.05, are referred to as 'high-confidence genes'. The high-confidence gene set contained 10 genes located at or near genome-wide significant GWAS loci: *ASTN2*, *C7orf10*, *FHL5*, *MEF2D*, *TRPM8*, *LRP1*, *STAT6*, *NAB2*, *PRDM16* and *UFL1* (Anttila et al., 2013). *LRP1*, *STAT6* and *NAB2* at chromosome 12q13 share the same genome-wide significant SNP, and the top SNPs for *FHL5* and *UFL1* at chromosome 6q16 are in strong linkage disequilibrium (LD). The remaining high-confidence genes *LEPROTL1*, *DCLRE1C*, *SUV39H2*, and *MBOAT4* are located near SNPs that did not reach the level of genome-wide significance in the migraine GWAS, and gain from a reduced multiple testing burden in our gene-based analysis compared to a SNP-based analysis. GWAS hits *MTDH*, *PHACTR1*, *TGFBR2*, *MMP16*, *TSPAN2* and *AJAP1* did not reach a multiple testing corrected gene-based P-value below 0.05, possibly due to a larger distance between the GWAS locus and the gene, and were therefore not designated as high-confidence gene.

2.2.3 Migraine-associated loci converge into five co-expression modules

We performed an enrichment analysis of the 2,116 migraine candidate genes in the 18 co-expression modules to identify the modules that have the strongest link with migraine. Five modules labelled A to E showed enrichment of candidate genes in a Fisher exact test ($P < 0.05$) (Figure 2.1; Table S1). To verify that the identified enrichments were not the result of bias in the Fisher exact test introduced by LD between SNPs in the GWAS data and by SNPs

assigned to multiple genes, we performed a second, LD-corrected Fisher exact test. These results confirm the association of modules A – E with migraine (Table S1).

Module A showed the highest enrichment of migraine candidate genes (enrichment $P = 9.44 \cdot 10^{-4}$, LD-corrected enrichment $P = 5.47 \cdot 10^{-4}$) and contains 1,556 genes with high expression in cerebral cortex, very low expression in cerebellum, and low expression in hippocampal formation and subcortical cerebrum (Figure 2.2). Module B (enrichment P = 0.015, LD-corrected enrichment $P = 7.18 \cdot 10^{-3}$) consists of 1,595 genes with high expression in cerebellum, low expression in subcortical regions and an intermediate expression in cerebral cortex (Figure 2.2). Module C (enrichment P = 0.02, LD-corrected enrichment $P = 7.77 \cdot 10^{-3}$) contains only 497 genes. Genes from module C have an expression pattern similar to that of module A with higher expression in hippocampal formation and claustrum (Figure 2.2). Module D (enrichment P = 0.024, LD-corrected enrichment $P = 5.82 \cdot 10^{-3}$) is the largest module with 1,984 genes that are preferentially expressed in subcortical regions and the white matter, with low expression in cerebellar and cerebral cortex (Figure 2.2). Module E (enrichment P = 0.03, LD-corrected enrichment P = 0.04) contains only 179 genes with high expression in cerebellar cortex, pons and hypothalamus (Figure 2.2).

2.2.4 Migraine-associated modules show enrichment of functions involved in neurotransmission, mitochondria, gene expression regulation and oligodendrocytes

Next, we performed a functional enrichment analysis of modules A – E to identify gene functions associated with migraine pathophysiology (Figure 2.2; Tables S2-S5). We studied pathways from KEGG, Reactome and PANTHER, and gene ontology (GO) terms from PANTHER and the GO FAT database using the Functional Annotation Clustering tool in DAVID. GO term and pathway groups were considered significant when Benjamini-corrected P-value was below 0.05 (reflected in an EASE score of 1.3 or higher). Functions enriched in module A included energy metabolism, protein catabolism and synaptic functions (Table S2). Genes in module B showed enrichment of multiple functions all involved in gene expression regulation (Table S3). Module C contains

a large set of genes involved in purine nucleotide binding, and also showed enrichment for several brain developmental and synaptic functions (Table S4). Genes in module D showed highest enrichment of functions involving energy supply, apoptosis and myelination (Table S5). Module E did not show any significant functional enrichments. Most enriched functions are module-specific; of modules A-D only module C shares most of its enriched functions with other modules (A, L and P) (Figure S2).

2.2.5 Enrichment of oligodendrocytic and neuronal genes in migraine-associated modules

Expression patterns in the brain are co-determined by cell type composition (Grange et al., 2014; Hawrylycz et al., 2012). Consequently, we expected to find enrichment of cell type-specific genes in the co-expression modules (Figure 2.1). Notably, modules A and C showed significant enrichment of genes specifically expressed in neurons (119 genes, $P = 8.00 \cdot 10^{-15}$; 40 genes $P = 3.12 \cdot 10^{-6}$, respectively), which is in line with the preferential expression in cerebral cortex of genes in this module and the enrichment for synaptic functions. Module D is significantly enriched for oligodendrocyte-specific genes (103 genes, $P = 1.37 \cdot 10^{-55}$), and also showed enrichment for genes specifically expressed in microglia and endothelial cells. This finding seems well in line with the observed high expression in white matter of genes in this module and the enrichment of several functions related to myelination. Module E is enriched for neuron-specific genes (18 genes, $P = 1.09 \cdot 10^{-4}$). Module B did not show enrichment of cell type-specific genes.

2.2.6 Confirmation of the association of modules A – D with migraine using a local seed network

The association of modules A – E with migraine may be the result of low migraine association signals, and may therefore not have a direct link to the genome-wide significant GWAS loci, as only module B (*LRP1*) and module D (*UFL1*) contain a high-confidence gene (Figure 2.1). To leverage the information in the high-confidence genes, we used them as seeds for a local co-expression network. The local co-expression network therefore contains only the high-confidence genes and their co-expression partners (Figure 2.3).

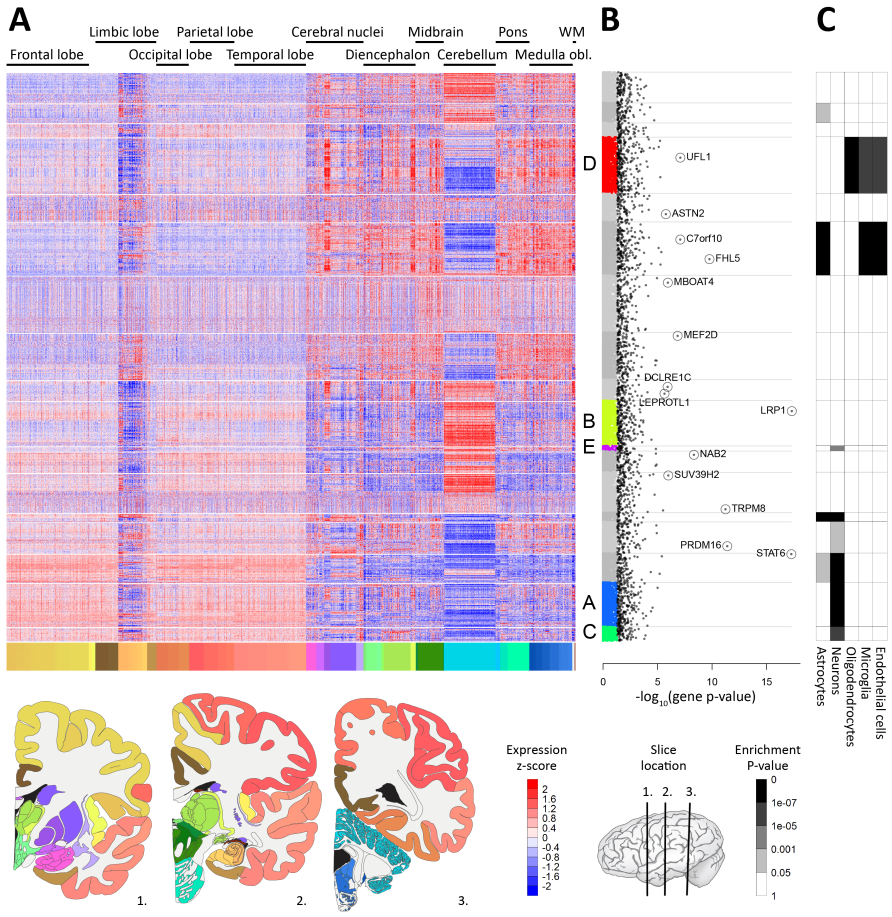


Figure 2.1: ***(A)*** *Heat map of the clustered gene expression data, with the 3,702 concatenated human brain samples in columns and the 19,972 genes in rows, ordered according to their clustering. The brain samples are ordered based on their location in the brain, which is noted above the heat map and illustrated with the colour coding from the Allen Brain Institute below the heat map. The colour coding is also illustrated in the three coronal brain sections below the heat map (for brain region names in the coronal sections, see Figure S3). Low expression is shown as blue, high expression is shown as red. The genes are clustered into 18 modules, here separated by white rows.* ***(B)*** *Log-transformed gene-based P-values for the association with migraine are shown for all genes with: 1) genes with P-values below 0.05 in the colour corresponding to modules A – E or in grey for the other modules; 2) migraine candidate genes in black; and 3) high-confidence genes circled and named. Gene modules A – E are the five modules enriched for candidate genes.* ***(C)*** *the table shows the enrichment of cell type-specific genes in the 18 modules from white (P-value* > 0.05*) to black (P-value* $< 10^{-7}$*).*

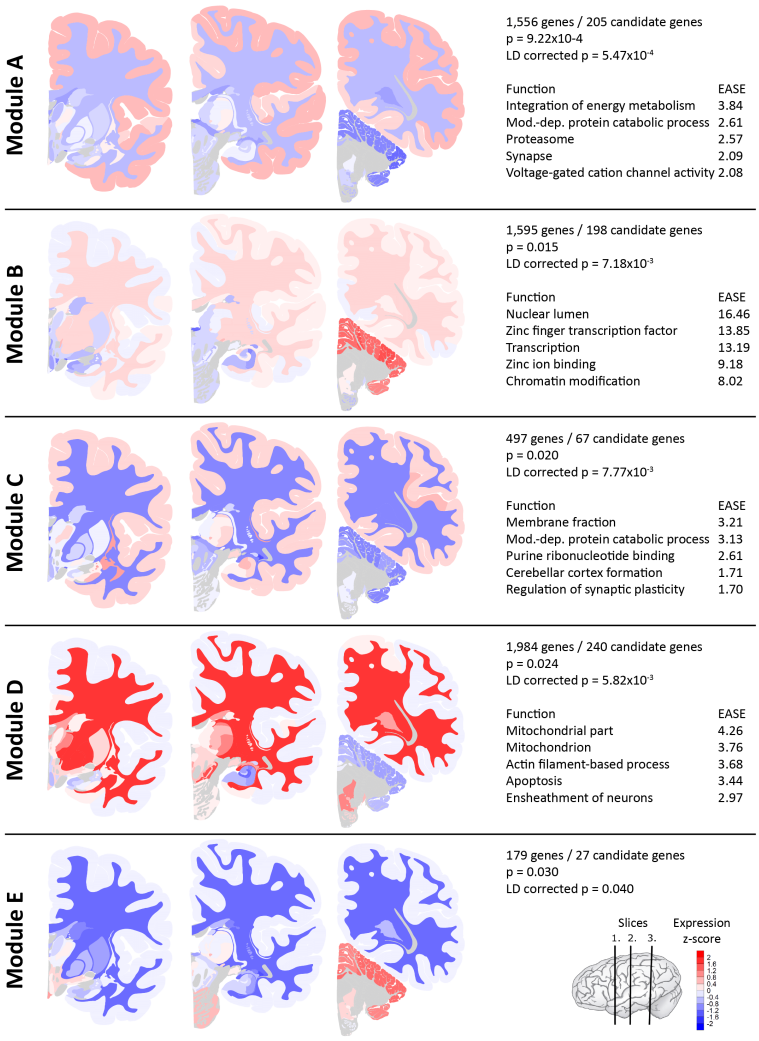


Figure 2.2: Average gene expression levels are shown for each module from blue (low) to red (high) in the different brain regions represented in the three coronal brain sections (for brain region names in the coronal sections, see Figure S3). Regions that lack gene expression information are depicted in grey. The lists on the right show: 1) the numbers of genes and migraine candidate genes; 2) the P-values for the enrichment of migraine candidate genes; and 3) the top 5 enriched functions in each module, as identified using the Functional Annotation Clustering tool in DAVID, with their corresponding EASE score. The EASE score is the geometric mean of the Benjamini-corrected negative log (base 10) P-values of its pathways and GO terms, so a score above 1.3 corresponds to a Benjamini-corrected P-value below 0.05. Module E has no significant functional enrichments.

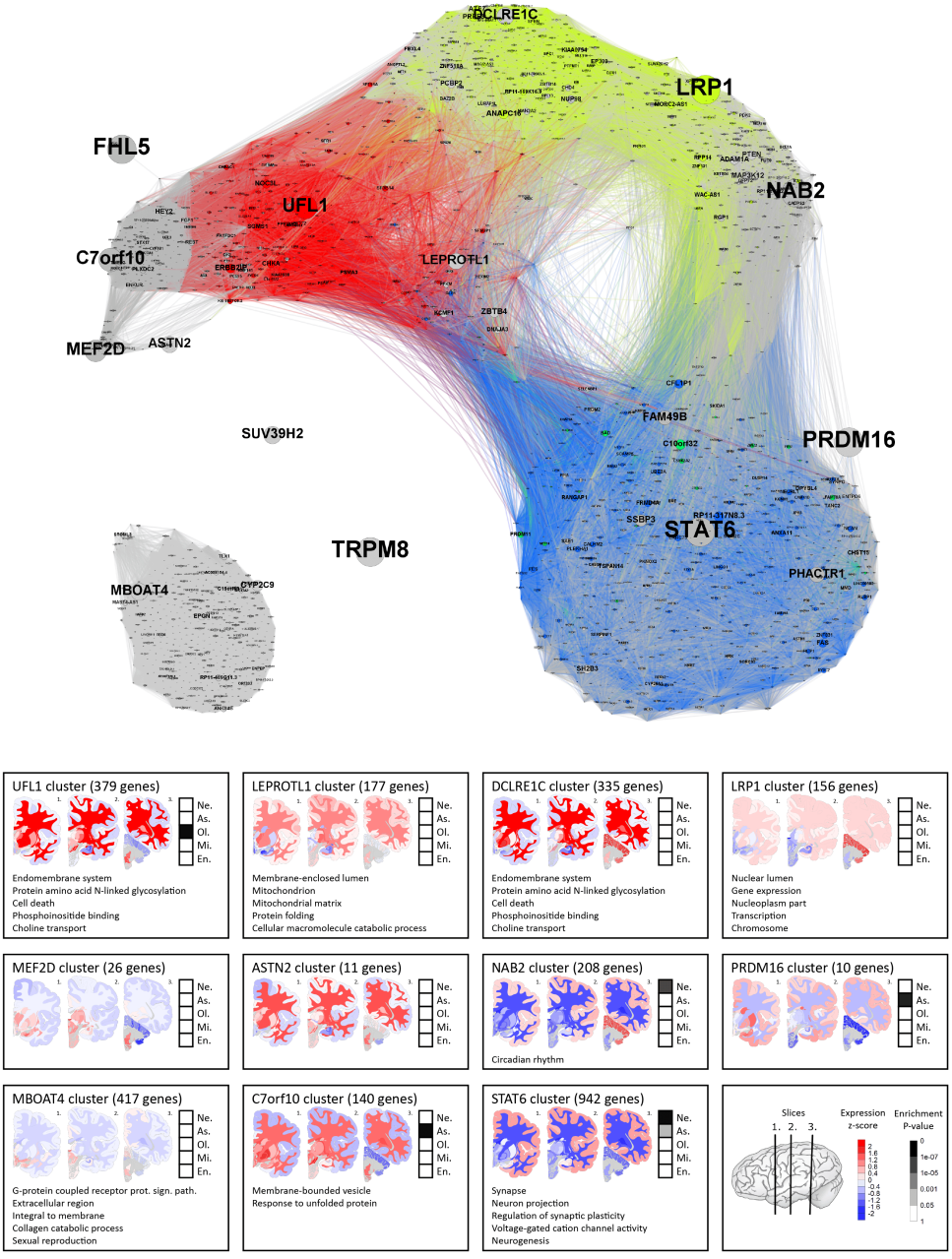

DCLRE1C
LRP1
FHL5
NAB2
UFL1
C7orf10
LEPROTL1
MEF2D
ASTN2
SUV39H2
FAM49B
PRDM16
TRPM8
STAT6
MBOAT4
PHACTR1
UFL1 cluster (379 genes)
Endomembrane system
Protein amino acid N-linked glycosylation
Cell death
Phosphoinositide binding
Choline transport
LEPROTL1 cluster (177 genes)
Membrane-enclosed lumen
Mitochondrion
Mitochondrial matrix
Protein folding
Cellular macromolecule catabolic process
DCLRE1C cluster (335 genes)
Endomembrane system
Protein amino acid N-linked glycosylation
Cell death
Phosphoinositide binding
Choline transport
LRP1 cluster (156 genes)
Nuclear lumen
Gene expression
Nucleoplasm part
Transcription
Chromosome
MEF2D cluster (26 genes)
ASTN2 cluster (11 genes)
NAB2 cluster (208 genes)
Circadian rhythm
PRDM16 cluster (10 genes)
MBOAT4 cluster (417 genes)
G-protein coupled receptor prot. sign. path.
Extracellular region
Integral to membrane
Collagen catabolic process
Sexual reproduction
C7orf10 cluster (140 genes)
Membrane-bounded vesicle
Response to unfolded protein
STAT6 cluster (942 genes)
Synapse
Neuron projection
Regulation of synaptic plasticity
Voltage-gated cation channel activity
Neurogenesis
Ne.
As.
Ol.
Mi.
En.
Slices
Expression z-score
Enrichment P-value

Figure 2.3: (Previous page) ***(top)*** *The network consists of the high-confidence genes and their co-expression partners that are connected if they have a co-expression value* > 0.6*. Each gene is shown as a circle and named with its gene name, with the size of both corresponding to its gene-based P-value (larger size corresponding to a lower P-value). The colours of the circles correspond to those of modules A – E in Figure 2.1: blue for module A, yellow for module B, green for module C, red for module D, purple for module E and grey for all other modules. The edge colours are matched to (a mixture of) the colours of the connecting genes.* ***(bottom)*** *For each high-confidence gene and its co-expressing partners are shown: 1) the number of genes in the local co-expression network around the high-confidence gene; 2) the average brain gene expression level from blue (low expression) to red (high expression) mapped in the three coronal brain sections (for brain region names in the coronal sections, see Figure S3); 3) the enrichment of cell type-specific genes in the table from white (P-value* > 0.05*) to black (P-value* $< 10^{-7}$*); and 4) the top five enriched gene functions. Not shown are boxes for high-confidence genes TRPM8, SUV39H2 and FHL5 because these genes have no or only few co-expressed genes. Ne: neuron; As: astrocyte; Ol: Oligodendrocyte; Mi: microglia; En: endothelial cell.*

The most highly connected high-confidence gene is STAT6, which has strong co-expression with genes from module A (connections marked in blue in Figure 2.3) and two genes from module C (connections marked in green), but is not part of either of these modules. Genes *DCLRE1C* and *LRP1* lie in a sub-network containing genes from module B (connections marked in yellow). *LEPROTL1* and *UFL1* are directly connected to genes from module D (marked in red). *SUV39H2* and *TRPM8* have no strongly co-expressed genes in the Allen Human Brain Atlas and remain unconnected. *MBOAT4* lies in a disconnected sub-network. The remaining 6 high-confidence genes are indirectly connected to the genes of modules A – D. The smallest module of interest, module E, has no genes in the local seed network.

2.2.7 Local seed network shows enrichment of functions and cell types similar to modules A – D

We performed a functional enrichment analysis in the local seed network, thereby focussing on each high-confidence gene and its co-expressing partners (Figure 2.3; Table S6). Briefly, a local network for each high-confidence gene was constructed by connecting it to genes with which it has a spatial gene co-expression larger than 0.6. The network around *STAT6*, *C7orf10* and

MBOAT4 showed enrichment of functions involved in the synapse and signal transduction. The network around *LEPROTL1* showed enrichment of mitochondrial genes. Functions involved in gene expression regulation were found in the networks around *DCLRE1C*, *LRP1* and *UFL1*. Other enriched functions were "circadian rhythm" (*NAB2* network), "apoptosis" (*UFL1* network), and "protein catabolism" (*LEPROTL1* network).

Finally, we investigated the enrichment of brain cell type-specific genes in the local seed network (Figure 2.3; Table S7). The co-expression network around *STAT6*, that shares many genes with module A, is highly enriched for neuron-specific genes ($P = 4.37 \cdot 10^{-32}$), as is the network around *NAB2* ($P = 2.50 \cdot 10^{-4}$). The sub-network connected to *UFL1*, overlapping with module D, contains many oligodendrocyte-specific genes ($P = 1.26 \cdot 10^{-8}$). The sub-networks connected to *PRDM16* and to *C7orf10* are enriched for astrocyte-specific genes ($P = 3.82 \cdot 10^{-7}$ and $4.34 \cdot 10^{-10}$, respectively).

2.3 Discussion

We performed a gene-based analysis of migraine GWAS data from a large meta-analysis of in total 23,285 migraine cases and 95,425 population-matched controls available through the International Headache Genetics Consortium (Anttila et al., 2013) aimed at identifying brain regions, cell types and pathways involved in migraine pathophysiology. To this end, we used detailed spatial brain gene expression data from 3,702 samples of six normal adult human brains from the Allen Human Brain Atlas to group genes into co-expression modules. We identified five modules enriched for migraine-associated genes that show involvement in cortical neurotransmission, protein catabolism and energy supply (Modules A and C); in gene transcription regulation in cortex and cerebellum (Module B); and in myelination and energy supply in subcortical areas (Module D) (Figure 2.4).

The lack of causal variants with large effect sizes for common migraine may explain, at least partly, the low enrichments of candidate genes in the co-expression modules. The conversion of the migraine GWAS data to the gene-based P-values may have caused inaccuracies as we may have associated SNPs to genes just because they are nearby these genes, although they may not have a functional effect on them; and, similarly, we may not have associated SNPs to genes simply because we considered them too far away to be

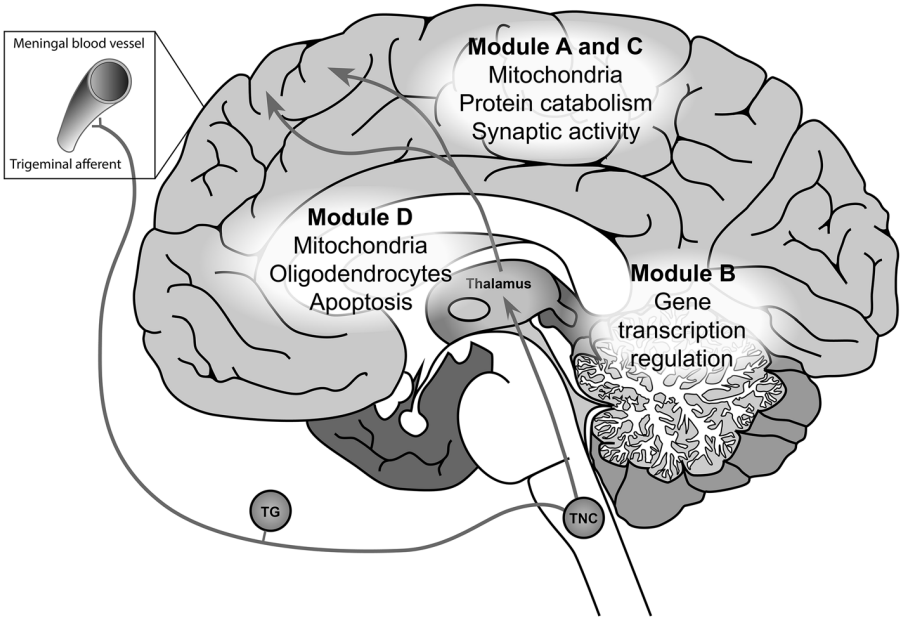


Figure 2.4: The migraine-associated modules A – D, which also overlap with the local migraine-related co-expression gene network, point to three distinct locations in the brain: the cortex (modules A, B and C), the cerebellum (module B) and the white matter and subcortical regions including the thalamus (module D), and multiple gene functions or cell types. Several brain regions overlap between the migraine-associated modules and the trigeminovascular system that is thought to generate the migraine headache. This system consists of trigeminal afferents that innervate the blood vessels in the meninges, whose signals are transmitted through the trigeminal ganglion (TG), the trigeminal nucleus caudalis (TNC), and the thalamus to the cortex where they can produce the sensation of pain.

functionally involved. To reduce these limitations we chose a 15-kb boundary around the genes, as it was shown that most SNPs that affect gene expression are located within this boundary (Pickrell et al., 2010). However, currently, no methods are available to calculate gene-based P-values that can fully surmount these limitations. To increase the reliability of our results, we used the largest migraine GWAS dataset currently available (Anttila et al., 2013). Furthermore, we used a second method to confirm the link between migraine and the

brain regions and gene functions identified by building a migraine-related co-expression gene network around the high-confidence migraine genes. Although the enrichment of migraine associated genes in the modules cannot proof that these brain regions, cells and pathways are dysfunctional in migraine patients, it can provide genetic evidence for processes already implicated in migraine, and may indicate new areas of interest for migraine research.

Two modules enriched for migraine-associated genes contained genes highly expressed in cortex that are largely involved in neurotransmission and that are highly enriched for neuron-specific genes (Modules A and C). Furthermore, module A contains many components of the glutamatergic system (*GLS*, *GRIK3*, *GRIN2A* and *GRM7*). The cell type enrichments in the modules were based on gene expression data from isolated mouse brain cells (Y. Zhang et al., 2014). Similar data from mouse studies have been used previously for characterization of human brain co-expression modules (Hawrylycz, Miller et al., 2015). These results confirm the link between cortical neurotransmission and migraine that had previously been identified in genetic studies in FHM (Ferrari et al., 2015). Several genes (*MTDH*, *LRP1*, *MEF2D*) identified by GWAS hits for common migraine could also be linked to glutamate signalling (Tolner et al., 2015), although these genes are not part of modules A or C.

The enrichment of genes involved in mitochondria in modules A and D form the first genetic link between mitochondrial function and common migraine. As neurotransmission requires a large amount of energy, it is not surprising that mitochondrial deficiencies have been implicated in a wide range of neurological disorders, including migraine (Sparaco, Feleppa, Lipton, Rapoport & Bigal, 2006). In migraine patients, magnetic resonance spectroscopy studies have consistently identified a depletion of brain high-energy phosphates, indicative of a disturbed energy metabolism (Reyngoudt, Achten & Paemeleire, 2012). Impaired mitochondrial activity has also been found in muscle and platelets of migraine patients (Reyngoudt et al., 2012; Sangiorgi et al., 1994). Also the efficacy of riboflavin and coenzyme Q10, two enhancers of mitochondrial function, in migraine prophylaxis in two small clinical trials points towards a possible causal role for mitochondria in migraine (Sandor et al., 2005; Schoenen, Jacquy & Lenaerts, 1998).

Module B shows high expression in cerebellum and medium expression in cortex, and is highly enriched for genes involved in aspects of gene expression regulation (i.e. transcription factors, chromatin remodellers, RNA processing).

Migraine pathophysiology has already been associated with actions of a specific set of transcription factors, i.e. female hormone receptors and receptors for the stress hormone cortisol (MacGregor, 2004; Sauro & Becker, 2009). Although the stress hormone receptor gene *NR3C1* is a member of module B, the other stress hormone receptor gene *NR3C2* and the female hormone receptor genes *ESR1*, *ESR2*, *RXFP1*, *RXFP2* and *PGR* are members of modules F, N, M, P, R and H, respectively. These transcription factors can thus not explain the association of module B with migraine. As to the high expression in the cerebellum, there are several lines of evidence that indicate a role for the cerebellum in migraine. (Subclinical) cerebellar abnormalities have been recognised in migraine patients, including lack of fine coordination (Sándor, Mascia, Seidel, De Pasqua & Schoenen, 2001) and vestibulocerebellar problems (Harno et al., 2003). Furthermore, studies using magnetic resonance imaging (MRI) identified cerebellar infarcts (Kruit, 2004) and microstructural cerebellar abnormalities (Granziera et al., 2014) in migraine patients. Cerebellar mechanisms causative of migraine are not known, but may possibly include signalling cascades that regulate gene expression as identified in module B.

Module D contains genes highly expressed in several subcortical brain regions and in the white matter and is enriched for gene functions involving myelin formation and genes specifically expressed in oligodendrocytes. Oligodendrocytes play key roles in the formation of axons and neuronal connections (Debanne, Campanac, Bialowas, Carlier & Alcaraz, 2011), and can also actively communicate with neurons in order to regulate their activity (Butt, Fern & Matute, 2014; Fields, 2008; Stys, 2011). The genes from module D are expressed in multiple brain regions that are implicated in the processing of migraine pain signalling: the trigeminovascular pathway (Noseda & Burstein, 2013)(Noseda and Burstein 2013). This pathway transmits nociceptive signals from meninges to thalamus and higher brain areas via several brainstem nuclei, including the trigeminal nucleus caudalis (TNC), (Figure 2.4). A recent study identified disrupted myelin sheets in the trigeminal nerve of migraine patients (Guyuron et al., 2014), providing first evidence for disturbed oligodendrocyte functioning in the trigeminovascular pathway. Furthermore, a high-field MRI study identified thalamic microstructural abnormalities in migraine patients that could indicate an increase of myelin (Granziera et al., 2014).

In summary, we performed a gene-based analysis of the migraine GWAS data, using detailed spatial gene expression data to define gene modules with

similar expression patterns in the normal human brain. Our results showed enrichment of migraine-associated genes in modules involved in cortical neurotransmission, mitochondrial and oligodendrocyte function that provide further evidence that these mechanisms play a causal role in migraine and deserve to be investigated in more detail by (functional) studies in patients and experimental animal models.

2.4 Materials and Methods

2.4.1 GWAS dataset

Summary statistics of migraine GWAS data from 23,285 cases and 95,425 controls from the meta-analysis available through the International Genetics Headache Consortium (Anttila et al., 2013) were used for this study. The quality control of the genotype data was described previously (Anttila et al., 2013). Autosomal SNPs were imputed against the HapMap CEU population (release 21-24 depending on the cohort). To convert the genomic coordinates of the SNPs from human reference genome build 36 to build 37, we used CrossMap (`http://crossmap.sourceforge.net/`) (Zhao et al., 2014). A total of 1,853,579 SNPs with high quality GWAS data and converted to build 37 were used in the calculation of gene-based P-values.

2.4.2 Gene-based P-values

Gene-based P-values were calculated from GWAS data using the gene-based test GATES (M.-X. X. Li, Gui, Kwan & Sham, 2011) implemented in the whole-genome analysis platform Fast ASsociation Test (FAST) (Chanda, Huang, Arking & Bader, 2013). GATES is a Simes test extension that integrates SNP P-values into a gene-based test statistic, based on SNP positions and LD information (1,000 Genomes data (Phase 1)) by taking the top SNP per gene and correcting its P-value for the effective number of independent tests. Gene location information based on the GRCh37.p13 build reference sequence was obtained from Biomart (version 75: Feb 2014 archive site). A flanking region of 15 kilobase (kb) up- and downstream of the gene was used to include SNPs located in regulatory regions. The size of the flanking region was based on the identification that most SNPs that influence the expression of a gene are located within 15 kb of the gene (Pickrell et al., 2010). Genes with a gene-based

$P < 0.05$ were considered migraine 'candidate genes'; genes with a Bonferroni corrected $P < 0.05$ were considered 'high-confidence genes'.

2.4.3 Spatial gene expression

Spatial gene expression data from six healthy adult human brains was obtained from the Allen Human Brain Atlas (`http://human.brain-map.org/`) (Hawrylycz et al., 2012). For each brain, RNA had been extracted from 363 to 946 different brain samples and measured on custom Agilent microarrays containing the 4x44K Agilent Whole Human Genome probes as well as an additional 16,000 custom probes. The expression data was matched to the GATES output based on Biomart associations of 4x44K Whole Genome microarray probe IDs with genes. If a probe was matched to multiple genes, it was excluded from the analysis. If multiple probe IDs were associated with the same gene, average expression levels were calculated for that gene. The spatial expression of a gene for a particular brain is thus described by the expression levels of that gene across all samples in that brain. Since the number of brain samples differs per brain, the spatial gene expression vector of a gene differs in length between brains.

2.4.4 Spatial gene co-expression and hierarchical clustering

Spatial co-expressions between genes were first calculated for each brain separately. For this, robust bi-weight mid-correlations were calculated across all brain samples for each of the six donors separately (Langfelder and Horvath 2012). Subsequently, these correlations were averaged across the donors to obtain co-expression values that only reflect spatial expression patterns and ignore between-brain differences. We then performed hierarchical clustering to obtain modules of spatially co-expressed genes. The linkage and distance measures, and the threshold at which the tree is cut, were chosen to maximise the enrichment of migraine candidate genes (see Supplementary Material and Methods for different combinations of linkage and distance measures). We chose for this independent evaluation over traditional cluster evaluation measures (like WGCNA (B. Zhang & Horvath, 2005)) as we are interested in finding modules (clusters) that are related to migraine genes. Eventually, clustering was done with complete linkage, with one minus the bi-weight mid-correlation as a distance measure, and the tree was cut into 18 clusters.

2.4.5 Enrichment of candidate genes in the modules

Enrichment of migraine-associated genes within a module was determined using a Fisher exact test, that calculated whether the number of migraine candidate genes in a module is higher than expected based on the total number of genes and migraine candidate genes. Neighbouring genes on the genome might have similar expression patterns due to local regulatory DNA elements, as well as similar gene-based P-values due to LD between their top SNPs or overlapping flanking regions. Therefore, we performed a second LD-corrected Fisher exact test in which we included only the number of independent genes in the calculation. As a measure for the number of independent genes in a gene set, we took the top SNP of each gene and used the Genetic type I Error Calculator (GEC) (M.-X. X. Li, Yeung, Cherny & Sham, 2012) to calculate the effective number of independent SNPs based on LD information from the HapMap project release 23. In this way, the LD-corrected Fisher exact test had as input the corrected estimates for the number of independent genes with gene-based P-values below and above 0.05, both in the cluster of interest and in the full set of genes. See Supplemental Material and Methods for additional information on the enrichment analysis.

2.4.6 Functional annotation

Gene Ontology (GO) term and pathway enrichment analysis in the modules was performed with DAVID (version 6.7; `http://david.abcc.ncifcrf.gov/`). We used the Functional Annotation Clustering tool in DAVID to group significant GO terms and pathways based on co-associated genes to remove redundant terms (D. Huang et al., 2007). Pathway information from KEGG, Reactome and PANTHER, and GO term information (biological processes, molecular functions and cellular components) from PANTHER, and the FAT subsets of GO terms was used. GO term and pathway groups were considered significant when the EASE score was larger than 1.3 (corresponding to a geometric mean Benjamini-corrected P-value of the clustered GO terms and pathways below 0.05). Significant groups were named after the most significant term in the group. Comparison of GO term and pathway enrichments between modules was performed in ToppCluster, a multiple gene list feature enrichment analyser (Kaimal, Bardes, Tabar, Jegga & Aronow, 2010). In ToppCluster, we performed GO term (biological processes, molecu-

lar functions and cellular components) and pathway enrichment analyses for all modules, which were considered significant when Bonferroni-corrected P-values were below 0.05. Functional enrichments and overlap in enrichments between modules were visualized in Cytoscape (version 3.2.1).

2.4.7 Cell type enrichment

For enrichment analysis of cell type-specific genes we made use of cell type-specific genes identified in gene expression data from isolated mouse brain cells (Q. Zhang, Burdette & Wang, 2014). We selected the gene expression data from neurons, astrocytes, myelinating oligodendrocytes, microglia, and endothelial cells. Genes were considered cell type-specific if they had more than 10-fold higher gene expression (reads per kilobase per million (RPKM)) levels compared to the mean expression in the other cell types. We obtained 818 neuron-, 380 astrocyte-, 198 oligodendrocyte-, 692 microglia-, and 546 endothelial-specific genes for which human orthologs were present. Enrichment was determined with Fisher exact tests.

2.4.8 Local modules from seed genes

Local co-expression networks were built from high-confidence genes by adding genes to the network whose co-expression exceeds a threshold (similar to Willsey et al. (2013)). Genes were only selected if they had co-expression values higher than 0.6 with a high-confidence gene. The threshold was chosen to: 1) maintain only reasonably strong links between genes, especially given the fact that we use robust bi-weight mid-correlations; and 2) have linking genes for most of the seed genes (see Supplementary Material and Methods for information on how the threshold value was selected). Co-expressions were measured as bi-weight mid-correlations, the same co-expression values which were used to determine the genome-wide co-expression modules, and local modules were defined as all genes connected to a single high-confidence gene. If a gene is connected to two high-confidence genes, it is part of the modules of both genes.

2.5 Conflict of interest

The authors declare that they have no conflict of interest.

2.6 Funding

This research was supported by the Dutch Technology Foundation STW, as part of the STW project 12721: “Genes in Space” under the IMAGENE perspective program; the Spinoza (2009) grant to M.D.F.; European Union Seventh Framework Programme projects EUROHEADPAIN project [grant number 602633] & Human Brain Project [grant number 604102]; and the Center for Medical Systems Biology (CMSB) established in the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO) [project nr. 050-060-409].

— 3 —

BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome

Abstract

Spatial and temporal brain transcriptomics has recently emerged as an invaluable data source for molecular neuroscience. The complexity of such data poses considerable challenges for analysis and visualization. We present BrainScope: a web portal for fast, interactive visual exploration of the Allen Atlases of the adult and developing human brain transcriptome. Through a novel methodology to explore high-dimensional data (dual t-SNE), BrainScope enables the linked, all-in-one visualization of genes and samples across the whole brain and genome, and across developmental stages. We show that densities in t-SNE scatter plots of the spatial samples coincide with anatomical regions, and that densities in t-SNE scatter plots of the genes represent gene co-expression modules that are significantly enriched for biological functions. We also show that the topography of the gene t-SNE maps reflect brain-region specific gene functions, enabling hypothesis and data driven research. We demonstrate the discovery potential of BrainScope through three examples: 1) analysis of cell type specific gene sets, 2) analysis of a set of gene co-expression modules that are stable across the adult human donors, and 3) analysis of the evolution of co-expression of oligodendrocyte specific genes over developmental stages. BrainScope is publicly accessible at www.brainscope.nl.

This chapter has been published as: Huisman, S. M. H., van Lew, B., Mahfouz, A., Pezzotti, N., Höllt, T., Michielsen, L., ... Lelieveldt, B. P. F. (2017). BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic Acids Research, 45*(10), e83. `https://doi.org/10.1093/nar/gkx046`
All supplemental materials can be found in the online publication.

3.1 Introduction

The field of molecular neuroscience has seen a sharp rise in the availability of spatially mapped molecular data, accessible through public databases. General databases such as GTEx (Lonsdale et al., 2013) and Encode (Dunham et al., 2012), but also brain-specific databases like PsychENCODE (Akbarian et al., 2015), contain anatomically annotated gene expression and epigenetic data across the brain. Where some projects focus on specific diseases (such as Huntington's disease (Neueder & Bates, 2014) and autism spectrum disorder (Voineagu et al., 2011)), others aim to capture general patterns in the healthy brain. A strong example of the latter are the efforts of the Allen Institute for Brain Science (Sunkin et al., 2013) to measure spatially mapped gene expression in mouse, macaque and human brain, both in the healthy adult individual and throughout brain development. These genome-wide studies of the transcriptome aim to elucidate relationships between brain structure and brain function, and identify genes that play a role in this.

Understanding brain transcriptome data is challenging, since it encompasses RNA expression over all genes, across many spatial coordinates of the brain, and through development in time. A powerful way to obtain insight into such complex multi-way data sets is by visually exploring the data using principles of presenting, browsing, and selecting. Currently available tools for analyzing gene expression in the brain that incorporate visualization include the Allen Institute's AGEA (L. Ng et al., 2009) and Neuroblast (Hawrylycz et al., 2011). These two portals represent two distinct views on the data. With AGEA, researchers can explore the interplay between anatomical connections and the gene expression similarities of brain areas. It shows sample-sample similarities and provides a parcellation of the brain entirely based on transcriptome data. A different view on the same data is offered by Neuroblast. Here, the focus lies on gene-gene comparisons: it shows which genes have similar spatial expression patterns in the healthy brain. Both AGEA and Neuroblast are valuable tools that have been used to study, for instance, bipolar disorder (McCarthy, Liang, Spadoni, Kelsoe & Simmons, 2014). However, these tools focus either on relationships between genes, or on the relationships between brain regions, while the interplay between these two is an essential part of the data. A suitable representation of brain transcriptome data that links a gene-centric and a sample-centric view is currently lacking.

The relationships between genes or samples can intuitively be represented in plots, where these elements are shown as points. The closeness of the points then represents their similarity. However, with a large number of samples and thousands of genes, a plot that reflects similarities needs to capture a high-dimensional space in a two-dimensional map. Common ways to reduce this dimensionality are multi-dimensional scaling (MDS) (Tzeng, Lu & Li, 2008) and principle component analysis (PCA) (Ma & Dai, 2011). A more recently introduced non-linear dimension reduction method is t-distributed stochastic neighborhood embedding (t-SNE) (Maaten & Hinton, 2008). The power of t-SNE comes from the fact that it tries to accurately represent the local neighborhoods of points, so neighbors in the plot match those in the original high dimensional data. In return, the distances between dissimilar points are less well-preserved. This is in marked contrast to, for example, PCA where the important components capture the direction of the largest variance across the points, which is generally reflected in distant (dissimilar) points. t-SNE has been used to produce transcriptional maps of brain regions in the Allen Brain Atlas (ABA) (Ji, 2013; Mahfouz et al., 2015), and it is popular in the analysis of single-cell molecular data (Macosko et al., 2015; Shekhar, Brodin, Davis & Chakraborty, 2014; van Unen et al., 2016; Wong et al., 2015).

Here, we present BrainScope, a portal that uses t-SNE maps of both samples and genes in an interactive visualization of the transcriptional landscape of the brain. It gives a brain- and transcriptome-wide view of gene co-expression and transcriptional similarity of brain regions, based on the human brain data of the Allen Institute (Hawrylycz et al., 2012; Miller et al., 2014). It allows for interactive analysis of gene expression in the human brain, in an intuitive visual way. To connect the gene-centered and the sample-centered views, we make use of *linked maps*: t-SNE plots where a selection of points is rendered as a visual change in the linked plots. The first instance of this is the *dual explorer* (see Figure 3.1**a**), which has a single transcriptome-wide *gene map* and a brain-wide *sample map*. Users can select genes or samples and show their mean expression patterns in the other map. In addition, this part of the portal contains brain *choropleths*: user-selected slices of the human brain that are used to localize samples and illustrate spatial expression patterns. In addition to the dual explorer, the portal contains the *comparative explorer* (see Figure 3.1**b**), which focuses on the comparison of several gene

maps, representing distinct donor brains. Therefore, the comparative explorer reveals inter-donor similarities in co-expression. Using the adult human data it shows robustness of co-expression modules, while for the developmental human data it shows changes in co-expression through time. Each part of the portal contains a direct link to enrichment tools Enrichr (E. Y. Chen et al., 2013) and ToppGene (J. Chen, Bardes, Aronow & Jegga, 2009), to provide a functional interpretation of selected gene sets.

The linked t-SNE maps of the BrainScope can conceptually be used in several ways. Selection of a single point reveals the corresponding expression pattern, either of a gene throughout the spatially mapped samples, or of a sample across all genes. Selection of points in the sample map reveals gradients of expression in the gene map, which elucidate gene-gene relationships. In addition to single point selection, a set of points can be selected to study the relationships between these points (co-expression or transcriptional similarity) and characterize sub-clusters by their mean expression patterns. In the comparative explorer, any selection of genes is carried over to all gene maps, showing differences in co-expression between brains. We demonstrate the usefulness of BrainScope by exploring the major patterns of gene expression in the adult human brain, and the way these reflect gene function and cell type composition of brain samples. In addition, user-supplied gene sets can be examined for structure. With the comparative explorer, we highlight the stability of the gene t-SNE maps over the six donor brains of the Allen Brain Atlas, in line with recently published consensus modules (Hawrylycz, Miller et al., 2015). Finally, the spatio-temporal transcriptome shows that the changes in expression of oligodendrocyte marker genes reflect the development of the brain. Combined, these applications enable a unique view of the rich gene expression data of the Allen Human Brain Atlases.

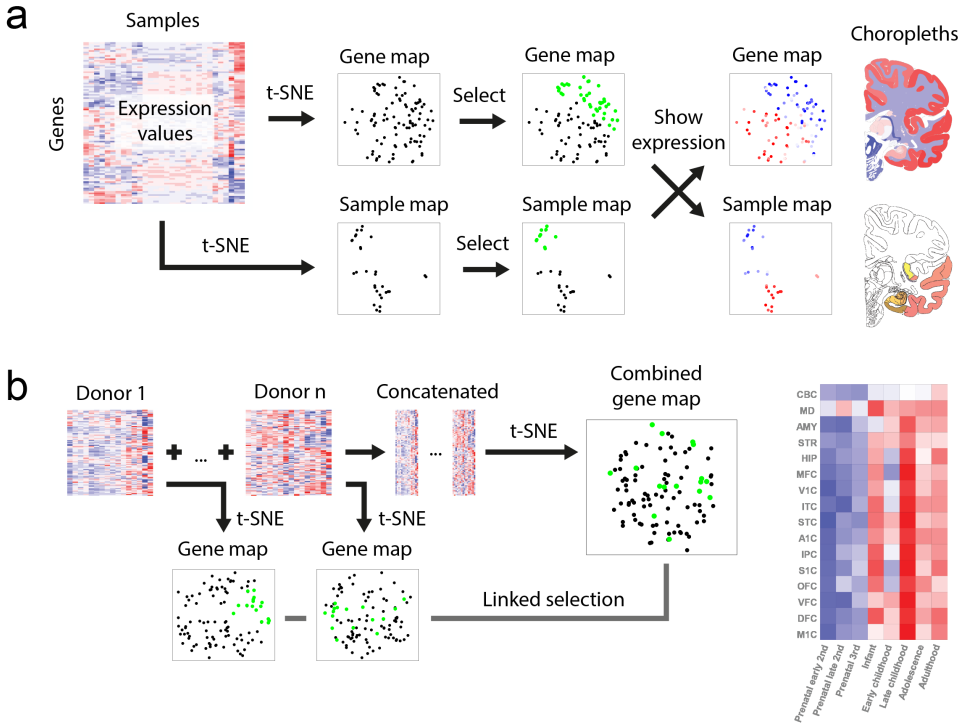


Figure 3.1: BrainScope views: ***(a)*** *In the* dual explorer*, the gene expression data is visualized in two directions: a map for genes and a map for samples. Points that are close in the map have a high similarity. The portal allows for selection of points in either of the maps and shows the expression in the other map (red is high, blue is low): a set of genes has a profile across samples, which is averaged; and a set of samples has expression values in all genes, which are also averaged. When genes are selected, their average expression is also shown on brain slice choropleths; and when samples are selected, their location is shown on the same choropleths.* ***(b)*** *In the* comparative explorer*, only gene t-SNE maps are shown, but it contains data for multiple donor brains (replicates, or developmental stages). A t-SNE map is made for each donor and, in addition, one map is made for the combined data sets. When a selection of genes is made in either of the maps, this selection is carried over to the other maps and the average regional expression of these genes is shown in a heatmap.*

3.2 Materials and methods

3.2.1 Gene expression data

Gene expression data was obtained from the Allen Institute for Brain Science. The Adult Human Brain Atlas (Hawrylycz et al., 2012) contains gene expression measurements of six healthy adult donors. Samples were taken using macro- and microdisection of anatomically annotated regions. The number of samples differs per donor, from 363 to 946, with a total of 3 702 samples. The expression values in each sample were determined with a customized microarray chip, measuring 58 692 probes. Initial data processing was performed by the Allen Institute, and the data were made available on their website (http://human.brain-map.org/static/download).

The Developing Human Brain Atlas (Miller et al., 2014) has a lower spatial resolution, but samples were taken from human donors of a broad range of ages. In total 42 brains were sampled, ranging in age from 8 weeks post-conception to 40 years old. The number of samples per brain ranges from 1 to 16, with a total of 524 anatomically annotated samples. Gene expression was determined using RNA sequencing, and RPKM values are available online for 52 376 genes (http://www.brainspan.org/static/download).

3.2.2 Data preprocessing

In the adult human brain data the 58 692 probes were mapped to 19 992 genes, using their Entrez identifiers. For genes that have two probes, the probe with the highest variance was selected. For genes with more than two probes, we picked the probe with the highest connectivity to all other probes (defined as the sum of Pearson correlations). The number of samples differs per donor brain. To enable combination of the data for dual t-SNE, all expression sets were reduced to have 105 values per gene, corresponding to the annotated regions that were sampled in each brain. Finally, to obtain a single gene and sample map in the dual explorer, the expression values for each combination of sample and gene were averaged over the six donors. The comparative explorer of the adult brain instead uses processed data for all brain samples (Hawrylycz, Miller et al., 2015). To enable a direct comparison between densities in the gene t-SNE maps and previously defined WCGNA-based gene

modules (Hawrylycz, Miller et al., 2015), both were computed from identical gene-sample data matrices.

In the developing human dataset samples were pooled into eight age windows, to obtain subsets with higher sample sizes. Supplementary Table S1 shows which donor brains were combined into each age window, with sample sizes and donor characteristics. Anatomical regions were required to have at least one sample for each age group, giving 16 regions with eight measurements each. From the 52 376 genes, only the 18 233 genes were selected that had an RPKM-value above 1 in at least 20% of all samples.

The dimension reduction results of t-SNE are dependent on scale and location of the data. For the gene maps all genes were z-score normalized, to have zero mean and a standard deviation of 1. For the sample maps the values for each sample were instead z-scored.

3.2.3 Dimension reduction

Dimension reduction was performed with t-distributed stochastic neighborhood embedding (t-SNE), a non-linear embedding technique (Maaten & Hinton, 2008). It creates a low-dimensional map of high-dimensional data, while preserving as much of the local structure as possible. The method has one main parameter, the perplexity value, which determines the variances of the Gaussian kernels that are used to calculate similarities in the high-dimensional space. The higher the perplexity value, the larger the number of neighboring points to which similarities are preserved. Because t-SNE only aims to preserve neighborhoods, the rotation of the maps is arbitrary. In the comparative explorers, the maps were rotated to be as similar as possible (defined by the mean Euclidean distance of all points). In many applications a PCA reduction to a somewhat lower dimensional space is performed, for computational and noise reduction reasons. In our analyses we did not perform this step, in order to retain the original neighborhoods. The gene t-SNE maps were made with the default perplexity value of 30. The sample t-SNE maps were made with a lower perplexity value of 10, due to the lower number of points.

3.2.4 Gene set clustering and analyses

We characterized the 3 000 genes in the regions of highest density of the gene map for the adult human data. These were identified in a Gaussian density

estimate of the map, with an identity covariance matrix. The 3 000 genes with highest local densities were then hierarchically clustered with Euclidean distance and complete linkage. The optimal number of clusters (27) was determined by maximizing the silhouette score, and the 23 clusters with more than 30 genes were characterized by their expression patterns and enrichments in ToppGene (J. Chen et al., 2009).

To define cell-type marker genes, we made use of a database of gene expressions from fluorescence-activated sorted cells of the mouse cerebral cortex (Y. Zhang et al., 2014). Genes were selected as markers when they had a 20 fold higher expression in the cell type of interest than the geometric mean of the other cell types. Mouse gene identifiers were matched to human orthologs using builds GRCm38.p4 and GRCh38.p3 in BioMart (Smedley et al., 2015).

Clustering of post synaptic density related genes (Bayés et al., 2011; Hawrylycz et al., 2012) in the gene t-SNE map was performed with a Gaussian mixture model, where the number of clusters was optimized using the Bayesian information criterion. The reported gene set enrichment analyses were performed in ToppGene (J. Chen et al., 2009).

3.3 Results

3.3.1 Dimension reduction of gene expression in the brain

BrainScope aims to visualize gene expression data of the brain, in an interactive and intuitive way (see Supplementary Video S1-3). It is built on spatially resolved gene expression data in the adult human brain, and the Brainspan atlas of the developing human brain, both provided by the Allen Institute for Brain Science. The adult human brain atlas contains genome-wide expression values of six donors, five males and one female, aged 24 to 57 years old. The in total 3 702 samples cover a wide range of anatomical regions, with 105 distinct regions that are sampled at least once in every donor. For the dual explorer, we averaged the expression values to these 105 regions for 19 992 genes (see the Materials and Methods section). For the comparative explorer of the developing human atlas, we grouped the measurements of 16 brain regions in eight developmental stages (see Supplementary Table S1). To produce two-dimensional maps of the expression data, we made use of t-SNE (Maaten & Hinton, 2008). A comparative analysis between t-SNE and PCA is provided

in the Supplementary Text and Supplementary Figure S2.

3.3.2 The dual explorer shows localized transcriptional similarity in human neuro-anatomy

The dual explorer contains two maps: a sample map and a gene map. In the sample map, samples are close together if they are similar in their gene expression profiles. By coloring samples with anatomical annotation colors, this map shows both anatomical relationships between samples, and their transcriptional similarity. For the adult human data, Figure 3.2**a** shows that samples of close spatial proximity are more likely to have similar expressions, so the sample t-SNE map reflects the anatomy of the brain. Note that this map was produced using only the transcriptional profiles of the samples, not their locations. All samples of the cerebral cortex are co-located in the map (cluster 8), while sub-clusters can be recognized for example for the frontal lobe and the hippocampal formation (cluster 7). The six regions of the amygdala cluster together (cluster 6), as do the five striatum regions (cluster 5), the three hypothalamus regions (cluster 4), and the seven dorsal thalamus regions (cluster 3). The ventral thalamus, on the other hand, is more similar to the anatomically adjacent globus pallidus and midbrain. The samples of the cerebellar cortex form a distinct cluster in the map (cluster 9), whereas the cerebellar nuclei (represented by the dentate nucleus) are most similar to samples from medulla and pons, the structures that anatomically connect the cerebellum to the midbrain.

The sample map reflects the similarities between samples, but the same transcriptome data can be used to infer gene-gene similarities. Figure 3.2**b** shows the transcriptional activity of 19 992 genes in the nine sets of samples that are selected in the sample map. The positions of the points (genes) in this t-SNE map capture brain-wide expression profiles, so the co-expression over the 105 selected regions. The colors of the points, on the other hand, show the activity of the genes in a selected subset of samples. For example, Figure 3.2**b**9 shows the average expression in cerebellar cortex samples, where we see a strong expression gradient from left-to right. These patterns of expression reflect how the gene map was made: genes with a similar expression across brain regions should be nearby in the t-SNE map.

The dual use of sample and gene maps (dual t-SNE) can give valuable

insights. The differences between anatomical annotation and the sample map highlight the importance of exploring similarities in the characterization of brain regions. For example, the cerebellar nuclei samples (which are part of cluster 1 in Figure 3.2) are very different from those of the cerebellar cortex (all part of cluster 9). If one were to look at the average expression pattern of the full cerebellum, this would be a mixture of two distinct expression patterns (that of Figure 3.2**b**1 and 3.2**b**9). The interplay between gene and sample map allows for quick exploration of brain region specific expression. For example, one group of spatially co-expressed genes are highly expressed in the hypothalamus samples (of cluster 4). When these genes are analysed for GO-term enrichment, we find they comprise several genes with hormone activity (Figure 3.2**c**). In addition to these specific analyses, one can also directly see large-scale patterns in the maps, such as the fact that few genes are highly expressed in both brain-stem and cerebral cortex (Figure 3.2**b**1 and 3.2**b**8).

As we have seen in the hypothalamus example, similarities in gene expression may point to similarities in function. The gene map captures gene co-expression networks, which may consist of functionally related genes. To test this, we considered parts of the gene map with large numbers of genes, so with a high density. Figure 3.3 shows the gene ontology enrichments and spatial expression patterns of 3 000 genes with the highest density values in the density map. Where Figure 3.2**a** shows similarities between neighboring brain regions, Figure 3.3 captures spatial co-expression networks. The results confirm the hypothesis that co-expression is related to shared functions, and it provides a global annotation of the gene map. The link between co-expression and function can also be used to characterize a gene by its neighbors in the gene map, as is illustrated for the APOE gene in the Supplementary Text and Supplementary Figure S1.

Gene-expression reflects cell type composition.

Gene expression measurements in the brain are partly determined by cell type composition of the samples (Grange et al., 2014). Therefore, the similarities between genes in their expression patterns may reflect cell type specific expression. As a result, cell type specific genes are likely to be co-located in the t-SNE map of the genes. To test this hypothesis, we obtained expression data

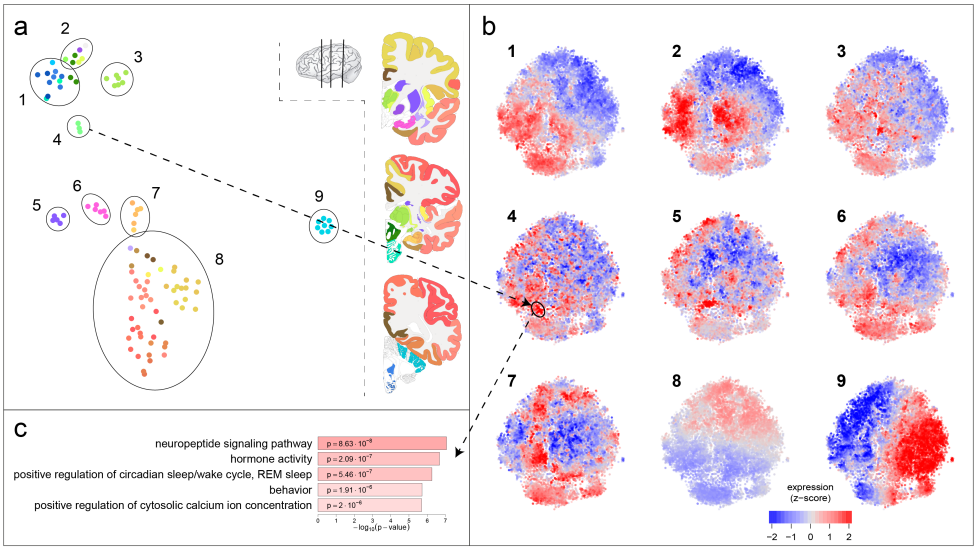


Figure 3.2: Characterization of the gene and sample maps in the dual explorer: ***(a)*** *The samples show a clustered pattern, matching the anatomical annotation shown in the color coding of the brain slices on the right. Nine groups of samples are highlighted in the map.* ***(b)*** *The mean expression of the nine groups of samples shown on the gene map. Each cluster of samples has its own distinct expression pattern, where red is high expression and blue is low expression. The dual explorer facilitates exploration of gene-sample relationships. The hypothalamus samples that are selected in* ***a4*** *have a high expression in the gene cluster highlighted in* ***b4****.* ***(c)*** *The five strongest GO-term enrichments in the hypothalamus related gene cluster, which point to well-known hypothalamus functions.*

from a cell-sort experiment of mouse cerebral cortex samples (Y. Zhang et al., 2014). We selected expression data of five major cell types present in the brain: astrocytes, endothelial cells, microglia, neurons and oligodendrocytes. Genes are labelled as cell type specific if they have at least a 20-fold expression in a specific cell type compared to the geometric mean of the other cell types.

The cell type specific genes (or cell type "markers") are co-located in the gene t-SNE map. Figure 3.4**a** shows the location of these genes, where neuronal markers are found at the top of the map, which contains genes with high expression in cerebral cortex (Figure 3.4**b,c**1). The endothelial cell markers

are also strongly co-expressed, with high average expression in thalamus, striatum and medulla (3.4**b,c**3). The microglia and oligodendrocyte markers form distinct clusters that share a high expression in the white matter (3.4**b,c**2 and 3.4**b,c**4). Microglia are known to be prevalent in the corpus callosum, which contains the "fountain of microglia", from which these cells migrate to other parts of the brain (Gehrmann, Matsumoto & Kreutzberg, 1995). Compared to the microglia markers, the oligodendrocyte markers have a somewhat higher expression in cerebral cortex and thalamus, but lower in hippocampal formation and amygdala. Oligodendrocytes are responsible for myelination in the central nervous system, so they are prevalent in white matter of the brain. Combined, these results show that the maps in the BrainScope portal pick up the detailed patterns of cell type specific expression that partly underlie the transcriptome of the brain.

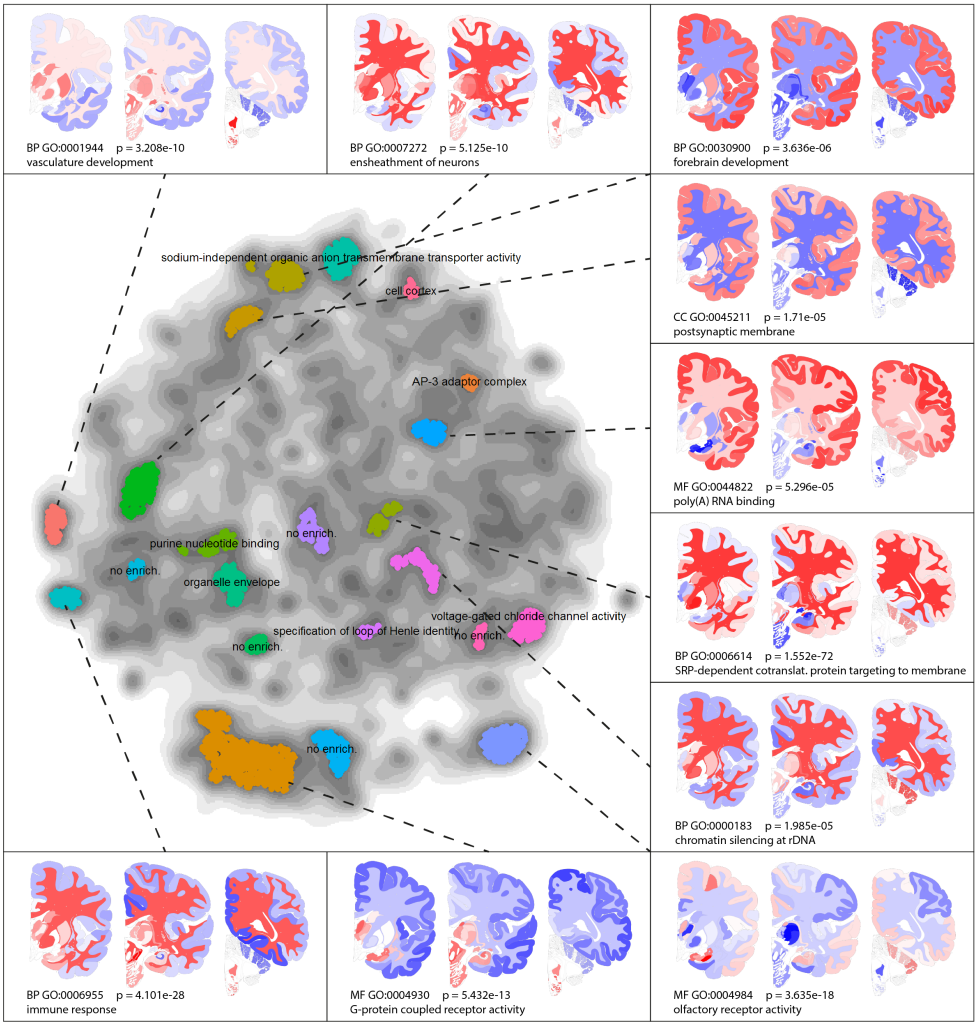


Figure 3.3: Functional characterization of the gene map: 3 000 genes within the highest densities were clustered, and clusters containing over 30 genes were characterized using ToppGene. Only the most significant GO-term is shown for each cluster, while the 10 clusters with strongest enrichments are provided with spatial expression choropleths. Most high-density areas in the gene map contain genes with common functions. All p-values are Bonferroni corrected in ToppGene, and the gene modules are provided in Supplementary Table S2.

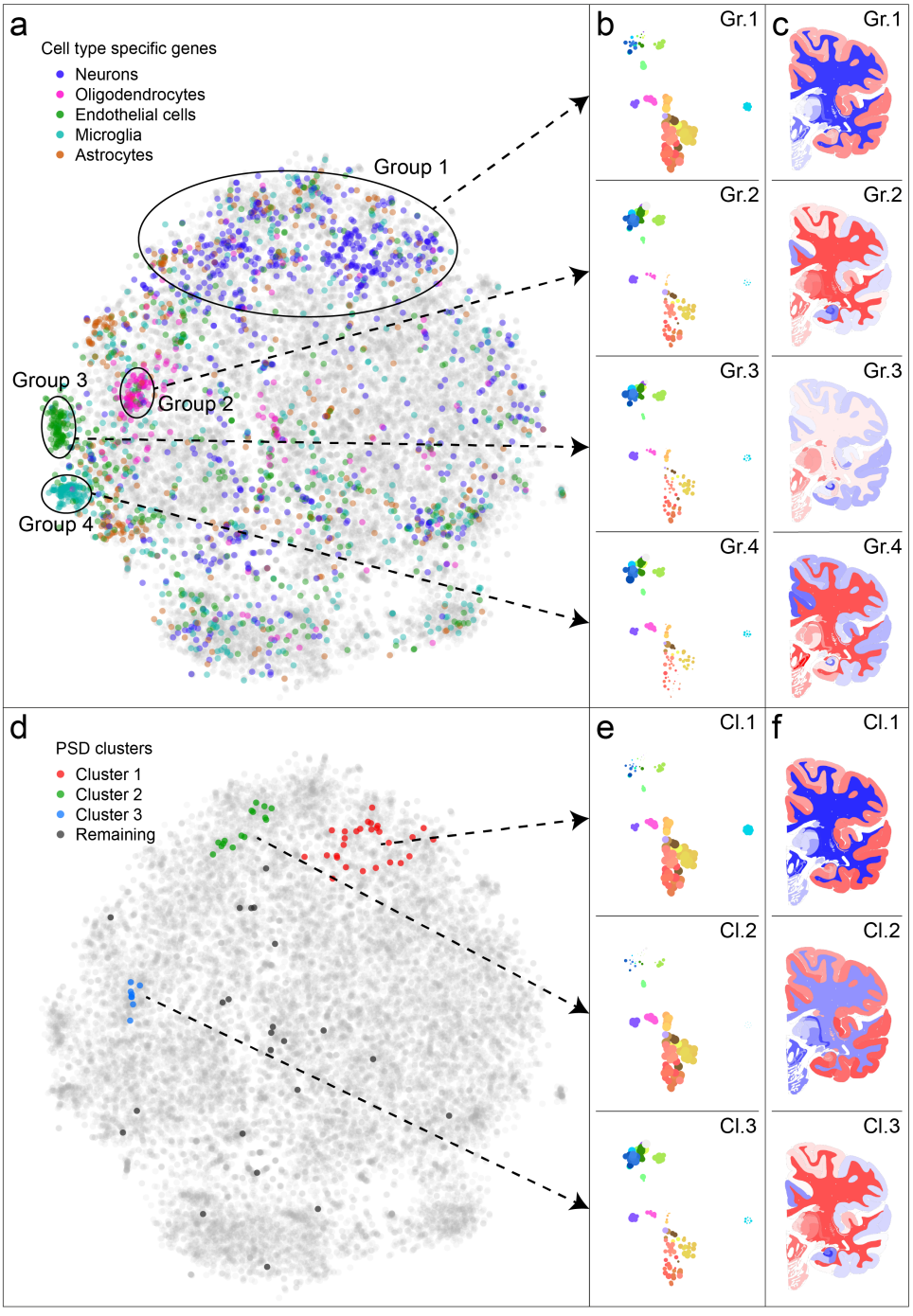

a
Cell type specific genes
Neurons
Oligodendrocytes
Endothelial cells
Microglia
Astrocytes
Group 1
Group 2
Group 3
Group 4
b
Gr.1
Gr.2
Gr.3
Gr.4
c
Gr.1
Gr.2
Gr.3
Gr.4
d
PSD clusters
Cluster 1
Cluster 2
Cluster 3
Remaining
e
Cl.1
Cl.2
Cl.3
f
Cl.1
Cl.2
Cl.3

Figure 3.4: (Previous page) Gene sets show cell type specificity of spatial gene expression and clusters of post synaptic density related genes: ***(a)*** *The adult human gene t-SNE map, with highlighted cell type markers. The cell type markers were picked based on data from fluorescence-activated cell sorted brain cells from mouse Y. Zhang et al. (2014), as genes with 20-fold expression in one of the types compared to the geometric mean in the other types. In the map, 4 groups of genes are highlighted, which correspond to areas of the map with high numbers of cell type markers.* ***(b)*** *The mean expression of all genes in the 4 groups in the gene map, shown by point sizes in the sample maps.* ***(c)*** *The mean expression of all genes in the same 4 groups, shown on a brain slice.* ***(d)*** *The adult human gene t-SNE map with highlighted post synaptic density (PSD) related genes Hawrylycz et al. (2012). Three clusters of co-expressed PSD related genes are highlighted.* ***(e)*** *and* ***(f)*** *The expression patterns of the three clusters, shown in the sample t-SNE map and choropleths. Clusters 1 and 2 contain genes mostly expressed in the cerebral cortex, where cluster 2 is distinct from cluster 1 because of its stronger expression in cerebellar cortex. Genes in cluster 3 are expressed most strongly in subcortical regions such as thalamus, hypothalamus, and brain stem.*

Dual explorer is an instrument for visual exploration of a set of genes of interest.

The dual explorer captures robust patterns of spatial co-expression in the brain. This allows for characterization of sets of genes with respect to their shared expression, and therefore potentially shared brain specific functions. To illustrate this, we selected the 74 genes that were identified to have strong regional expression in the brain and presence in post synaptic density (PSD) (Hawrylycz et al., 2012), using data from a proteomic profiling of human neocortex (Bayés et al., 2011). Post-synaptic densities connect neuronal cells and are essential to signal transmission in the brain. The 74 genes may all be specific to the PSD, but they do not all have identical spatial expression patterns in the brain. Figure 3.4**d** shows the PSD related genes in the gene map, where they can be separated into three clusters (and a remainder of unclustered genes). Cluster 1 contains 28 genes that are preferentially expressed in the cerebral cortex, and compared to all genes in the genome are enriched for the GO-term synapse part (*GO:0044456*, $p = 4.52 \cdot 10^{-23}$). Cluster 2 contains 15 genes that are similar in expression pattern, but have lower expression in the cerebellar cortex. They have the strongest GO-enrichment for synapse

(*GO:0045202*, $p = 8.82 \cdot 10^{-8}$). The 8 genes in cluster 3 have low expression in the cerebral cortex and cerebellum and high expression in subcortical regions, such as the thalamus and brainstem. This cluster is enriched for GO-term myelin sheath (*GO:0043209*, $p = 6.94 \cdot 10^{-9}$). In fact, 5 out of the 8 genes share this annotation, which is surprising for genes that have been identified as being PSD related. Hence, the dual explorer gives a clear view of the clusteredness of this gene set of interest and the number of discernible clusters. In general, it allows for rapid interactive exploration of spatial expression patterns and gene function.

3.3.3 Comparative explorer shows expression stability across donors in the adult human brain

The Allen Brain Atlas contains gene expression measurements for six adult brains. We compared the gene t-SNE maps of these donors, using the comparative explorer. The explorer contains a consensus gene map, made by concatenating all samples before dimension reduction, and six per-donor gene maps that use all samples taken from a single donor. Details on data processing can be found in the Materials and Methods section. Figure 3.5 shows all seven gene maps side-by-side. To allow for visual comparison, the genes are labeled using gene modules that have been found to be consistently co-expressed in each of the six donor brains (Hawrylycz, Miller et al., 2015). These previously published modules were created by first assessing each gene for stability, defined as the correlation between the expression vectors for each pair of donor brains. The 50% of genes with the highest differential stability were then selected for an initial clustering of genes. Subsequently, weighted gene co-expression analysis (WGCNA) (B. Zhang & Horvath, 2005) was used to obtain 32 modules, which were characterized by module eigengenes. To obtain genome-wide gene modules, the remaining genes (with lower differential stability) were then linked to their most similar modules, defined on the correlation with the module eigengenes. Figure 3.5 shows that many of the previously reported WCGNA modules consistently form clusters in the consensus t-SNE map, as well as in the per-donor maps, pointing to the robustness of these maps. The t-SNE method, with only one main parameter, offers a visual representation of the data that is strongly in line with the results of the more parameter sensitive WGCNA algorithm. The relative positions of the

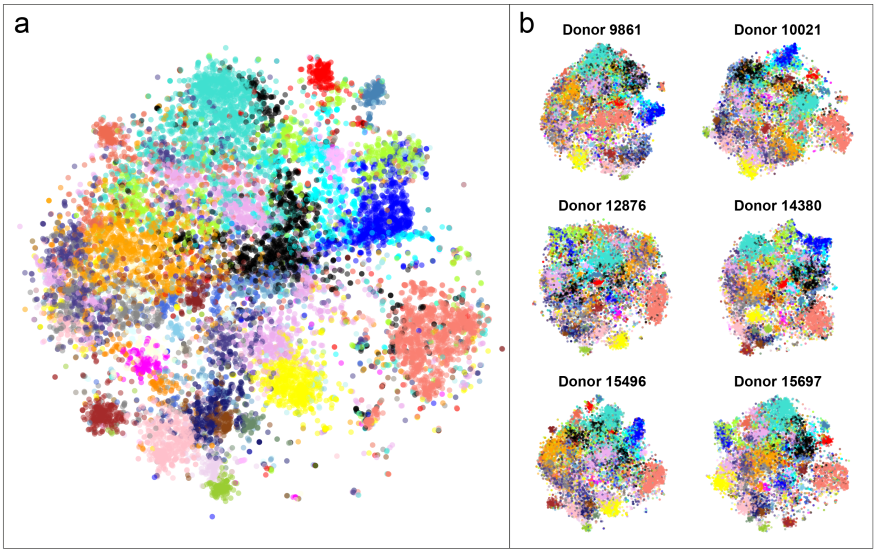


Figure 3.5: Gene t-SNE maps are robust and reproducible across donors: ***(a)*** *The combined gene t-SNE map, showing previously reported stable gene modules Hawrylycz, Miller et al. (2015). The map separates the 32 modules and shows their relationships.* ***(b)*** *The gene maps for each of the 6 donor brains. The maps are made using independent data sets, so they reflect the robustness of spatial gene expression patterns in the human brain. Data was pre-processed as in the original publication Hawrylycz, Miller et al. (2015) to enable direct comparison of the WGCNA modules to the gene t-SNE maps.*

modules in these maps vary to some extent, which is a result of the limited importance of large distances in t-SNE. In addition, the differences in brain region sampling may account for variability between donors.

3.3.4 Developmental comparative explorer captures spatio-temporal co-expression patterns

Thus far, we have only considered spatial gene expression patterns in the adult human brain. The Brainspan atlas of the developing human brain contains spatially and temporally resolved transcriptome data. To visualize this atlas, we developed the Brainspan comparative explorer (Figure 3.6). The Brain-

span human developmental atlas contains gene expression data for 42 brains, ranging in donor age from 8 weeks post-conception to 40 years after birth. From each brain, up to 16 anatomical regions were sampled. We summarized the data to contain mean expression values for each of the 16 anatomical regions, for 8 developmental stages: early second trimester of pregnancy, late second trimester, third trimester, infancy, early childhood, late childhood, adolescence, and adulthood.

This summarized data set is visualized with the comparative explorer in Figure 3.6**a**, i.e. genes are close together in the map if they behave similarly through time and anatomical regions simultaneously. It also shows gene maps for each developmental stage individually (Figure 3.6**b**), i.e. genes are close together in a map when they behave similarly across anatomical regions within that developmental stage. The comparative explorer gives insight in the transcriptional background of development. For example, Figure 3.6 shows that oligodendrocyte marker genes are spatio-temporally co-expressed, but before birth these genes are not co-expressed. In fact, these marker genes have a very low expression before birth, which reflects the fact that myelination is largely a post-natal process. The rise in expression of myelination related genes after birth has been observed before (Kang et al., 2011), and BrainScope's comparative explorer shows that this is also reflected in changes in co-expression over time.

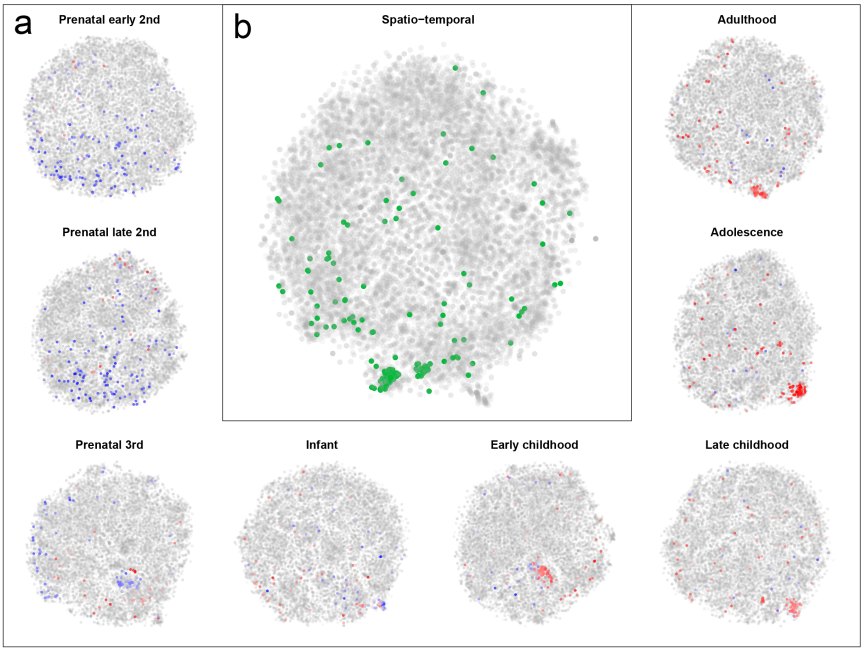


Figure 3.6: Developmental gene expression patterns show oligodendrocyte activity after birth: ***(a)*** *The gene t-SNE maps per developmental stage. These maps reflect spatial co-expression of genes at each stage of development from early second trimester to adulthood. Oligodendrocyte marker genes are highlighted by their average expression across the brain at that stage of development. The oligodendrocyte marker genes have a low pre-natal expression (*blue*) and, as a result, a weak co-expression. After birth these genes become more active (*red*), and more co-expressed, which reflects the formation of white matter after birth.* ***(b)*** *The spatio-temporal gene t-SNE map of gene expression. Genes cluster together if they have a similar expression pattern through developmental time and anatomical space. The highlighted genes (*green*) are those that are oligodendrocyte specific.*

3.4 Discussion

We present the BrainScope portal for interactive visual analysis of gene expression in the brain. Through the use of linked t-SNE maps both global and local patterns in the data can be elucidated. Specific cell types give rise to expres-

sion patterns, which can be explored in both the sample and gene map using cell type marker genes. Users can upload their own gene sets of interest to find the spatial expression and co-expression patterns in the healthy human brain. The fact that neighboring genes in the gene map reflect co-expression, and therefore possible functional links, means that genes can be studied in their co-expressional context. In addition, the comparative explorer for the adult brain allows for the assessment of inter-donor stability of co-expression. The maps show transcriptomic robustness over donors, in a similar manner to the widely-used WGCNA algorithm. Finally, the developmental comparative explorer captures transcriptional patterns through development and age. Taken together, BrainScope gives an instant overview of similarities of all genes and of all brain regions.

The non-linearity and focus on local neighborhood structure of t-SNE make it well-suited for the visualization of similarities between samples and between genes in 2D plots. A practical advantage of t-SNE is that it has only one main parameter, the perplexity value, which controls the relative size of the neighborhood that is taken into account. It performs better in separating co-expression gene modules than PCA.

The gene and sample maps in BrainScope are based on all samples, and hence are affected by the anatomical distribution and sampling density. In addition, the portal is genome-wide. This means users are likely to find their brain regions and genes of interest represented in the portal. A filtering of genes could, however, give a stronger signal for specific applications, and a selected gene set may provide tailored sample-sample relationships. In addition, the current maps are affected by the strong difference between cerebellar cortex and cerebral samples. Therefore, an extension to the portal would be the option to recalculate the t-SNE maps on a subset of samples or genes, in an interactive manner (Pezzotti et al., 2017). A user could select points based on prior knowledge or visual inspection of expression and update the maps. This would require more investment in server-side calculations.

Currently the portal contains only the gene expression data for the Allen Atlases of the adult human and developing human brain. However, the Allen Institute also provides spatial transcriptomic data for mouse (Lein et al., 2007), developing mouse (Thompson et al., 2014), macaque (Bernard et al., 2012), and developing macaque (Bakken et al., 2016). In addition to these large scale datasets by the Allen Institute, spatially resolved epigenetic data of the brain

is now available from the PsychENCODE project (Akbarian et al., 2015). The concepts of the BrainScope portal are applicable to these datasets as well. To illustrate this, we have applied the methodology to the spatial gene expression data of the UK Brain Expression Consortium (Ramasamy et al., 2014) (see Supplementary Text and Supplementary Figure S3).

The amount of data available to molecular neuroscientists is rapidly growing. The availability of increasingly high-dimensional data, even on a single-cell level, calls for visualization tools that can offer both a birds-eye view of the full data, and an entry point to formulating specific questions. Consequently, BrainScope is a valuable tool for neurologists to gain a deeper understanding of the interactions between brain anatomy and molecular function.

3.5 Acknowledgements

This research has received partial funding from the Dutch Technology Foundation STW, as part of the STW project 12721: “Genes in Space” and 12720: “VANPIRE”, under the IMAGENE STW Perspective program. The authors gratefully acknowledge Mike Hawrylycz of the Allen Institute for Brain Science.

— 4 —

A structural equation model for imaging genetics using spatial transcriptomics

Abstract

Imaging genetics deals with relationships between genetic variation and imaging variables, often in a disease context. The complex relationships between brain volumes and genetic variants have been explored both with dimension reduction methods and model based approaches. However, these models usually do not make use of the extensive knowledge of the spatio-anatomical patterns of gene activity. We present a method for integrating genetic markers (single nucleotide polymorphisms) and imaging features, which is based on a causal model and, at the same time, uses the power of dimension reduction. We use structural equation models to find latent variables that explain brain volume changes in a disease context, and which are in turn affected by genetic variants. We make use of publicly available spatial transcriptome data from the Allen Human Brain Atlas to specify the model structure, which reduces noise and improves interpretability. The model is tested in a simulation setting, and applied on a case study of the Alzheimer's Disease Neuroimaging Initiative.

This chapter has been published as: Huisman, S. M. H., Mahfouz, A., Batmanghelich, N. K., Lelieveldt, B. P. F., Reinders, M. J. T. (2018). A structural equation model for imaging genetics using spatial transcriptomics. *Brain Informatics, 5*(2), 13. `https://doi.org/10.1186/s40708-018-0091-0`
All supplemental materials can be found in the online publication.

4.1 Introduction

The aim of imaging genetics studies is to find associations between genetic variants and imaging features, often in a disease context (J. Liu & Calhoun, 2014). This scheme extends beyond traditional genome wide association studies (GWAS) by identifying genetic associations of imaging biomarkers with the assumption that these biomarkers are a more direct reflection of the genetic effects. Thus, they could provide a stronger association signal (Hibar, Kohannim, Stein, Chiang & Thompson, 2011). Additionally, the identified associations are likely to provide new insights into the underlying disease mechanisms as well as new hypotheses about the anatomical and/or functional locations involved in complex diseases (Franke et al., 2016).

So far, imaging genetics studies have been largely focused on the brain (Calhoun, Liu & Adalı, 2009; Franke et al., 2016; J. Liu & Calhoun, 2014; Stein et al., 2012; Vounou et al., 2012), despite efforts to extend their application to other fields (Batmanghelich, Saeedi, Cho, Estepar & Golland, 2015). Several large consortia have gathered data from thousands of subjects to understand the effects of genetic variants on brain structure and function (Medland, Jahanshad, Neale & Thompson, 2014). One of the hallmark sources for imaging genetics studies is the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Mueller et al., 2005). This database contains single nucleotide polymorphism (SNP) and structural MRI data for Alzheimer's patients, individuals with late mild cognitive impairment, and cognitive normal controls.

One of the largest challenges facing imaging genetics studies is the statistical power needed to identify reliable associations. In a typical GWAS, researchers have to correct for the number of independent tests performed (i.e. number of independent SNPs tested) in order to limit the number of false positive discoveries. However, a genome-wide brain-wide imaging genetic study will not only have to correct for the number of independent SNPs, but also for the number of independent imaging features tested. This requirement yields most of the studies underpowered to identify reliable associations. One of the largest imaging genetics studies (Hibar et al., 2015) analyzed over 30,000 individuals within the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA) consortium. They performed a genome wide association of SNPs with seven brain volumes, and identified only eight genome wide significant

SNPs.

Despite the high dimensionality of the imaging data (millions of voxels), the actual number of independent tests for which we need to correct in an imaging genetics study is far smaller than the number of voxels. Due to the spatial relationships between voxels, measurements from neighboring voxels are usually highly correlated. A common approach is to test genetic associations for anatomically defined brain regions (Hibar et al., 2011). Several studies have shown that both neuroanatomical parcellation and connectivity of the brain are strongly reflected in gene expression patterns across the brain (Hawrylycz et al., 2012; Ko et al., 2013; Richiardi et al., 2015). The public availability of brain transcriptome atlases from the Allen Institute for Brain Science (Sunkin et al., 2013) provides an opportunity to use these transcriptional signatures to group brain regions, limiting the number of effective tests.

Several methods have been proposed to identify associations between genetic variants and imaging features by applying dimension reduction, such as variations of canonical correlation analysis (Du et al., 2016), and independent component analysis (commonly used in a functional MRI context) (Calhoun et al., 2009). Others have opted to model the interactions between the different data types explicitly. Both Stingo, Guindani, Vannucci and Calhoun (2013) and Batmanghelich, Dalca, Quon, Sabuncu and Golland (2016) pose graphical Bayesian models which capture a more mechanistic causal view of the data. These models consist of a directed acyclic graph, which can easily be made to incorporate covariates, including possible confounding factors. Both studies use relatively small candidate SNP sets, because they aim for understanding SNP brain relationships rather than the discovery of genome wide associations. However, these Bayesian models are quite challenging to specify and fit.

In this work, we propose a method to identify associations between candidate genetic variants and imaging features allowing for the incorporation of prior knowledge. The proposed method combines a graphical model with dimension reduction to model the effect of SNPs on brain imaging features through a set of latent variables. We use a maximum likelihood structural equation modelling (SEM) approach to find the edge weights of our model (Bollen, 1989). By performing dimensionality reduction within the model, we reduce the number of parameters to be estimated. In addition, the model allows for easy incorporation of information from the Allen Human Brain Atlas (Hawrylycz et al., 2012) to inform the grouping of brain regions based on the

similarity of their transcriptional profiles.

Our model uses the transcriptional profiles for grouping because we consider gene expression to be an intermediate phenotype, that links SNPs to brain imaging features. Most disease associated SNPs are located near regulatory regions of the genome (Maurano et al., 2012), and the effects of SNPs on expression tend to be tissue and cell type specific (Ardlie et al., 2015). Gene expression data of brain regions reflects cell type composition and anatomical similarity (Hawrylycz et al., 2012), and captures a wide range of brain specific molecular pathways (S. M. Huisman et al., 2017). For these reasons, the region groups in the dimension reduction are based on spatial gene expression data of the brain.

4.2 Materials and methods

The interplay between genetic variation, brain anatomy, and disease symptoms is complex. We use a structural equation model with latent variables (Bollen, 1989) to model these relationships. We pose that the genetic variation is exogenous, in other words: the genetic variation in a study population is not caused by disease or brain anatomy. This variation does have an effect on the brain. For example, in Alzheimer's disease, genetic variants may influence the immune response and amyloid β concentrations in the brain, which may in turn lead to shrinkage in several brain areas (Bettens, Sleegers & Van Broeckhoven, 2013). Large scale imaging initiatives, such as ADNI, offer a possibility to study this shrinkage of brain regions. This can be estimated from MRI data of diseased individuals and controls, and expressed in cortical thickness and subcortical volume measurements.

In our graphical model, we define groups of brain regions, based on the transcriptional profiles of these areas in the healthy brain. Areas that share patterns of gene expression in a normal brain may be similarly affected by genetic variations. For each of the region groups, we introduce one latent variable. This latent variable is affected by the genetic variations, and causes changes in relevant brain regions. This makes our model similar to principal component analysis (PCA) on sets of brain regions, combined with a regression for the latent variables. However, in our model the weights are estimated together, and the latent variables reflect not only the correlations between the

regions (as in a conventional PCA), but also those between regions and SNPs and among the SNPs.

4.2.1 Variables used

We model the relationship between single nucleotide polymorphisms (SNPs) and brain region measurements. Let $\mathbf{g}_i \in \mathbb{R}^p$ be a vector of centred (zero-mean) SNP values (originally coded as $0, 1, 2$), and $\mathbf{x}_i \in \mathbb{R}^q$ a vector of centred (zero-mean) and scaled (sd = 1) brain region measurements, both for individual i. The reason both types of measurements are centred, is to eliminate intercepts from the model. The brain measurements are, in addition, scaled to unit variance to compensate for the considerably larger variance in thickness or volume for larger brain areas. The genetic variants and brain measurements are connected in the model by a set of latent variables, $\mathbf{z}_i \in \mathbb{R}^m$.

In addition to the variables included in the model, we have two other sources of information. In defining the model structure, we make use of external information on the brain region measurements, in the form of brain region groups with a shared transcriptional profile. These brain regions can be defined based on spatial gene expression data of the healthy adult brain. Finally, the goal is to understand disease related phenotypes. The disease labels are not used in the modelling stage. However, we hypothesize that if the variation in the data is related to a disease state, the latent variables will reflect this. After model fitting, we therefore associate each individual's estimated latent variable score to his or her disease status.

4.2.2 The graphical model

We model the relationship between brain SNP values and brain region measurements in a structural equation model (SEM). It consists of two parts. The first part is a linear model for brain region measurements as a function of the latent variables,

$$\mathbf{x}_i = \mathbf{B}\mathbf{z}_i + \boldsymbol{\zeta}_i, \tag{4.1}$$

where $\mathbf{x}_i$ contains the observed brain region measurements, $\mathbf{z}_i$ the latent variables, and $\boldsymbol{\zeta}_i$ is a zero-mean normally distributed error variable. The matrix $\mathbf{B}$ contains the weights of the latent variables that explain the brain

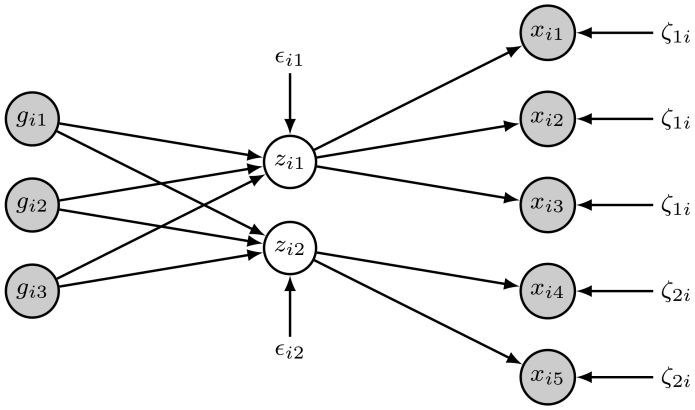


Figure 4.1: The graphical structural equation model. Observed variables are shown in grey circles, latent variables in white circles, and error variables without circles. This example contains two latent variables, both with their own set of observed brain region measurements. This structure, where the latent variables define groups of region measurements, is defined by prior knowledge on these brain regions. We use spatially resolved gene expression data of the healthy human brain to define these region groups.

region measurements. The second part of the SEM is a linear model for these latent variables as a function of the SNP values,

$$\mathbf{z}_i = \mathbf{A}\mathbf{g}_i + \boldsymbol{\epsilon}_i, \tag{4.2}$$

where $\mathbf{g}_i$ contains the observed SNP measurements, and $\boldsymbol{\epsilon}_i$ is a zero-mean normally distributed error variable. The matrix $\mathbf{A}$ contains regression weights, representing the effects of the SNPs on the latent variables. Combined, these equations mean that region changes are viewed as a manifestation of the latent values, while the SNP values are considered causal to them. The latent variables represent some intermediate phenotype, related to the molecular state of the connected brain regions.

The number of latent variables is equal to the number of brain region groups, which are defined based on external spatial gene expression data. A region group contains the brain regions with a similar transcriptional profile, as these may react similarly to differences in genetic background. We restrict

each latent variable to only predict the brain region measurements for its own region group. This results in a restriction on the weight matrix $\mathbf{B}$, where each latent variable (corresponding to a column in $\mathbf{B}$) has a unique set of non-zero entries. Fig. 4.1 shows the model for two latent variables, where we can see that each latent variable is connected to its own set of brain regions.

4.2.3 Model implied covariance

In linear Gaussian structural equation modelling, we learn the parameters of a model by optimising the correspondence between the observed covariance $\mathbf{S}$ (from the data), and the model implied variance Σ. The model implied variance can be divided in a block matrix, by defining

$$\Sigma = \begin{bmatrix} \Sigma_{gg} & \Sigma_{gx} \\ \Sigma_{xg} & \Sigma_{xx} \end{bmatrix}.$$

Note that this implied covariance does not contain any components for the latent variables in $\mathbf{z}$. The latent variables are not observed, and therefore we cannot use their observed covariance in fitting the model.

The elements of the implied covariance can be parametrised in terms of the model coefficients. The first element is

$$\Sigma_{gg} = \mathbf{E}[\mathbf{g}\mathbf{g}^T].$$

This is the covariance of the SNPs (since these values are centred). The SNPs are exogenous in our model: $\mathbf{g}$ does not have any causal variables within our model. As a result, the implied covariance of the SNPs is not parametrised in terms of model coefficients. We can estimate this covariance term simply by taking the observed covariance between the SNPs.

The next element of the implied covariance matrix is

$$\begin{aligned} \Sigma_{xg} &= \mathbf{E}[\mathbf{x}\mathbf{g}^T] \\ &= \mathbf{E}[\mathbf{B}\mathbf{A}\mathbf{g}\mathbf{g}^T + \boldsymbol{\epsilon}\mathbf{g}^T + \boldsymbol{\zeta}\mathbf{g}^T], \end{aligned}$$

and similarly

$$\begin{aligned} \Sigma_{gx} &= \mathbf{E}[(\mathbf{x}\mathbf{g}^T)^T] \\ &= \mathbf{E}[\mathbf{g}\mathbf{g}^T\mathbf{A}^T\mathbf{B}^T + \mathbf{g}\boldsymbol{\epsilon}^T + \mathbf{g}\boldsymbol{\zeta}^T]. \end{aligned}$$

The final element of the implied covariance matrix is the model implied covariance among the brain regions. This is given by

$$\begin{aligned}\Sigma_{xx} &= \mathbf{E}[\mathbf{x}\mathbf{x}^T] \\ &= \mathbf{E}[\mathbf{B}\mathbf{A}\mathbf{g}\mathbf{g}^T\mathbf{A}^T\mathbf{B}^T + \mathbf{B}\mathbf{A}\mathbf{g}\boldsymbol{\epsilon}^T\mathbf{B}^T + \mathbf{B}\mathbf{A}\mathbf{g}\boldsymbol{\zeta}^T \\ &\quad + \mathbf{B}\boldsymbol{\epsilon}\mathbf{g}^T\mathbf{A}^T\mathbf{B}^T + \mathbf{B}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{B}^T + \mathbf{B}\boldsymbol{\epsilon}\boldsymbol{\zeta}^T \\ &\quad + \boldsymbol{\zeta}\mathbf{g}^T\mathbf{A}^T\mathbf{B}^T + \boldsymbol{\zeta}\boldsymbol{\epsilon}^T\mathbf{B}^T + \boldsymbol{\zeta}\boldsymbol{\zeta}^T].\end{aligned}$$

4.2.4 Model assumptions and estimation

Some elements of the implied covariance are often assumed to be zero. These assumptions lead to a strong simplification of the implied covariance. It is common in a regression setting to pose that the predictor variables and error variables are independent. In our case, the error independence assumption leads to $\mathbf{g}\boldsymbol{\epsilon}^T = \boldsymbol{\epsilon}\mathbf{g}^T = 0$. In addition, we assume that the errors in the brain region predictions (equation (4.1)) are independent of the errors in the latent variable predictions (equation (4.2)). This means that $\boldsymbol{\zeta}\boldsymbol{\epsilon}^T = \boldsymbol{\epsilon}\boldsymbol{\zeta}^T = 0$. Finally, we assume that the errors in brain region prediction are independent of the SNPs, so $\mathbf{g}\boldsymbol{\zeta}^T = \boldsymbol{\zeta}\mathbf{g}^T = 0$.

As a result of these assumptions, the full implied covariance matrix of the model reduces to

$$\begin{bmatrix}\Sigma_{gg} & \Sigma_{gx} \\ \Sigma_{xg} & \Sigma_{xx}\end{bmatrix} = \mathbf{E}\begin{bmatrix}\mathbf{g}\mathbf{g}^T & \mathbf{g}\mathbf{g}^T\mathbf{A}^T\mathbf{B}^T \\ \mathbf{B}\mathbf{A}\mathbf{g}\mathbf{g}^T & \mathbf{B}\mathbf{A}\mathbf{g}\mathbf{g}^T\mathbf{A}^T\mathbf{B}^T + \mathbf{B}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{B}^T + \boldsymbol{\zeta}\boldsymbol{\zeta}^T\end{bmatrix}. \tag{4.3}$$

For normally distributed data, the maximum likelihood estimate of the covariance matrix is

$$\max_{\Sigma}(-\log(|\Sigma|) - \mathrm{tr}(\mathbf{S}\Sigma^{-1})), \tag{4.4}$$

where is $\mathbf{S}$ the observed covariance matrix. The SNP data we use is discrete, and can therefore not be considered normally distributed. To compensate for this, we will estimate robust standard errors. In equation (4.4), the covariance Σ is parametrised according to equation (4.3), so we can perform the optimisation over the parameter values.

Model fitting is performed in the *lavaan* package in *R* (Rosseel, 2012). For identifiability, we fix the loading of the first brain region measurement per region group (latent variable) to 1. This does not only fix the scales of the latent variables, but it also has the advantage that the resulting latent variables will have the same direction of effect as the first brain region measurement. For example, a reduction in volume of the first brain region will result in a reduction in the corresponding latent variable. All the error variances on the brain region measurements (variance of $\boldsymbol{\zeta}$) are assumed to be equal, which is the same as in principal component analysis.

The model fit in *lavaan* yields estimates for $\mathbf{B}$, $\mathbf{A}$, and the covariance matrices of the error variables $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$. Each of these parameter estimates is provided with robust p-values (for the hypothesis of being equal to zero), when using the *MLM* estimation procedure (Rosseel, 2012). Using the estimated model parameters, one can then calculate unbiased Bartlett scores for the latent variables (Distefano, Zhu & Mîndrilă, 2009).

4.2.5 Data

Simulated data The model is evaluated both on simulated and real data. In the simulation, we first generated SNP values ($\mathbf{g}_i$) in accordance with Hardy-Weinberg equilibrium. The minor allele frequencies were independently drawn from a beta distribution with shape parameters $\alpha = 1$ and $\beta = 2$. Then we simulated latent variables ($\mathbf{z}_i$) as a linear combination of the SNP values, with Gaussian noise ($sd = 2$). Each of these latent variables determined the region measurements ($\mathbf{x}_i$) of a set of regions (a region group), with added Gaussian noise ($sd = 2$). This part of the simulation is in line with equations (4.1) and (4.2) and Fig. 4.1. Finally, we used a logistic model in which a linear combination of some of the latent variables determined the probability of observing a phenotype. These binary phenotypes (disease versus healthy) were then drawn from a Bernoulli distribution.

We simulated 100 independent datasets for 500 individuals. Each time, we set the number of SNPs to 20 and the number of latent variables (and therefore region groups) to 5. We randomly selected 10 SNP-to-latent weights ($\mathbf{A}$) to be either 1 or -1. The 5 region groups contain 20, 10, 10, 5, and 5 regions respectively, for a total of 50 brain region measurements. Each latent variable has latent-to-brain-region weights (in $\mathbf{B}$), which were uniformly

sampled between 0.5 and 1.5. All other elements of $\mathbf{B}$ were set to zero, which effectively restricts each latent variable to affect only its own region group. Finally, two out of the five latent variables were randomly selected to affect the disease probability, with weights of either 10 or -10. All other latent-to-phenotype weights were set to zero.

To test the robustness of our method, we also simulated data for a range of alternative parameter settings. We varied the amount of noise in the latent variables ($\mathbf{z}_i$) and the region measurements ($\mathbf{x}_i$) between 1 and 5. The number of non-zero SNP-to-latent weights (in $\mathbf{A}$) was varied from 2 to 20. Finally, we constructed data sets with misspecified latent-to-brain-region weights (in $\mathbf{B}$). For this end, we swapped links between latent variables and regions. In each swap, a region was disconnected from its original latent variable and instead connected to another latent variable. To retain the sizes of the regions groups, another region of that second latent variable was then connected to the first latent variable. Each swap therefore resulted in two misspecified links. We made sure not to swap regions back to their original latent variables.

ADNI data and preprocessing The real data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`) (Mueller et al., 2005). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see `www.adni-info.org`.

The ADNI database contains measurements on a large number of cognitive normal (CN) controls, individuals with late mild cognitive impairment (LMCI), and individuals with Alzheimer's disease (AD). The measurements in the database include patient demographics, raw and processed MRI data, biomarker data and SNP data. For the brain volumes we made use of the UCSF cross-sectional FreeSurfer (Version 4.3) cortical thickness and white matter parcellation measurements. For the SNPs we made use of the ADNI 1 Illumina Human 610-Quad BeadChip data, with imputation as previously described (Batmanghelich et al., 2016). In the end, we selected volumes, SNPs

and diagnoses for 746 individuals. This data was split in two equal parts of 373 individuals, one as a training and one as a validation, to prevent over-fitting in the modelling process.

Our methodology is not suited to genome wide analysis. Instead, it tries to find the effects of specific SNPs on a set of latent variables. As candidate SNPs we selected a set of 35 polymorphisms associated with Alzheimer's disease according to the International Genomics of Alzheimer's Project (IGAP) study results (Lambert et al., 2013). IGAP is a two-stage GWAS on individuals of European ancestry for Alzheimer's disease. In stage 1, IGAP used genotyped and imputed data on 7,055,881 SNPs of 17,008 Alzheimer's disease cases and 37,154 controls. In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1&2. We selected the known SNPs, stage 1 discoveries, and stage 1&2 discoveries from table 2, and the suggestive SNPs from supplemental table 4 of Lambert et al. (2013).

The volume data was present for 112 regions. We corrected it for individual age, gender, and whole brain volume (using linear regression), with the goal of maintaining all meaningful variation in brain region volumes, possibly related to the disease phenotype. For our latent variable model, the brain regions volumes were linked to region groups. We defined these regions groups based on the transcriptional profiles in the healthy adult human brain, as provided by the Allen Atlas (Hawrylycz et al., 2012). This gene expression resource contains anatomically labelled measurements taken from six human brains. Regions with measurements in each of the six brains were selected, and the expression values were averaged to obtain a single value for each of the 19 992 genes in each of the 105 regions of the Allen Atlas (S. M. Huisman et al., 2017). We then performed a t-distributed neighbourhood embedding (t-SNE) analysis to obtain a two-dimensional map of the brain regions. Brain regions are placed nearby in this map if they have a similar expression profile across all genes. This map was then used to manually define nine groups of brain regions, as is shown in figure 2 of S. M. Huisman et al. (2017). The regions of the ADNI data were manually linked to the nine region groups, as shown in Supplemental Table 1. The anatomical atlas used for the Allen Atlas is hierarchical: it has a tree-like structure with large regions containing smaller regions. Table 4.1 shows a higher level description of the regions in the nine

region groups. In most cases the Allen Atlas regions were more general (larger) than the FreeSurfer regions of the ADNI data. Out of the 112 regions, 105 regions were linked to a region group, while the other 7 regions did not have corresponding samples in the Allen Atlas data and were therefore left out.

Table 4.1: The nine region groups (corresponding to the latent variables), with the brain regions they contain. These are higher level labels of the Allen Atlas (Hawrylycz et al., 2012). A full subdivision of the ADNI FreeSurfer regions into these region groups is provided in Supplemental Table 1.

Region group code	ABA region
CrCortex	cingulate gyrus
CrCortex	frontal lobe
CrCortex	insula
CrCortex	middle frontal gyrus
CrCortex	occipital lobe
CrCortex	parahippocampal gyrus
CrCortex	parietal lobe
CrCortex	temporal lobe
Hippocam	hippocampal formation
Amygdala	amygdala
Striatum	striatum
DorsThal	dorsal thalamus
SubCort1	myelencephalon
SubCort2	globus pallidus
SubCort2	white matter
ClCortex	cerebellar cortex
SulcSpac	sulci & spaces

4.3 Results

4.3.1 Simulation

To evaluate the performance of our model, the SEM was fitted to each of the simulated datasets. We considered two measures for model comparison. First, we set out to assess the prediction of phenotypes (disease status) from the latent variables, with a logistic regression. In each of the 100 simulated datasets, we estimated the latent variable scores, and used only those to predict the phenotype. For each of these 100 models, we obtained an Akaike information criterion (AIC) value. We compared our model to several logistic regression models that use only the simulated data, instead of the SEM estimated latent scores. The first alternative model uses only all the brain region measurements, the second only all the SNP measurements, and the third a combination of all regions and SNP measurements. As a fourth alternative model, we performed a PCA on the volume measurements, and extracted the first five principal components. Fig. 4.2 shows that, on average, our latent variables obtain a lower AIC than models using either all brain region data, all SNP data, or both. The model using the first five principal components of the brain region data is most similar to our model, and it only has a slightly lower AIC on average than our model.

The second measure for model comparison is the ability to retrieve the correct SNPs. In each of our simulation datasets, two of the five latent variables have an effect on the phenotype (disease status). All SNPs that affect either of these two latent variables effectively impact the phenotype. We consider those SNPs to be the SNPs with a true effect. We now consider how these SNPs are ranked for importance in our SEM analysis, and two alternative approaches. From our SEM fit, we extracted the robust SNP p-values for predicting the latent variables (so the p-values for the estimates in $\mathbf{A}$). These give an impression of the importance of a SNP in predicting the latent variables. In addition, we used the latents' logistic regression p-values for the phenotype. These show the importance of a latent variable in predicting the phenotype. As a result, the path from a SNP to the phenotype contains two p-values per latent variable: one for the latent variable prediction, and one for the phenotype prediction.

We considered combining these p-values in two ways: 1) for each SNP

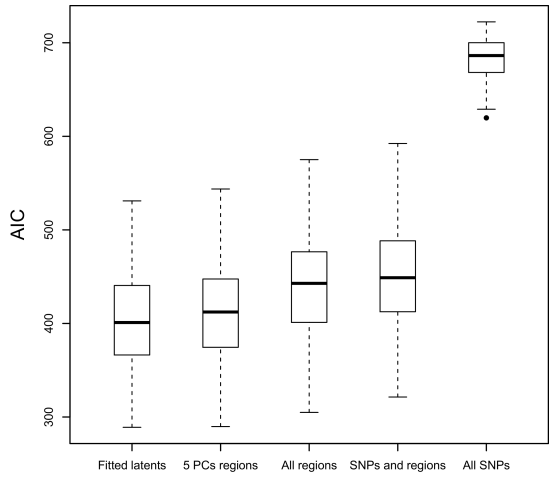


*Figure 4.2: Simulation AIC model fit. The logistic regression models use either the SEM estimated latent variable scores (*Fitted latents*), the first five principal components of the brain region data (*5 PCs regions*), all brain region data, all SNP and brain region data, or all SNP data.*

we took the maximum p-value of the two per latent variable, and then the minimum p-value over the five latent variables; or 2) for each SNP we used Fisher's method (R. Fisher, 1950) to combine the two p-values per latent variable ($-2\sum\log(p_i)$), and then took the minimum p-value over the five latent variables. Note that Fisher's method is meant for p-values testing the same null-hypothesis, which is not the case here. Both methods yield a score (p-value) for SNP importance. We varied a threshold for this score from 0 to 1 and compared the set of SNPs with values below this threshold to the set of SNPs with a known true effect. In this way, we constructed a receiver operating characteristic curve for SNP retrieval, and calculated the corresponding area under the curve (AUC).

We compared the performance of our methodology to a straightforward modelling approach: a logistic regression to predict the disease status phenotype from the SNPs. This was performed both in a univariate way (as in a GWAS), and a multivariate way. Fig. 4.3 shows the performance of our SEM based methods, using the maximum p-value per SNP-latent combination (*SEM max*) or using Fisher's method (*SEM Fisher*), and of the GWAS-like

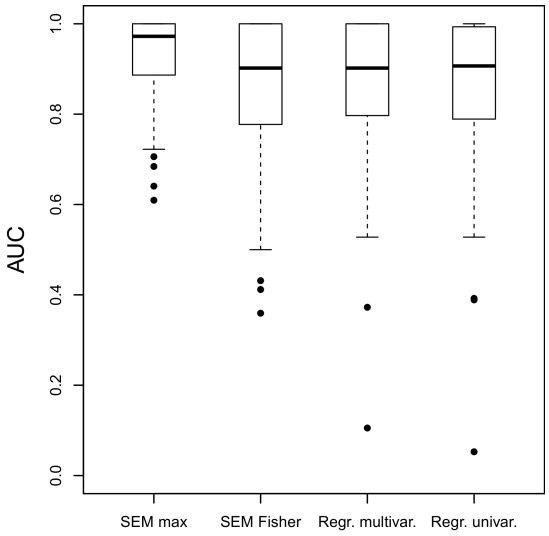


*Figure 4.3: Simulation AUC for SNP selection. Shown are the results for two methods of p-value integration for our model (*SEM max *and* SEM Fisher*), for multivariate logistic regression, and univariate logistic regression. A high AUC means that the method correctly ranks the importance of the SNPs for the phenotype (disease state).*

approaches. The *SEM max* method has the highest average AUC, indicating that it is best able to rank the SNPs on their importance for the phenotype. Note that the *SEM Fisher* method has the disadvantage that either a strong SNP-to-latent or a strong latent-to-phenotype effect can lead to a low combined p-value, regardless of the other value. The observed difference between the univariate an multivariate approach is very small, which is to be expected since the simulated SNP values are independent.

To test the robustness of our model, we also compared the models for a range of alternative simulation settings. Supplemental Fig. 2 shows the results of these simulations. The amount of noise on the latent variables has a similar impact on all compared methods. With a large amount of noise on the brain region measurements, the prediction of phenotypes remains best with our model, but the identification of SNPs is better with methods that do not make use of this region volume data. The number of SNPs with a non-zero effect on the latent variable has little impact on the simulation results.

Misspecification of the region groups, on the other hand, has a negative impact specifically on the performance of our method. This shows that our approach is somewhat sensitive to the specification of brain region groups.

4.3.2 ADNI application

We apply our methodology to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data (Mueller et al., 2005). We selected 35 SNPs and 105 brain region volumes for 746 individuals. The brain regions were divided into nine region groups based on the gene expression patterns of matching brain areas in the healthy human brain (Hawrylycz et al., 2012; S. M. Huisman et al., 2017). Each of the nine brain region groups has one corresponding latent variable, and each latent variable has a unique set of brain region measurements attached to it. Supplemental Fig. 1 shows the volume loadings for each of the latent variables. Since the first loading for each latent variable is set to 1, the latent variables will have the same direction of effect as this variable. All but two of the region volumes have a positive loading. Two regions in the subcortical group 2 (*SubCort2*) are negatively correlated to the latent variable scores, reflecting a more heterogeneous signal in this group.

Fig. 4.4 shows the association between the nine latent variables and the selected SNPs. Only those SNPs are shown that have a nominally significant ($p < 0.05$) association with at least one of the latent variables. After correction for multiple testing, the only significant effect is that for rs429358, located in *APOE*, on the hippocampal region group (Bonferroni corrected $p = 2.28 \cdot 10^{-4}$). In the validation set, here used as a replicate, this effect was again significant (Bonferroni corrected $p = 8.66 \cdot 10^{-3}$). None of the other associations are significant after multiple testing correction. This *APOE* allele is known to be associated with a decrease in the hippocampal volume, both in individuals with mild cognitive impairment (Farlow et al., 2004) and in Alzheimer's disease (Schuff et al., 2009).

The latent variables reflect differences in brain region volumes across the ADNI dataset. To test whether these differences in brain region volumes are related to the disease phenotype, we compared the latent variable scores between the CN, LMCI, and AD individuals. Fig. 4.5 shows the distribution of latent variable scores for the validation set. To calculate these, we used the fitted SEM of the training data, and used its parameter estimates to cal-

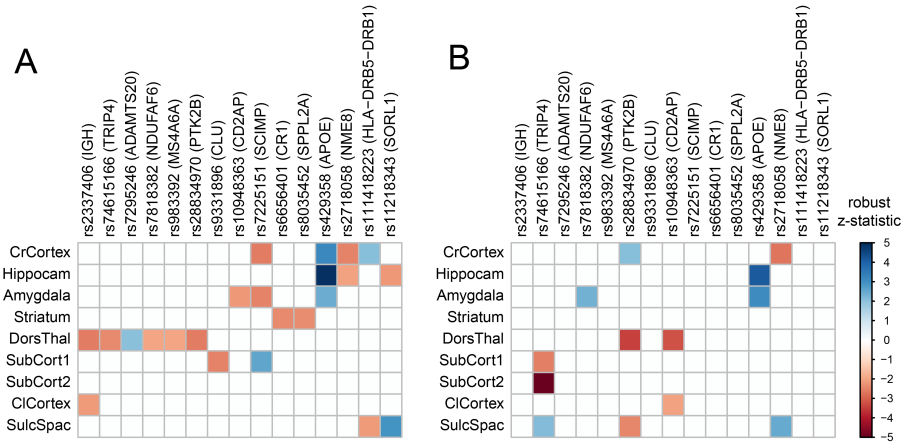


Figure 4.4: Association between SNPs and latent variable scores, as found by the robust maximum likelihood fit of the SEM. All nominally significant associations ($p < 0.05$) are coloured by their robust z-statistic values (Rosseel, 2012). The linked genes (Lambert et al., 2013) are shown in brackets. ***(A)*** *The results for the training set. After Bonferroni correction for the 315 tests, only the effect of rs429358 (*APOE*) on the hippocampus region group remains significant.* ***(B)*** *The validation results confirm the significant effect of rs429358 (*APOE*) on the hippocampus region group.*

culate latent variable scores for the validation data. For three region groups the latent variable scores were significantly lower in LMCI than in controls, and even lower in AD. These regions are the cerebral cortex, the hippocampal formation, and the amygdala. This reflects significant shrinkage in these areas during Alzheimer's disease progression. The region group of sulci and spaces (SulcSpac) has a latent variable that significantly increases in LMCI and AD. The significant association between the SNP rs429358 and the latent variable scores for hippocampus reflects the importance of *APOE* for Alzheimer's disease.

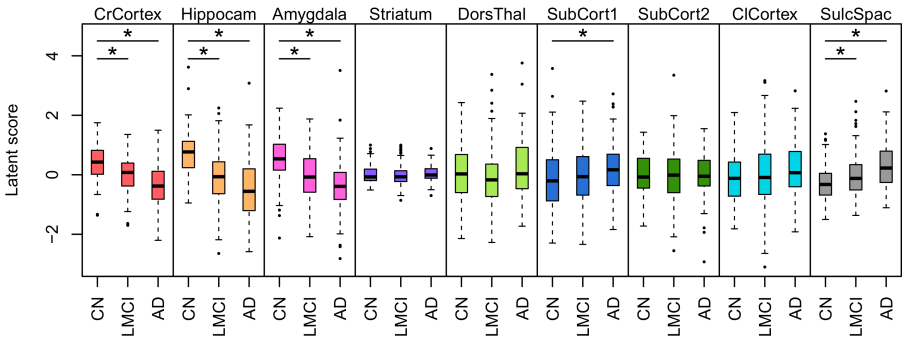


Figure 4.5: Association between the validation latent variable scores and diagnosis. Diagnosis is cognitive normal (CN), late mild cognitive impairment (LMCI), or Alzheimer's disease (AD). Nominally significant differences (ANOVA $p < 0.05$) are indicated with asterisks. The cerebral cortex (CrCortex), hippocampus (Hippocam), and amygdala (Amygdala) latent volume variables are lowered with disease progression, while the latent variable score for sulci and spaces (SulcSpac) is increased.

4.4 Conclusion

We have proposed the use of a maximum likelihood structural equation model for combining SNP data and structural brain area measurements. The model makes use of external gene expression data, to define groups of brain regions that may respond similarly to genetic variation. For each of these region groups, we define a latent variable, which captures the relationship between the regions in a group and genetic variation. We have applied the model on a simulated dataset, to show it can capture disease relevant variation and identify causal SNPs. In addition, we have applied the model to the ADNI dataset, containing Alzheimer's patients, individuals with late mild cognitive impairment, and cognitive healthy controls. One SNP, linked to *APOE*, shows a reproducible significant relationship to the latent variable that captures hippocampal volume change. This latent variable, and that of the cerebral cortex, amygdala, and sulci & spaces also significantly associate with the disease diagnosis. This shows that our approach can be used to integrate several data types, and yield interpretable results.

The fitting process of the structural equation model has relatively high

computational cost. It is truly multivariate, which makes it infeasible at the moment to perform genome-wide analysis. It does have advantages for incorporating a large number of variables, since it allows for straightforward inclusion of constraints on the parameter estimates (Rosseel, 2012). With a constraint on the sum of squared weights, one could for instance implement a ridge regression. In addition, the model allows for the inclusion of additional data. This can be done either in the specification of the model structure, as we have done for the region groups, or by adding observed variables to the model. In our model, we chose to group brain regions based on the similarity of their expression profiles in the healthy brain. An interesting extension to the model would be to incorporate a layer of latent variables to reflect a grouping of the SNPs. These groups could also be based on the similarity of the brain-wide expression patterns of the associated genes.

These results show that maximum likelihood SEM is a versatile approach for data integration, which can be used to elucidate the relationships between genetic variation, structural brain phenotypes, and brain disease.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private

sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding was provided by the Dutch Technology Foundation STW, as part of the STW project 12721: *Genes in Space* under the *ImaGene* STW Perspective Program, and from the European Union Seventh Framework Programme (FP7/2007-2013) under Grant Agreement 604102 (Human Brain Project).

— 5 —

A spatial transcription based model search in schizophrenia imaging genetics

Abstract

The main goal in neuroimaging genetics studies is to find the relationships between genetic markers and brain imaging measurements. To do this, they have to deal with a large number of variables on both the genetic and the imaging side. This leads to challenges in multiple testing correction, and in the interpretation of results. We propose to use brain-specific prior information on the activity of genes, to find a set of promising multivariate models that predict imaging measurements from single nucleotide polymorphism (SNP) data. First, the prior information is used to obtain biologically meaningful cross-clusters of genes and brain regions, and then these are utilised in a Bayesian analysis of SNP and MRI measurements from a schizophrenia case study.

5.1 Introduction

Neuroimaging genetics studies try to find relationships between genetic markers and image derived measurements of the brain (J. Liu & Calhoun, 2014). These brain measurements can be based on functional or structural magnetic resonance imaging (MRI) scans for instance. The analyses can aim to study normal brain function or to find out what goes wrong in a disease context. If the analyses are genome-wide, this presents a practical challenge. If every variable, i.e. genetic variation, is tested for association, the statistical multiple testing burden is high. Therefore, genome-wide studies require large sample sizes. For polygenic phenotypes the required number of samples to detect

small univariate effects ranges from tens of thousands, to several orders of magnitude more if the minor allele frequencies are low (Visscher et al., 2017). In genome-wide and brain-wide imaging genetics studies this problem is even bigger. On the one hand, the number of univariate tests to consider now also grows with the number of brain measurements. On the other hand, it is more time consuming and expensive to obtain a large sample size if individuals' brains are to be scanned.

In addition to this lack in statistical power, standard genome wide association studies (GWAS) are univariate. They therefore ignore the inherent additive effect that variations in functional groups of genes have. These groups of genes can be specific pathways or very large collections (Boyle, Li & Pritchard, 2017). To address these challenges, several methodologies have been proposed. Partial least squares regression (Beaton, Kriegsman, Dunlop, Filbey & Abdi, 2016; Bouhaddani et al., 2016; Le Floch et al., 2012), parallel independent component analysis (Pearlson, Liu & Calhoun, 2015) and sparse canonical correlation analysis (Du et al., 2016) all assume some latent space in between the genetic and imaging features. These latent variables can also be modelled in a Bayesian network (Bouhaddani et al., 2016; Chekouo, Stingo, Guindani & Do, 2016) or a structural equation model (S. M. H. Huisman, Mahfouz, Batmanghelich, Lelieveldt & Reinders, 2018). The methods proposed in these studies may use candidate gene sets obtained from other studies to reduce the number of considered features. However, they usually make no use of prior knowledge on gene-gene or brain region-region similarities. Here, we propose an alternative approach, where we use external data of gene expression in the brain to search for promising Bayesian multivariate linear regression models.

We will apply our method on data of the Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) consortium (Blokland et al., 2017). Schizophrenia has a high heriability, with estimates varying from 41 to 87% (Chou et al., 2017; Hilker et al., 2017), and it involves a large number of genes (Kavanagh, Tansey, O'Donovan & Owen, 2015), which are part of a range of molecular pathways (C. Liu et al., 2017). In addition, the sizes of several subcortical brain regions have been found to differ between schizophrenia patients and healthy controls, with schizophrenia patients having for instance smaller hippocampus, amygdala and thalamus, and a larger pallidum (Haijma et al., 2013; Van Erp et al., 2016). The links between genetic variants and brain region measurements in schizophrenia are less clear (Franke et al.,

2016).

We developed a methodology to find multivariate linear regression models that predict brain region volume and thickness measurements from single nucleotide polymorphisms (SNPs). We set out to find the specific combinations of brain regions and SNPs where a linear multivariate relationship can be found. The space of possible models is far to large to analyse extensively, so we propose models that have an a-priori interest. For this end, we analyse gene expression data of the healthy adult human brain (Hawrylycz et al., 2012). Co-expression of genes across the brain reflects functional similarities between these genes, specially for brain specific molecular processes and phenotypes (Hawrylycz, Sunkin & Ng, 2015). On the other hand, transcriptional similarities between brain regions reflect anatomical and functional similarity (Mahfouz et al., 2015; Richiardi et al., 2015). Therefore, we performed a cross-clustering, making groups of genes and brain regions simultaneously with a consistent expression pattern. For each of the cross-clustering partitions in the gene expression data, we fitted a single regression model in the imaging genetics data. As a result, we explore an informative subspace of all possible models, and find groups of genetic variants that predict brain region measurements in a schizophrenia setting.

5.2 Methods

We modelled the effects of genetic variation on the volumes or thicknesses of brain regions in the context of the GENUS schizophrenia study. This was done in a multivariate way, where multiple SNPs can effect a part of the brain together. To find out which sets of variations have a combined effect, we linked all variations to genes and sampled groups of genes based on their co-expression in the healthy human brain. At the same time, we grouped brain regions into sets of regions with a similar transcriptional profile in the healthy brain. Since the most biologically relevant grouping of brain regions may depend on the genes of interest, we performed a cross-clustering (D. Li & Shafto, 2011). This cross-clustering divides a matrix of gene expression in the brain into partitions (cross-clusters) of genes and and brain regions. In a cross-clustering, each set of genes can have a unique subdivision of brain regions.

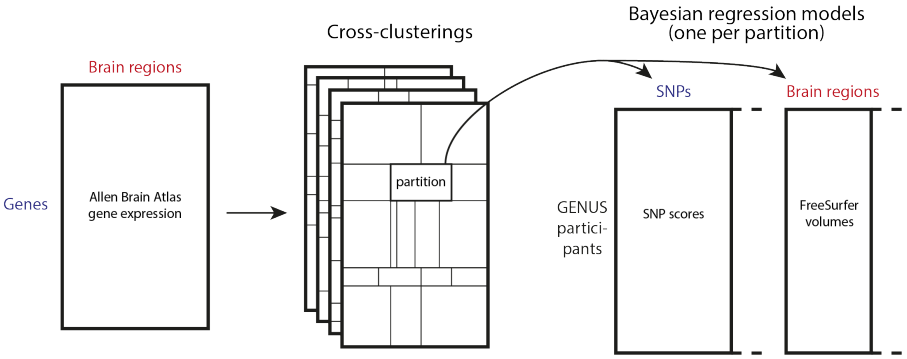


Figure 5.1: The Allen Brain Atlas gene expression data forms the basis of a cross-clustering of genes and brain regions. We sampled 385 cross-clusterings (4 shown in the figure), each of which provides a full partitioning of the gene expression matrix. For each of the partitions (in each cross-clustering) we selected the linked SNPs and FreeSurfer regions from the GENUS data to fit a Bayesian regression model.

A single cross-clustering cannot capture the complexity of gene expression in the brain. Genes may play a role in several pathways at the same time, in a complex network. As a result, a single partitioning of all genes does not contain all the meaningful groups of genes. To deal with this complexity, we repeated the cross-clustering a number of times (385) to obtain a distribution of cross-clusterings. Each clustering consists of a full partitioning of all genes and brain regions. Genes that cluster together in one clustering, may not do so in another clustering. The cross-clustering method (D. Li & Shafto, 2011) provides a likelihood for each clustering, which reflects the extent to which it can describe the gene expression data.

For each partition (cluster) of the Allen Brain Atlas data, we fitted a single Bayesian regression model on the GENUS data. Each gene was linked to SNPs that were either identified as one of that gene's expression quantitative trait loci (eQTLs) or located inside that gene. Each brain region of the Allen Brain Atlas was manually linked to its corresponding FreeSurfer region. By applying these mappings, a partition of genes and brain regions in the Allen Brain Atlas can be translated into a set of SNPs and a set of brain regions in the GENUS data. These GENUS subsets were then used in the Bayesian regression models. Figure 5.1 gives a graphical overview of this methodology.

5.2.1 Allen Brain Atlas data

Our model makes use of gene expression measurements of the healthy human brain from the Allen Institute for Brain Science (Hawrylycz et al., 2012). This dataset, the Allen Brain Atlas, consists of genome-wide micro-array measurements of 3 702 samples taken from across six postmortem adult human brains. These donor brains were selected to be unaffected by disease or substance use. The number of samples differed per brain, but in total 105 regions were sampled at least once in each of the six brains. Preprocessing of the microarray data was performed by the Allen Institute. We averaged the expression values across probes and samples to a transcription matrix containing 19 992 genes and 105 brain regions, as described in S. M. Huisman et al. (2017). This matrix of spatial gene expression was used as an input to the cross-clustering algorithm (D. Li & Shafto, 2011) to obtain partitions of genes and brain regions.

5.2.2 GENUS data

The schizophrenia data were obtained from the Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) consortium, which contains data from 19 different studies (Blokland et al., 2017). We included measurements of 1 192 healthy controls and 781 diagnosed schizophrenia patients, see Table 5.1. The imaging genetics models in this study were fitted on the data of these individuals. In fact, each model was fitted on the *all* dataset of 1 973 individuals, on the smaller set of *cases* only and on the *controls* only.

The individuals in the GENUS data had SNP measurements, with quality control and imputation as in Blokland et al. (2017). In addition, we filtered out variants with missingness $> 10\%$ using PLINK (Purcell et al., 2007) to obtain our full set of filtered SNPs. This filtered SNP data was used to calculate the genetic relationship matrix (GRM). This is a SNP based estimation of the genetic relationship between all individuals in the sample (Yang et al., 2010). The GRM was used in a SNP based REML heritability analysis (Yang, Lee, Goddard & Visscher, 2011), and its first six principal components capture population structure in the data for the correction of the brain measurements. Estimation of the GRM with its principal components and of the heritability values was performed in GCTA (Yang et al., 2011).

Table 5.1: Overview of the 19 different sites included in the dataset. For more information, see Blokland et al. (2017).

study	cases	controls	male	female	age_mean	age_sd
CAMH	88	103	106	85	45	18
CIDAR_P	1	39	23	17	21	4
CIDAR_VA	61	60	90	31	32	13
GAP	60	34	58	36	28	8
IMH_SIGNRP	131	0	81	50	33	9
LandR	0	64	26	38	26	3
MCIC_MGH	22	23	26	19	41	9
MCIC_UMN	30	19	34	15	32	11
MCIC_UNM	19	20	30	9	33	14
MGH_FB12_MTH	23	52	44	31	38	13
MGH_SSS_SZB	20	46	42	24	32	11
MTS_MH	2	15	5	12	44	14
MTS_SGH	4	6	8	2	45	9
NEFS_AVANTO	0	74	32	42	43	4
NEFS_GENESIS	0	4	4	0	36	3
NEFS_SONATA1	0	20	11	9	38	2
NEFS_SONATA2	0	59	26	33	40	2
NEFS_TRIO	0	18	7	11	48	2
NUIG	75	26	64	37	34	10
PHRS	0	24	10	14	18	4
TCIN	43	160	100	103	31	11
UMCU_SZ1	64	89	110	43	37	13
UMCU_SZ2	138	237	221	154	28	7

The 1 973 individuals also had structural (T1-weighted) MRI measurements, which were processed with FreeSurfer using the Destrieux atlas (Destrieux, Fischl, Dale & Halgren, 2010). We used 62 Destrieux atlas measurements (31 per hemisphere) for cortical thickness and a set of 14 volume measurements (7 per hemisphere) for cerebellar cortex, thalamus, caudate, putamen, pallidum, hippocampus, and amygdala. All measurements were corrected for age, sex, data collection site identifier, total intracranial volume, and the first 6 principal components of their GRM, by calculating the residuals of a multivariate ordinary least squares regression model with these predictors. The residuals of this correction model were used in the Bayesian regression models and additional analyses.

To perform our model search, we linked SNPs to genes in two ways. First, we selected known eQTLs for the human brain from three studies (Gibbs et al.,

2010; B. Ng et al., 2017; Ramasamy et al., 2014). Second, we identified SNPs located inside genes using MAGMA (de Leeuw, Mooij, Heskes & Posthuma, 2015). For both these SNP selections we then performed pairwise SNP tag selection with PLINK by shifting (each time 5 SNPs) a 1Mbp window across the genome and greedily pruning SNPs with a squared correlation > 0.5. Finally, we combined both sets of tags to end up with 70 538 SNPs for 12 033 genes in total (the tag set).

In addition, we linked the FreeSurfer regions to the Allen Brain Atlas regions. For the subcortical brain regions the Allen Atlas annotations are more specific than those used in FreeSurfer. In these cases, we selected the full FreeSurfer region even if only part of it was selected in a partition of the cross-clustering. In the cortical regions we found ambiguous mappings between the regions. Here, we selected all Destrieux regions with at least some overlap with a region that was part of a partition. See Table 5.2 for the links between the FreeSurfer and Allen Brain Atlas brain regions.

5.2.3 Effects of SNPs on FreeSurfer measurements

Each partition of genes and brain regions, based on the gene expression data of the Allen Brain Atlas, was linked to a subset of SNPs and brain regions in the GENUS data. We used Bayesian linear regression models to predict the brain region measurements (volumes or thicknesses) from the SNP data. For a set of brain region measurements in GENUS, Y, we performed a PCA dimension reduction (first component) to the vector y. The SNPs in the GENUS data were complemented with a unit vector to form the predictor matrix X. We then fitted the model

$$y = X\beta + \epsilon,$$

where β is a weight vector and ϵ an error variable with mean 0 and variance σ^2. For each model, corresponding to a partition, we calculated a Bayesian marginal likelihood as a measure of model fit. We used a Gaussian prior on β and an inverse-gamma prior on the error variance σ^2. To put all models on the same scale, we standardized each y to have a mean of 0 and variance of 1, and we centred the SNP values in X to give them a mean of 0 and imputed the missing values with the mean values (0s in this case).

The regression models were fitted on the data of the individuals in the GENUS data. However, the study population consists of both schizophrenia

patents (cases) and healthy controls. The associations between SNPs and brain measurements that we find may either be specific to individuals with schizophrenia, a result of the differences between the patients and controls, or present in the general population. Since the schizophrenia diagnosis label is not part of our models, we try to shed some light on these distinctions by fitting each regression model on the three sets of individuals: all schizophrenia patients ($n_{case} = 1\,192$), all controls ($n_{control} = 781$), and the combined data ($n_{all} = 1\,973$).

We fitted a model for each partition on each of the three datasets. Since the number of samples differs in the datasets, the likelihoods calculated from the models is not on the same scale. To remedy this, we calculated Bayes factors as measures for model quality. We define X_k as the SNP data belonging to the kth partitioning and y_k as the first PC of the measurements of the regions in partition k, to define a Bayes factor

$$BF_k = \frac{p(y_k|X_k, H_1)}{p(y_k|X_k, H_0)},$$

for the model H_1 where volumes are predicted by the SNPs, and the null-hypothesis H_0 where they are predicted by an intercept-only model. Note that $p(y_k|X_k, H_0)$, the null-likelihood, is constant for a given dataset, since it only depends on the fixed mean, variance and number of elements (n) in y_k.

5.2.4 Parameter prior settings

In a Bayesian regression analysis, the model parameters have prior distributions to reflect some prior information. We used informative priors on the regression weights to regularise our models, effectively stating that we expect most weights to be close to 0. Given the high computational costs of fitting this number of models, we did not explore a wide range of prior settings. For the Gaussian prior of the β values we used a mean of 0 and a precision value of either a data-dependent n (so n_{case}, $n_{control}$, or n_{all} for the respective datasets) or a fixed value of 30. With the data size dependent prior, we effectively give our prior equal weight to the selected data. For the error variance σ^2, we set the shape and scale parameter of the inverse gamma distribution both to a value of 0.001, which means this prior has little impact on the posterior.

5.3 Results

5.3.1 Preliminary analyses

We performed a model selection procedure, in which we used gene expression data in the healthy human brain to propose models for imaging genetics in a schizophrenia sample. This procedure gives a list of models of interest in which the volume or thickness of a set of brain regions is predicted by a set of genetic variants. However, before looking into these models, we characterise some aspects of the data.

We first explore the differences between schizophrenia patients and healthy controls with respect to the measured subcortical region volumes and cortical thicknesses. Figure 5.2A shows the results of t-test comparing the brain region measurements of cases and controls. These measurements were corrected by a linear regression for age, sex, data collection site, intracranial volume and 6 principal components of the GRM. We see a significantly larger pallidum and right hand side putamen in schizophrenia cases, and significantly smaller thalamus, hippocampus, and a range of cerebral cortex regions. These patterns are quite symmetrical, with similar t-values for both hemispheres.

Next, we estimated the overall SNP heritability of brain region measurements in our data. This was based on the GRM of the full set of filtered SNPs and data of cases and controls combined (the *all* data). Figure 5.2B shows that none of the regions show a significant heritability ($p < 0.05$) as estimated in our data, with the highest heritability point estimates for pallidum, putamen and a number of cerebral cortex regions. The number of samples in our data set may not be large enough to detect a heritability, or our SNPs do not fully capture the variation in causal variants.

We calculated the first six principal components (PCs) of the GRM to characterise the genetic population structure. Figure 5.3 shows these principal component scores for all individuals. We can see that the first PC is mainly influenced by being part of the IMH site (Institute of Mental Health – Singapore Translational and Clinical Research in Psychosis). Since this is the only study site in Asia, the pattern is likely due to population structure. Moreover, this study only includes cases (see table 5.1), so analyses based on the SNP data are likely to have a diagnosis related bias unless this is corrected for. The principal components were therefore used, together with other

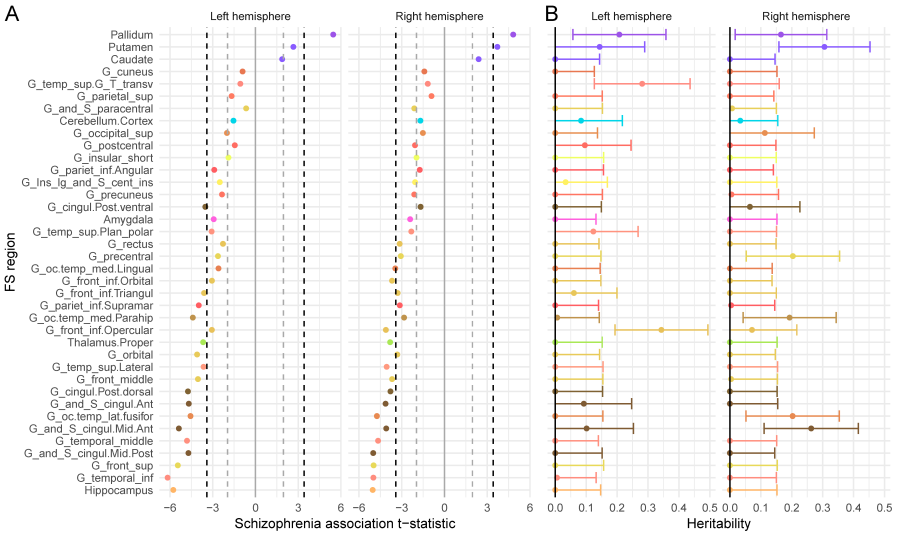


Figure 5.2: ***(A)*** *The t-test results for each of the 76 brain region measurements, comparing the schizophrenia group to the controls. The brain region measurements were corrected by a linear regression for age, sex, data collection site, intracranial volume and 6 principal components of the GRM. The vertical dashed lines show upper and lower cut-offs for* $p < 0.05$ *before (grey) and after (black) Bonferroni multiple testing correction.* ***(B)*** *SNP based heritability estimates for the corrected region measurements on the full dataset (*all*), with standard error bars (1 s.e.). See Figure 5.5A for a colour legend, and see Table 5.2 for the explanation of the region abbreviations.*

covariates, to correct the brain region measurements.

Even though we could not find significant heritability values in our data, we set out to see if any of the individual SNPs have a significant univariate association with the region measurements. Often genome wide association studies need a very large sample size to compensate for the multiple testing issues involved in a genome-wide analysis. In our imaging genetics case this is exacerbated by the fact that we perform the tests for a number of brain regions. Figure 5.4 shows that our analyses have no significantly associated SNPs after additional correction for the number of brain regions. The SNP rs72809913 for the superior parietal lobule (`G_parietal_sup`) is linked to the gene *SMYD5* by an eQTL study (B. Ng et al., 2017). It would be significant

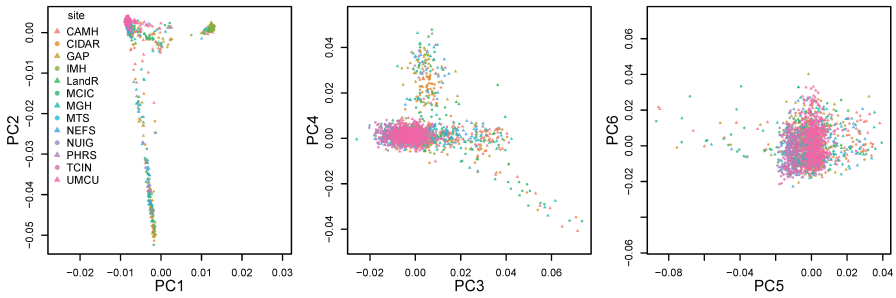


Figure 5.3: The first six principal component scores of the genetic relationship matrix based on the full set of filtered SNPs on all individuals (the all *data). Each dot is an individual and the colours indicate the study site they were part of.*

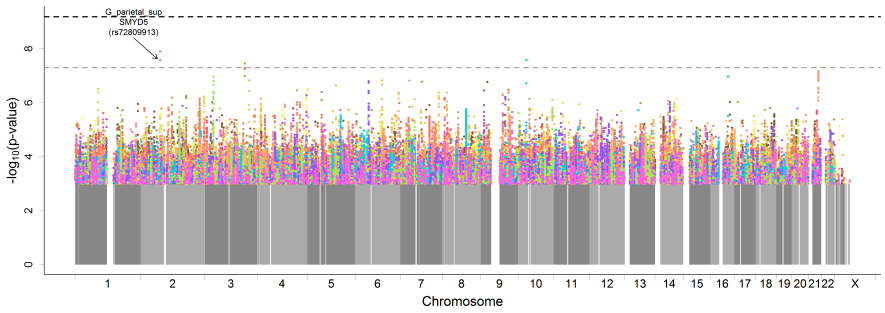


Figure 5.4: Genome wide association plot for univariate models predicting each of the 76 corrected region measurements. Only p-values < 0.001 *are shown as individual points, and their colour indicates the brain region that was tested (see Figure 5.5). Horizontal lines show the standard GWAS cut-off of* $p < 0.05 \cdot 10^{-6}$ *before (grey) and after (black) an additional Bonferroni correction for the number of brain regions being tested.*

($p = 1.25 \cdot 10^{-8}$) without multiple testing correction for the number of regions. With this correction, we find no associations in the genome wide univariate analysis. Our proposed method tries to find multivariate models that do have an impact on brain region measurements.

5.3.2 Clustering for proposed models

Our methodology relies on a cross-clustering of the gene expression data of the Allen Brain Atlas. Each clustering contains a number of partitions, which in turn consist of sets of brain regions and genes. Similarities between genes in the expression dataset tend to reflect shared pathways and functions in the brain (S. M. Huisman et al., 2017) and similarities between brain regions reflect anatomy and cell type composition (Mahfouz et al., 2015). Figure 5.5B shows how frequently brain regions cluster together in our cross-clustering approach. We can see that regions contained in larger structures, such as the cerebral and cerebellar cortex or the striatum tend to cluster together with the other regions within these structures. Figure 5.5C shows an example of a single cross-clustering.

From our clustering on the Allen Brain Atlas data we obtained a set 385 cross-clusterings, containing 101 694 cross-clustering partitions. These partitions varied widely in size, with the number of genes in a cross-clustering partition ranging from 1 to 19 971 and the number of brain region measurements from 1 to all 105. Figure 5.6A shows the 385 clusterings and their cross-clustering likelihoods, comparing the average number of gene and samples (brain regions) the partitions in these cross-clusterings contain. The best clustering, with the highest clustering likelihood, had an average size of 98 genes and 11 brain regions per partition. In the end, it is not the clusterings we are interested in, but the 101 694 partitions.

For each cross-clustering partition of genes and brain regions in the Allen Brain Atlas data, we selected the linked SNPs and brain regions in the GENUS data. As a result, we obtained model data for the 101 694 cross-clustering partitions. This model data of course also varied in size, with the number of SNPs in a cross-clustering partition ranging from 1 to 69 166 and the number of brain region measurements from 2 (all regions have measurements in two hemispheres) to all 76. In this way, each cross-clustering partition was translated to a single regression model. Each model, in turn, was fitted on each of the three datasets (control, case or both).

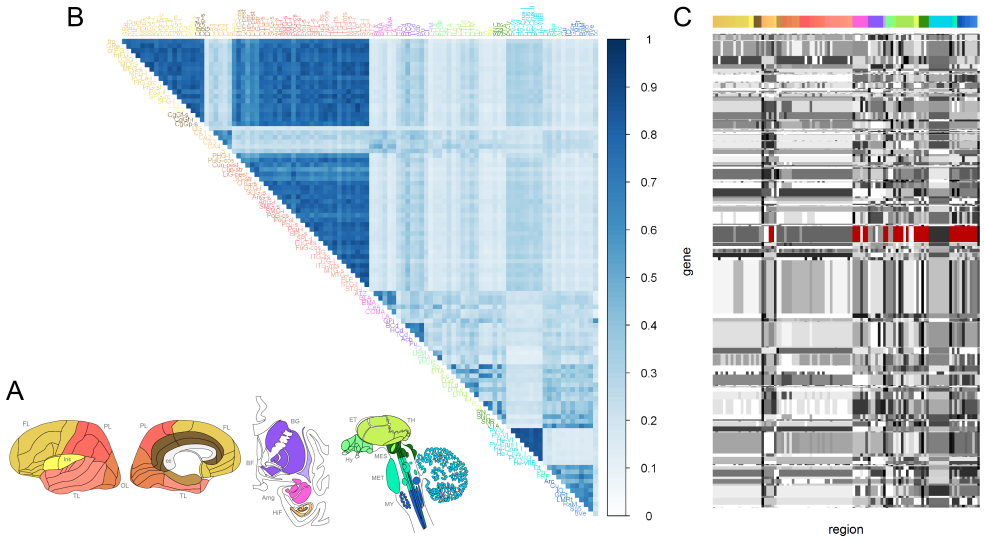


Figure 5.5: ***(A)*** *A schematic map of the brain, showing the colours used in the Allen Reference Atlas (Ding et al., 2016) and throughout this paper.* ***(B)*** *Co-clustering proportions of the Allen Brain Atlas regions in the cross-clustering approach. For each clustering we averaged the proportion of times each pair of regions was part of the same partition, and then averaged those proportions over all cross-clusterings. A value of 1 indicates regions are always found together in a partition.* ***(C)*** *An example of a cross-clustering on the ABA gene expression data. One of the partitions is highlighted in red. Note that the columns (regions) cannot be ordered to make all regions within them side-by-side, so ordering is based on anatomy (see* ***A****), and within each gene view the regions in a single partition share a shade of grey.*

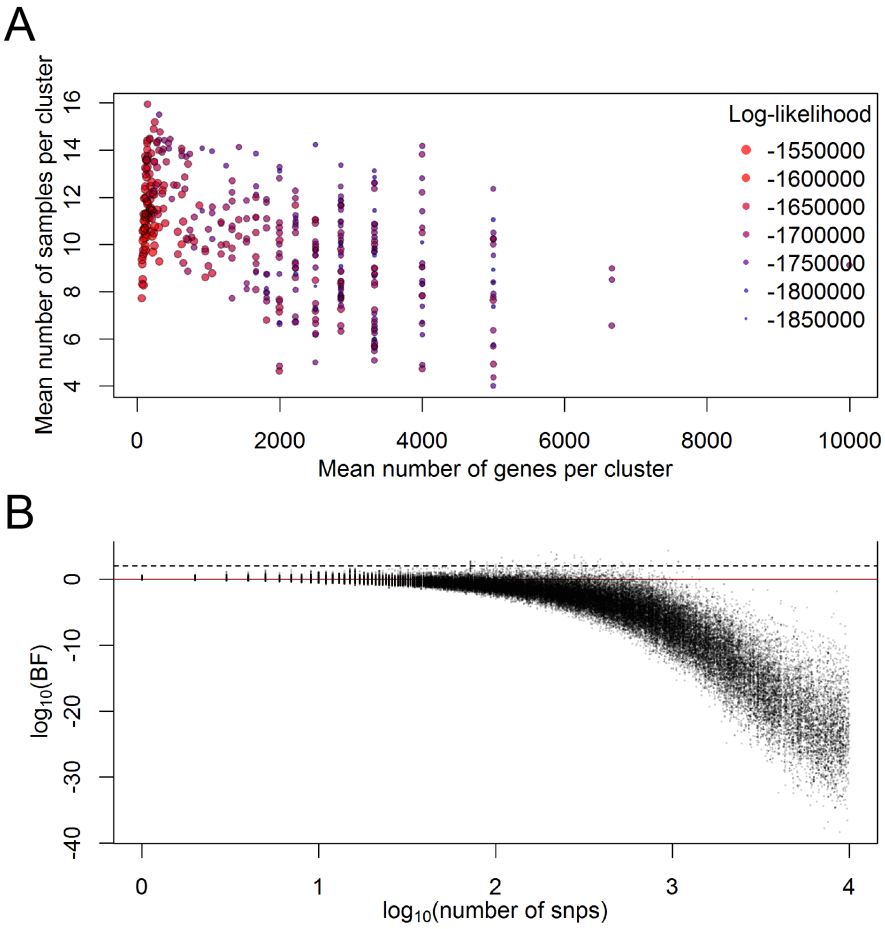


Figure 5.6: ***(A)*** *The likelihoods for each of the 385 cross-clusterings and the average number of genes and samples in the clusters they contain.* ***(B)*** *The log-Bayes factors for the models for the* all *dataset with prior n_{all}, as a function of the number of SNPs they contain. Larger models tend to have a lower Bayes factor. Models containing more than 10 000 SNPs were not considered.*

5.3.3 Selected models

We obtained a marginal likelihood for each model, with two different settings for the prior precision of the β values (see section 5.2.4). To interpret these likelihoods, we translated them into Bayes factors with respect to a null model which contained only an intercept to predict the region volumes. Figure 5.6B shows the likelihoods as a function of the number of SNPs in each model, for the *all* dataset, and the prior β precision value set to the number of individuals n_{all}. Models containing more than 10 000 SNPs (out of the total 70 538 SNPs) were not considered, because large models tend to have low likelihoods and are computationally intensive to fit. Note that the overwhelming majority of models has a negative log-Bayes factor and is therefore not of interest.

Each cross-clustering on the Allen Brain Atlas data was sampled independently, and two clusterings could therefore contain identical partitions. Even if two partitions were not identical on the Allen Brain Atlas data, after mapping to the GENUS data they could lead to the selection of the same sets of SNPs and brain regions. As models of interest, we selected all unique models with a log-10 Bayes factor larger than 2 in any of the three datasets (cases, controls, or all). This means such a model has a likelihood at least 100 times higher than the corresponding null model. As a result, we end up with two sets of models of interest, one for each of the β prior settings (203 models for prior n and 77 models for prior 30).

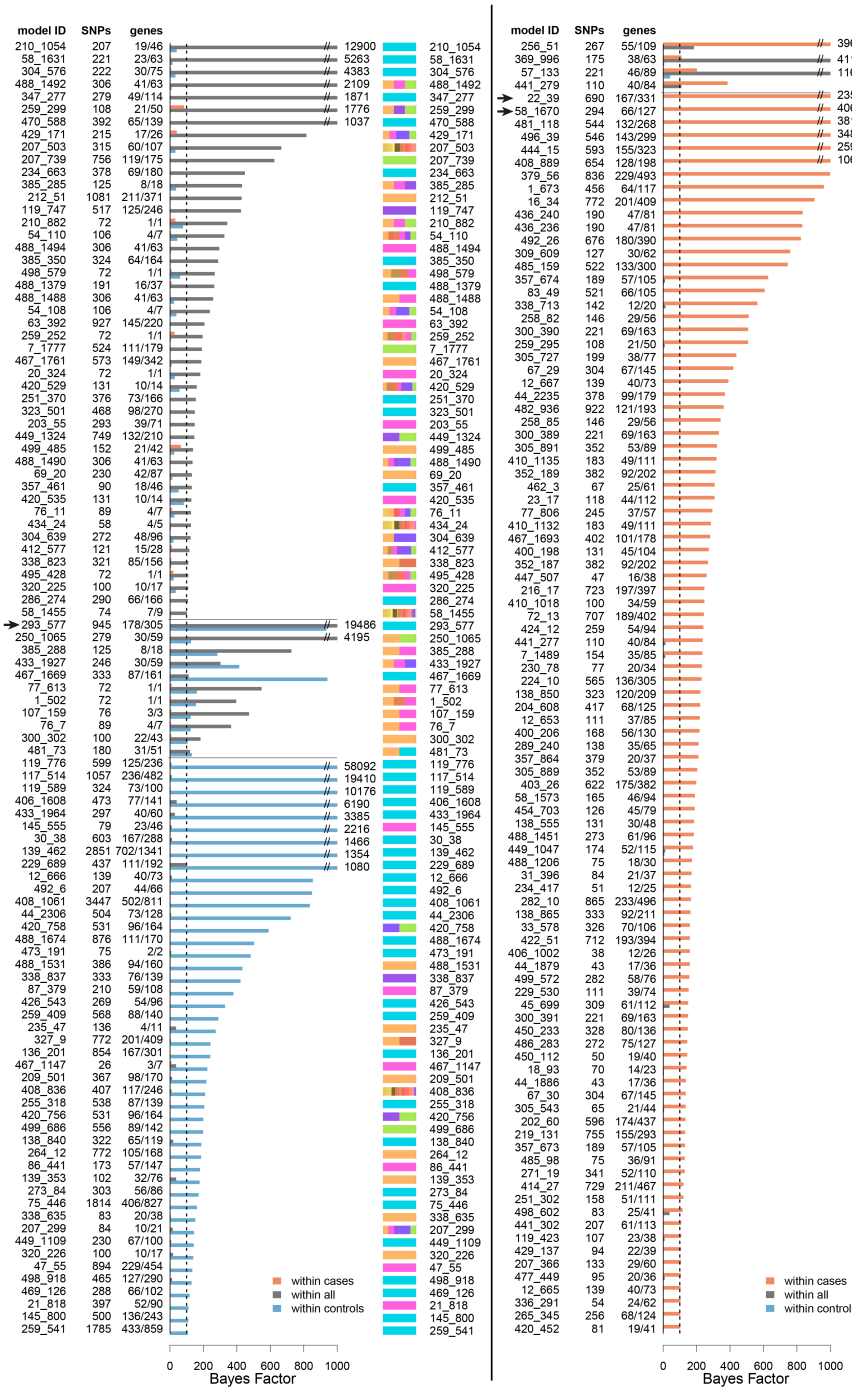
Figure 5.7 shows some statistics of the models of interest where the β prior precision is set to n. The number of SNPs in the selected models ranges from 26 to 3 447, and the number of regions from 2 to 74. The figures show a comparison of the likelihoods in each of the three datasets. In total, 46 models were selected in the *all* dataset only, 11 in the *all* data and the controls, 46 in the controls only, 4 in the controls and cases, and 96 in the cases only.

Before we go into a description of some models, we would like to stress that the selected models are not independent. They often overlap in the regions they predict, and in the SNPs that are selected. Figure 5.8 and Supplemental Figures 5.11 and 5.12 show the relationships between the models in model networks. Each node in the network is a model, so it contains a set of brain regions and a set of SNPs. Models are connected if they have any overlap in the set of selected SNPs. If a model has a high Bayes factor, other models with similar sets of SNPs and similar sets of regions might also do well. As a result,

we find groups of connected models (so shared SNPs) with an overlap in brain regions (indicated by the node colours). So model overlap in SNPs is captured in the node connections, and overlap in regions in shared node colours. Note for example a cluster of models (nodes) at the bottom of Figure 5.8 that all represent models for the cerebellum, with some overlap in SNPs. The cluster of models at the top right even contains a number of models with an identical set of 72 SNPs, all located in the gene *EYS*. These models differ in the brain regions that were selected, although they all contain the amygdala.

We do not provide a full interpretation of all selected models, but will highlight a few of them, indicated with arrows in Figure 5.7. Functional enrichment analyses are performed using DAVID 6.8 (D. W. Huang, Sherman & Lempicki, 2009). Model 58_1670 is selected in the cases only and it contains samples from left and right hand pallidum, see Figure 5.9. The model seems predictive for pallidum volume in the cases. The full set of 127 genes selected in the gene expression data contains 41 genes that are annotated as structural constituent of ribosome (GO:0003735), Benjamini-Hochberg corrected $p = 4.7 \cdot 10^{-47}$. Given that this model was only selected using the schizophrenia cases, the variations in pallidum size might be related to variation among schizophrenia patients. It is not selected in the *all* data, perhaps because the signal is drowned out by variation in the controls. Note that the volume of the pallidum is positively associated with schizophrenia in our sample, see Figure 5.2A. Interestingly, asymmetries in this region have been associated with schizophrenia in the ENIGMA Consortium data (Okada et al., 2016), and ribosomal gene DNA copy numbers and DNA damage seem to be higher in schizophrenia patients (Chestkov et al., 2018; Porokhovnik et al., 2015).

Figure 5.7: (Next page) The model Bayes factors in each of the three datasets for all selected models with the β prior precision set to the number of samples. The figure is split over two lists to fit the page. The number of SNPs in each model is shown on the left and the colour bars on the right indicate the regions that were part of that model. The column for genes shows the number of genes in the original partition, and to the left of that the number of those that actually contain SNPs. The three models that are preceded by a black arrow are discussed in the text.


model ID SNPs genes
210_1054 207 19/46
58_1631 221 23/63
304_576 222 30/75
488_1492 306 41/63
347_277 279 49/114
259_299 108 21/50
470_588 392 65/139
429_171 215 17/26
207_503 315 60/107
207_739 756 119/175
234_663 378 69/180
385_285 125 8/18
212_51 1081 211/371
119_747 517 125/246
210_882 72 1/1
54_110 106 4/7
488_1494 306 41/63
385_350 324 64/164
498_579 72 1/1
488_1379 191 16/37
488_1488 306 41/63
54_108 106 4/7
63_392 927 145/220
259_252 72 1/1
7_1777 524 111/179
467_1761 573 149/342
20_324 72 1/1
420_529 131 10/14
251_370 376 73/166
323_501 468 98/270
203_55 293 39/71
449_1324 749 132/210
499_485 152 21/42
488_1490 306 41/63
69_20 230 42/87
357_461 90 18/46
420_535 131 10/14
76_11 89 4/7
434_24 58 4/5
304_639 272 48/96
412_577 121 15/28
338_823 321 85/156
495_428 72 1/1
320_225 100 10/17
286_274 290 66/166
58_1455 74 7/9
293_577 945 178/305
250_1065 279 30/59
385_288 125 8/18
433_1927 246 30/59
467_1669 333 87/161
77_613 72 1/1
1_502 72 1/1
107_159 76 3/3
76_7 89 4/7
300_302 100 22/43
481_73 180 31/51
119_776 599 125/236
117_514 1057 236/482
119_589 324 73/100
406_1608 473 77/141
433_1964 297 40/60
145_555 79 23/46
30_38 603 167/288
139_462 2851 702/1341
229_689 437 111/192
12_666 139 40/73
492_6 207 44/66
408_1061 3447 502/811
44_2306 504 73/128
420_758 531 96/164
488_1674 876 111/170
473_191 75 2/2
488_1531 386 94/160
338_837 333 76/139
87_379 210 59/108
426_543 269 54/96
259_409 568 88/140
235_47 136 4/11
327_9 772 201/409
136_201 854 167/301
467_1147 26 3/7
209_501 367 98/170
408_836 407 117/246
255_318 538 87/139
420_756 531 96/164
499_686 556 89/142
138_840 322 65/119
264_12 772 105/168
86_441 173 57/147
139_353 102 32/76
273_84 303 56/86
75_446 1814 406/827
338_635 83 20/38
207_299 84 10/21
449_1109 230 67/100
320_226 100 10/17
47_55 894 229/454
498_918 465 127/290
469_126 288 66/102
21_818 397 52/90
145_800 500 136/243
259_541 1785 433/859
// 12900
// 5263
// 4383
// 2109
// 1871
// 1776
// 1037
// 19486
// 4195
// 58092
// 19410
// 10176
// 6190
// 3385
// 2216
// 1466
// 1354
// 1080
210_1054
58_1631
304_576
488_1492
347_277
259_299
470_588
429_171
207_503
207_739
234_663
385_285
212_51
119_747
210_882
54_110
488_1494
385_350
498_579
488_1379
488_1488
54_108
63_392
259_252
7_1777
467_1761
20_324
420_529
251_370
323_501
203_55
449_1324
499_485
488_1490
69_20
357_461
420_535
76_11
434_24
304_639
412_577
338_823
495_428
320_225
286_274
58_1455
293_577
250_1065
385_288
433_1927
467_1669
77_613
1_502
107_159
76_7
300_302
481_73
119_776
117_514
119_589
406_1608
433_1964
145_555
30_38
139_462
229_689
12_666
492_6
408_1061
44_2306
420_758
488_1674
473_191
488_1531
338_837
87_379
426_543
259_409
235_47
327_9
136_201
467_1147
209_501
408_836
255_318
420_756
499_686
138_840
264_12
86_441
139_353
273_84
75_446
338_635
207_299
449_1109
320_226
47_55
498_918
469_126
21_818
145_800
259_541
within cases
within all
within controls
0 200 400 600 800 1000
Bayes Factor
model ID SNPs genes
256_51 267 55/109
369_996 175 38/63
57_133 221 46/89
441_279 110 40/84
22_39 690 167/331
58_1670 294 66/127
481_118 544 132/268
496_39 546 143/299
444_15 593 155/323
408_889 654 128/198
379_56 836 229/493
1_673 456 64/117
16_34 772 201/409
436_240 190 47/81
436_236 190 47/81
492_26 676 180/390
309_609 127 30/62
485_159 522 133/300
357_674 189 57/105
83_49 521 66/105
338_713 142 12/20
258_82 146 29/56
300_390 221 69/163
259_295 108 21/50
305_727 199 38/77
67_29 304 67/145
12_667 139 40/73
44_2235 378 99/179
482_936 922 121/193
258_85 146 29/56
300_389 221 69/163
305_891 352 53/89
410_1135 183 49/111
352_189 382 92/202
462_3 67 25/61
23_17 118 44/112
77_806 245 37/57
410_1132 183 49/111
467_1693 402 101/178
400_198 131 45/104
352_187 382 92/202
447_507 47 16/38
216_17 723 197/397
410_1018 100 34/59
72_13 707 189/402
424_12 259 54/94
441_277 110 40/84
7_1489 154 35/85
230_78 77 20/34
224_10 565 136/305
138_850 323 120/209
204_608 417 68/125
12_653 111 37/85
400_206 168 56/130
289_240 138 35/65
357_864 379 20/37
305_889 352 53/89
403_26 622 175/382
58_1573 165 46/94
454_703 126 45/79
138_555 131 30/48
488_1451 273 61/96
449_1047 174 52/115
488_1206 75 18/30
31_396 84 21/37
234_417 51 12/25
282_10 865 233/496
138_865 333 92/211
33_578 326 70/106
422_51 712 193/394
406_1002 38 12/26
44_1879 43 17/36
499_572 282 58/76
229_530 111 39/74
45_699 309 61/112
300_391 221 69/163
450_233 328 80/136
486_283 272 75/127
450_112 50 19/40
18_93 70 14/23
44_1886 43 17/36
67_30 304 67/145
305_543 65 21/44
202_60 596 174/437
219_131 755 155/293
357_673 189 57/105
485_98 75 36/91
271_19 341 52/110
414_27 729 211/467
251_302 158 51/111
498_602 83 25/41
441_302 207 61/113
119_423 107 23/38
429_137 94 22/39
207_366 133 29/60
477_449 95 20/36
12_665 139 40/73
336_291 54 24/62
265_345 256 68/124
420_452 81 19/41
// 3961
// 4112
// 1164
// 23558
// 4060
// 3819
// 3487
// 2598
// 1061
256_51
369_996
57_133
441_279
22_39
58_1670
481_118
496_39
444_15
408_889
379_56
1_673
16_34
436_240
436_236
492_26
309_609
485_159
357_674
83_49
338_713
258_82
300_390
259_295
305_727
67_29
12_667
44_2235
482_936
258_85
300_389
305_891
410_1135
352_189
462_3
23_17
77_806
410_1132
467_1693
400_198
352_187
447_507
216_17
410_1018
72_13
424_12
441_277
7_1489
230_78
224_10
138_850
204_608
12_653
400_206
289_240
357_864
305_889
403_26
58_1573
454_703
138_555
488_1451
449_1047
488_1206
31_396
234_417
282_10
138_865
33_578
422_51
406_1002
44_1879
499_572
229_530
45_699
300_391
450_233
486_283
450_112
18_93
44_1886
67_30
305_543
202_60
219_131
357_673
485_98
271_19
414_27
251_302
498_602
441_302
119_423
429_137
207_366
477_449
12_665
336_291
265_345
420_452
within cases
within all
within controls
0 200 400 600 800 1000
Bayes Factor

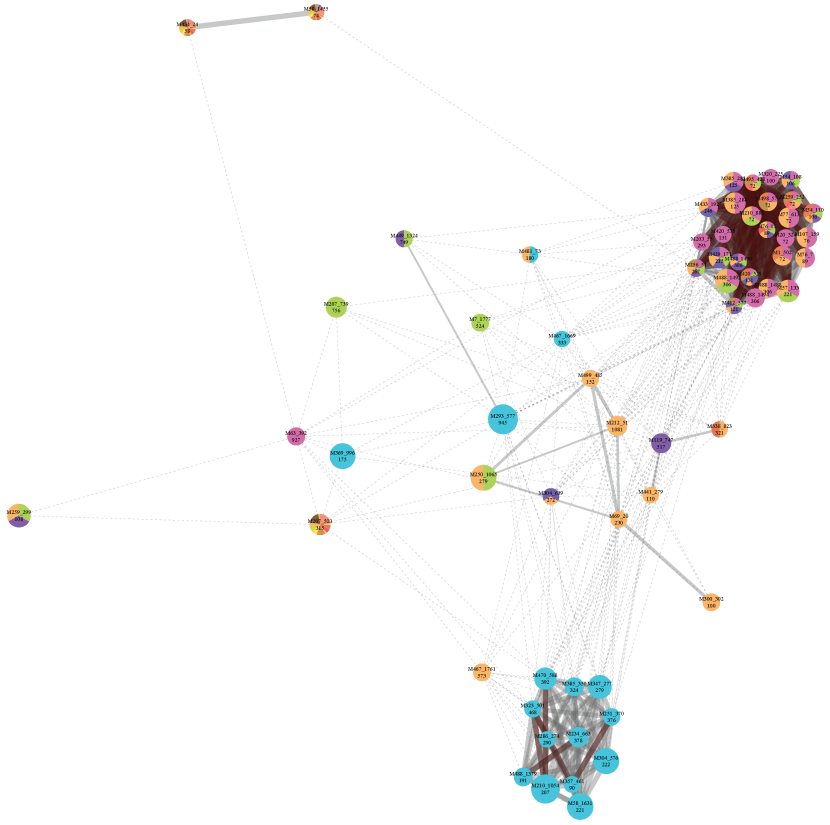


Figure 5.8: The model network for all models selected on the all *dataset (cases and controls) with the penalisation parameter set to the number of samples n_{all}. Each node represents one model, has a size proportional to the Bayes factor, and is coloured to indicate the brain regions included. The text on each node shows the node identifier and the number of SNPs it contains. Nodes (models) are connected if they share SNPs, either $< 20\%$ (dashed edge) or $\geq 20\%$ (solid edge) of the smaller of the two models. If all SNPs in the smaller model are contained in the other model the edge is coloured red.*

The set of selected models for cases-only contains a subset of overlapping models that predict the thickness of the inferior temporal gyrus, as can be seen in Figure 5.11. The model with the highest Bayes factor in this group is 22_39, see Figure 5.10. Of the 331 genes selected in this model, 37 have olfactory receptor activity (GO:0004984), Benjamini-Hochberg corrected $p = 4.5 \cdot 10^{-16}$. This model only had strong evidence within the cases, in predicting cortical thickness from the SNPs, but the inferior temporal gyrus is also the brain region most strongly associated to the schizophrenia label in our data (Figure 5.2A). Schizophrenia patients tend to have a thinner inferior temporal gyrus.

Finally, model 293_577 is selected in the *all* set and the controls, and it contains samples from the cerebellum, see Figure 5.14. The full set of 305 genes has 21 genes that have been linked to immunity (UniProtKB KW-0391), Benjamini-Hochberg corrected $p = 5.1 \cdot 10^{-3}$. Because the model performs well in the controls too, the possible link between the SNPs in this model and the brain region measurements is most likely not related to schizophrenia. Note that the maximum a posteriori probability (MAP) estimate of the error variance is also relatively low in the cases, but this is not reflected in the model likelihood and Bayes factor.

When we set the prior precision of β to 30, the models with a log-10 Bayes factor > 2 tend to be considerably smaller in number of SNPs. Figure 5.13 shows some characteristics of these models. The number of SNPs in the selected models ranges from only 1 to 29, and the number of regions from 4 to 74. Of the selected models, 1 had a log-10 Bayes factor > 2 in the cases and in the *all* data, 43 only in the cases, 31 only in the *all* data, and 2 only in the controls. On top of the bars showing the Bayes factors for the three datasets we list the genes that are represented in the model, either because they contain eQTLs or because model SNPs are located within the gene boundaries. Most of the models that were selected in the *all* dataset only contain SNPs in the *AFM* gene, and differ just in the exact combination of brain regions selected. Note that these are therefore highly correlated models.

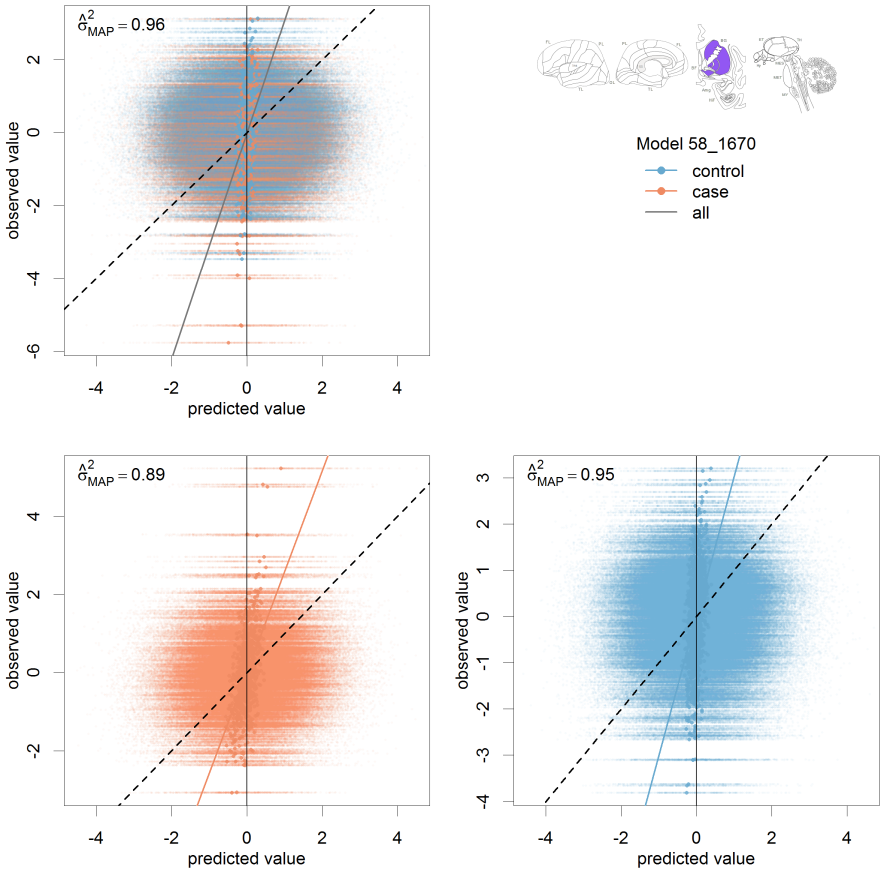


Figure 5.9: Characterisation of model 58_1670, with the β prior precision set to the number of samples. The plots show the observed first principal component score (y) of the region measurements versus the estimated posterior predictive distribution for each individual. We approximated these distributions by each time sampling a set of regression weights (β) from their multivariate normal posterior distribution, and an error variance (σ^2) from its inverse-gamma posterior distribution, and then taking a sample from the posterior predictive distribution using these parameters. The number at the top-left of each plot indicates the maximum a posteriori probability estimate of the error variance. Since all variables were scaled to unit-variance, an error variance < 1 indicates that at least part of the sample variance is explained by the model. The schematic view of the brain shows the selected brain regions.

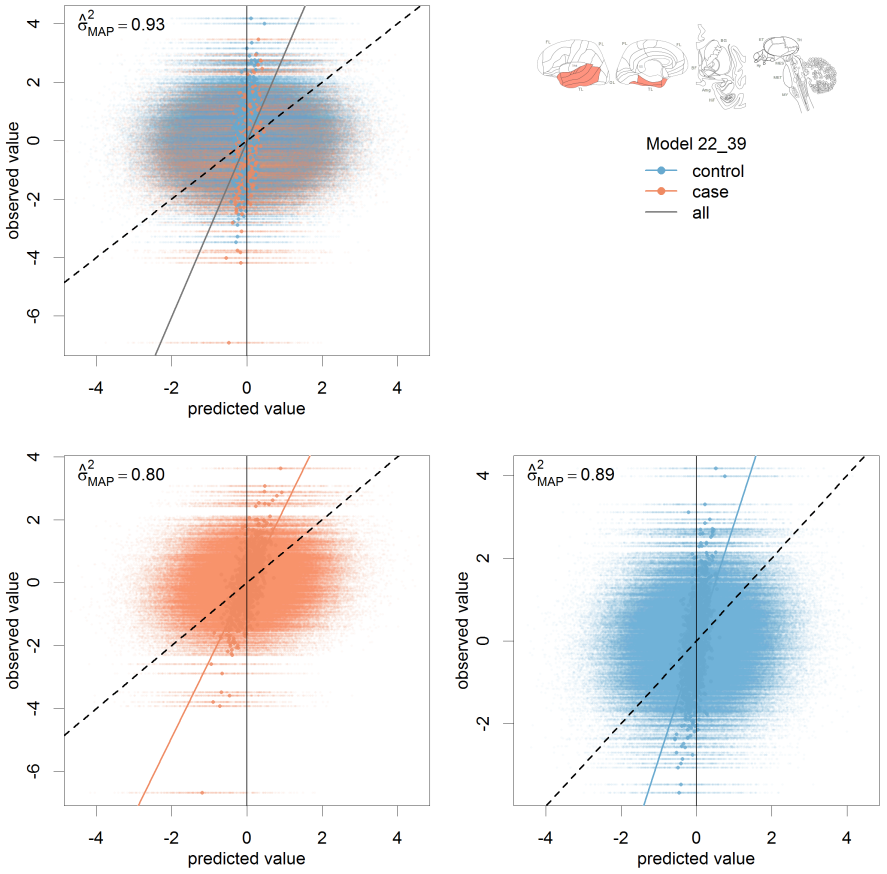


Figure 5.10: Characterisation of model 22_39, with the β prior precision set to the number of samples. For further explanation, see the caption of Figure 5.9.

5.4 Discussion

We set out to find the relationships between genetic variants and brain region measurements in a schizophrenia context. Rather than fitting univariate models, as in a standard genome wide association study, we explored a set of multivariate models. These models were proposed using a cross-clustering on the genome wide and brain wide gene expression data of the Allen Human Brain Atlas. Each partition of the gene expression data was translated into a model predicting a set of FreeSurfer measurements of schizophrenia patients and healthy controls from a set of SNP profiles.

We found a large set of models with a higher evidence than a null-model. Some of the results point to involvement of ribosomal genes in the pallidum, which was enlarged in schizophrenia patients, and olfactory receptor genes in the inferior temporal gyrus, which was thinner in schizophrenia patients. These results are exploratory, and do not show a direct causal link, but they illustrate the effectiveness of our method in finding imaging genetic associations. The effect sizes are small, as they commonly are in genome wide association studies (Visscher et al., 2017), which makes the models more sensitive to batch effects of datasets. The smaller datasets that are contained in the GENUS data differ widely in geographic provenance, demographic characteristics, and relative group sizes for cases and controls. Specially for studies that contain only either cases or controls, a simple linear regression for batch effect removal might not be sufficient.

An additional challenge in the interpretation of the biological results, is the fact that the models do not explicitly contain the schizophrenia label. Models are fitted on cases, controls or all data, but the relationship between the imaging genetic signals and the disease label is not modelled. Theoretically, any relationships between genetic markers and brain region measurements could be independent of schizophrenia. However, we pose that relationships that are present in the cases, but not in the controls, may be related to schizophrenia severity or type. Relationships that were found in the full dataset only (*all*) could be related to schizophrenia, or be a general result, picked up because of the increased statistical power in the *all* data. This complicates interpretation and is an additional reason why the results need independent verification.

Our method relies on the information provided by the gene expression data, both with respect to region-region similarities and gene-gene similarit-

ies. So one of the assumptions of our approach is that the clustering of genes is relevant for the phenotype of interest. That is, genes with a similar spatial expression pattern contain SNPs that, together, predict the variation in brain region volumes and thicknesses measured in the study population. The value of transcriptional similarities of genes in the brain has been shown in multiple studies (Hawrylycz, Miller et al., 2015; Mahfouz, Huisman, Lelieveldt & Reinders, 2017), but if the goal is to find useful groupings of genes, one could perform a pathway based analysis instead. Pathway gene groupings have been used in a range of complex disease studies (Kao, Leung, Chan, Yip & Yap, 2017). However, the advantages of our gene groupings are that they are brain tissue specific and that they include genes for which no pathway information is available. Alternatively, one could use a hybrid approach, where pathway shared membership is used as prior grouping information (Du et al., 2016).

In addition to the gene-gene similarities, the expression data we used for model selection also informs our grouping of brain regions. Gene expression in the Allen Brain Atlas is strongly influenced by cell type composition (Grange et al., 2014; Hawrylycz, Miller et al., 2015), and it reflects relative locations in the brain (Mahfouz et al., 2015). If differences in brain area sizes are related to development in specific cell types, this would support our way of clustering. Since SNPs appear to affect gene expression in a cell-type specific manner (Ardlie et al., 2015), brain regions with a similar cell type composition are more likely to be affected by the same SNPs. As an alternative to gene-expression based similarity, one could use connectivity information, to make groupings that are more closely related to functional networks in the brain.

A challenge of using the Allen Brain Atlas in combination with the GENUS data is that it requires mappings. First, each gene grouping is translated into a grouping of SNPs. This linking of SNP to gene is a common challenge in genetic research (Mooney et al., 2014), and for a real understanding of the biological mechanisms, it may require additional wet-lab experiments. Second, the regions in the Allen Atlas have to be mapped to FreeSurfer regions, also a common challenge in studies involving brain imaging data (Evans, Janke, Collins & Baillet, 2012). Although the Allen Reference Atlas (Ding et al., 2016) and the Destrieux atlas (Destrieux et al., 2010) are similar in their use of anatomical terms, they are not identical. In addition, some regions, notably the ventricles, have no gene expression data, and are therefore ignored in this study.

A problem often encountered in imaging genetics studies is the need for strong multiple testing corrections of p-values, on datasets with small sample sizes. Because our study is a Bayesian exploratory analysis without true hypothesis testing, multiple testing correction is formally not needed. However, it is still a challenge to point out which models are truly interesting, and to avoid proposing models that were selected purely by chance. This could be addressed by explicitly modelling the fact that we consider a large number of hypotheses (Westfall, Johnson & Utts, 1997).

Even though we already consider over 100 000 models, the total number of combinations of genes and brain regions is many orders of magnitude larger. Moreover, the proposed models are overlapping in their genes and regions. We could have proposed larger and more diverse set of models by using a true bi-clustering method rather than a cross-clustering. Having more models to consider would, however, make the multiple "hypothesis testing" issues stronger, as would having models that are less correlated.

Finally, our approach poses some practical challenges. As in any Bayesian analysis, we specify prior distributions for the model parameters. We used informative prior distributions for the regression weights, as a way to regularise the models. Our prior distribution reflects the expectation that most SNPs will have no effects. However, the strength of this prior distribution, i.e. the precision of the prior, has to be specified. We have used two different settings for this prior. The more informative prior favoured models with a larger number of SNPs. Determining which models are of most interest requires further investigation of the biological relationships between genetic variations and brain region measurements.

By using a cross-clustering of gene expression data in the healthy human brain, we explored a large set of models for imaging genetics in schizophrenia. This shows the potential for multivariate modelling in a high dimensional data problem, while making use of prior information. Since the prior knowledge of the gene expression data is brain specific, this also aids in the interpretation of identified models.

5.5 Supplement

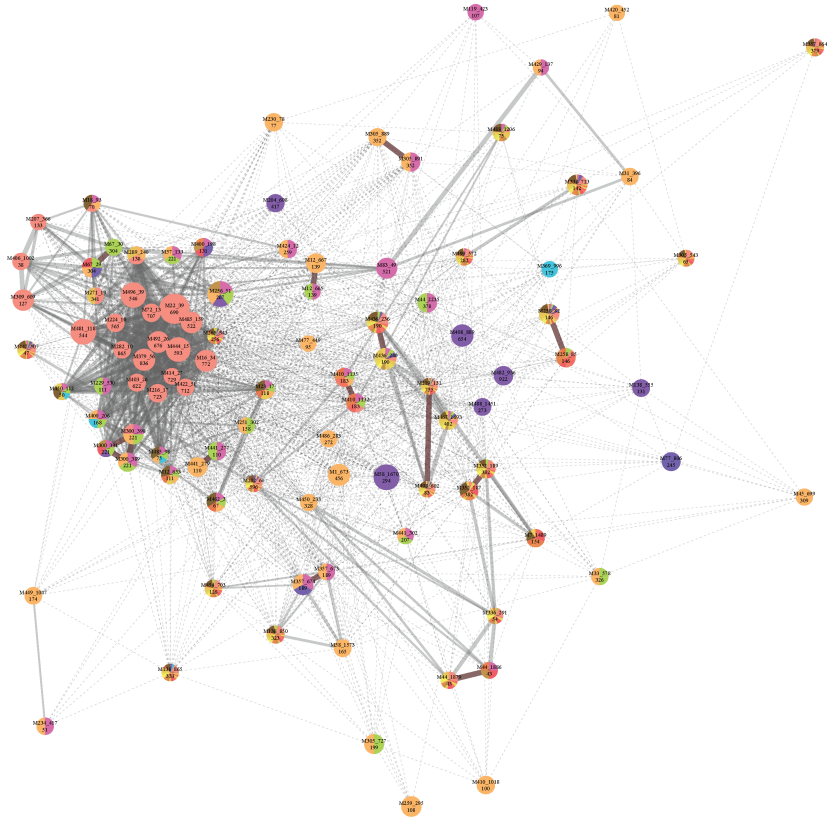


Figure 5.11: The model network for all models selected on the data of schizophrenia patents only, with the penalisation parameter set to the number of samples n_{case}. For further explanation, see the caption of Figure 5.8.

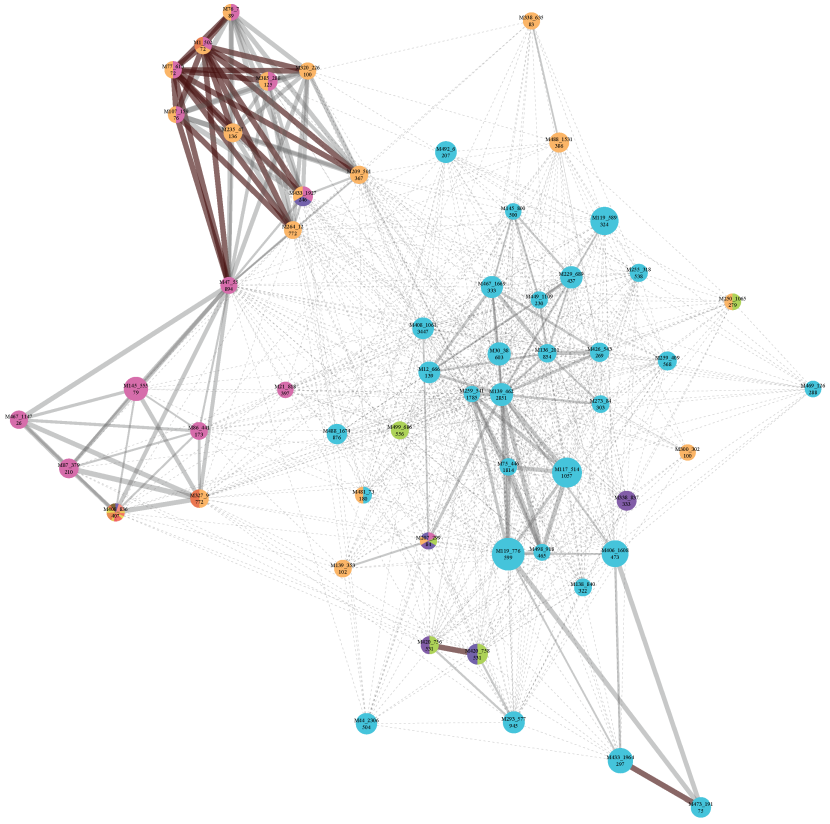


Figure 5.12: The model network for all models selected on the controls only, with the penalisation parameter set to the number of samples $n_{control}$. For further explanation, see the caption of Figure 5.8.

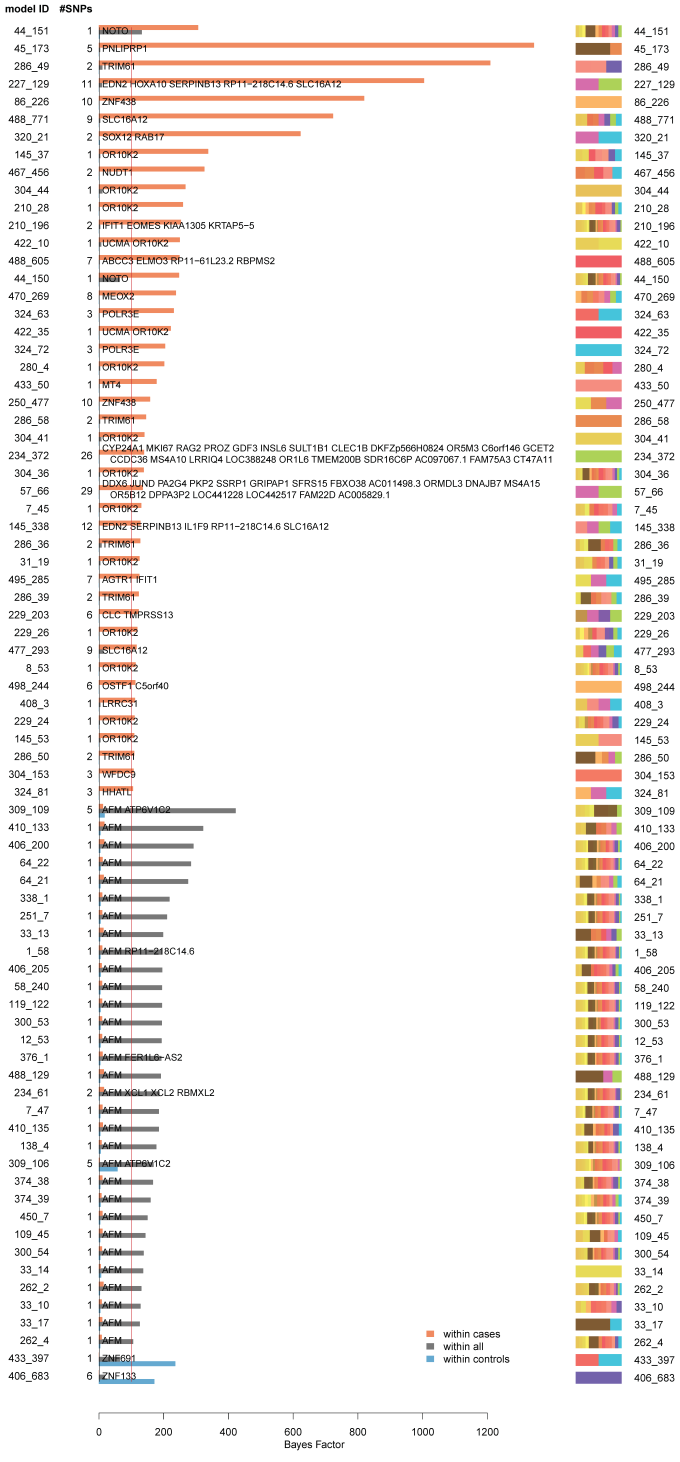

model ID
#SNPs
44_151 1 NOTO
45_173 5 PNLIPRP1
286_49 2 ITRIM61
227_129 11 EDN2 HOXA10 SERPINB13 RP11−218C14.6 SLC16A12
86_226 10 ZNF438
488_771 9 SLC16A12
320_21 2 SOX12 RAB17
145_37 1 OR10K2
467_456 2 NUDT1
304_44 1 OR10K2
210_28 1 OR10K2
210_196 2 IFIT1 EOMES KIAA1305 KRTAP5−5
422_10 1 UCMA OR10K2
488_605 7 ABCC3 ELMO3 RP11−61L23.2 RBPMS2
44_150 1 NOTO
470_269 8 MEOX2
324_63 3 POLR3E
422_35 1 UCMA OR10K2
324_72 3 POLR3E
280_4 1 OR10K2
433_50 1 MT4
250_477 10 ZNF438
286_58 2 TRIM61
304_41 1 OR10K2
234_372 26 CYP24A1 MKI67 RAG2 PROZ GDF3 INSL6 SULT1B1 CLEC1B DKFZp566H0824 OR5M3 C6orf146 GCET2 CCDC36 MS4A10 LRRIQ4 LOC388248 OR1L6 TMEM200B SDR16C6P AC097067.1 FAM75A3 CT47A11
304_36 1 OR10K2
57_66 29 DDX6 JUND PA2G4 PKP2 SSRP1 GRIPAP1 SFRS15 FBXO38 AC011498.3 ORMDL3 DNAJB7 MS4A15 OR5B12 DPPA3P2 LOC441228 LOC442517 FAM22D AC005829.1
7_45 1 OR10K2
145_338 12 EDN2 SERPINB13 IL1F9 RP11−218C14.6 SLC16A12
286_36 2 ITRIM61
31_19 1 OR10K2
495_285 7 AGTR1 IFIT1
286_39 2 TRIM61
229_203 6 CLC TMPRSS13
229_26 1 OR10K2
477_293 9 SLC16A12
8_53 1 OR10K2
498_244 6 OSTF1 C5orf40
408_3 1 LRRC31
229_24 1 OR10K2
145_53 1 OR10K2
286_50 2 TRIM61
304_153 3 WFDC9
324_81 3 HHATL
309_109 5 AFM ATP6V1C2
410_133 1 AFM
406_200 1 AFM
64_22 1 AFM
64_21 1 AFM
338_1 1 AFM
251_7 1 AFM
33_13 1 AFM
1_58 1 AFM RP11−218C14.6
406_205 1 AFM
58_240 1 AFM
119_122 1 AFM
300_53 1 AFM
12_53 1 AFM
376_1 1 AFM FER1L6−AS2
488_129 1 AFM
234_61 2 AFM XCL1 XCL2 RBMXL2
7_47 1 AFM
410_135 1 AFM
138_4 1 AFM
309_106 5 AFM ATP6V1C2
374_38 1 AFM
374_39 1 AFM
450_7 1 AFM
109_45 1 AFM
300_54 1 AFM
33_14 1 AFM
262_2 1 AFM
33_10 1 AFM
33_17 1 AFM
262_4 1 AFM
433_397 1 ZNF691
406_683 6 ZNF133
within cases
within all
within controls
0
200
400
600
800
1000
1200
Bayes Factor
44_151
45_173
286_49
227_129
86_226
488_771
320_21
145_37
467_456
304_44
210_28
210_196
422_10
488_605
44_150
470_269
324_63
422_35
324_72
280_4
433_50
250_477
286_58
304_41
234_372
304_36
57_66
7_45
145_338
286_36
31_19
495_285
286_39
229_203
229_26
477_293
8_53
498_244
408_3
229_24
145_53
286_50
304_153
324_81
309_109
410_133
406_200
64_22
64_21
338_1
251_7
33_13
1_58
406_205
58_240
119_122
300_53
12_53
376_1
488_129
234_61
7_47
410_135
138_4
309_106
374_38
374_39
450_7
109_45
300_54
33_14
262_2
33_10
33_17
262_4
433_397
406_683

Figure 5.13: (Previous page) The model Bayes factors in each of the three datasets for all selected models with the β prior precision set to 30. The number of SNPs in each model is shown on the left and the colour bars on the right indicate the regions that were part of that model. The names on the bars show all genes that were linked to the SNPs in that model.

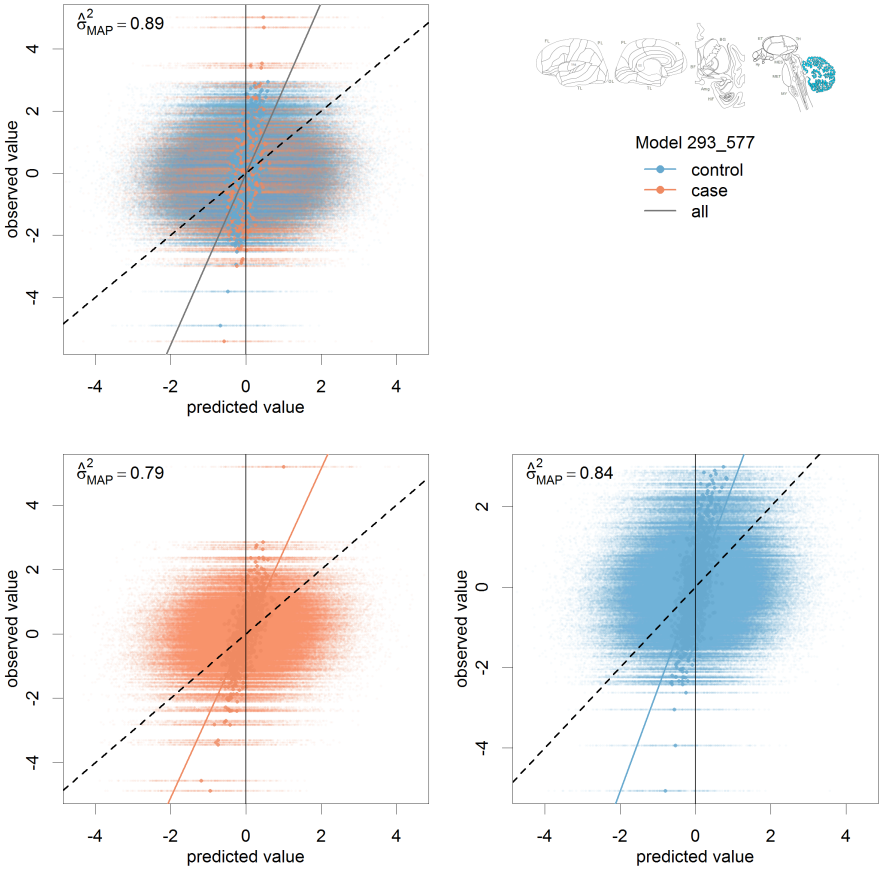


Figure 5.14: Characterisation of model 293_577, with the β prior precision set to the number of samples. For further explanation, see the caption of Figure 5.9.

Table 5.2: The FreeSurfer regions and their linked ABA regions.

short name	linked ABA identifiers	long name
G_and_S_paracentral	PCLa-i	Paracentral lobule and sulcus
G_and_S_cingul-Ant	CgGf-s; CgGf-i	Anterior part of the cingulate gyrus and sulcus (ACC)
G_and_S_cingul-Mid-Ant	CgGf-s; CgGf-i	Middle-anterior part of the cingulate gyrus and sulcus (aMCC)
G_and_S_cingul-Mid-Post	CgGf-s; CgGf-i	Middle-posterior part of the cingulate gyrus and sulcus (pMCC)
G_cingul-Post-dorsal	CgGp-s	Posterior-dorsal part of the cingulate gyrus (dPCC)
G_cingul-Post-ventral	CgGp-s	Posterior-ventral part of the cingulate gyrus (vPCC, isthmus of the cingulate gyrus)
G_cuneus	Cun-pest; Cun-str	Cuneus (O6)
G_front_inf-Opercular	fro	Opercular part of the inferior frontal gyrus
G_front_inf-Orbital	orIFG	Orbital part of the inferior frontal gyrus
G_front_inf-Triangul	trIFG	Triangular part of the inferior frontal gyrus
G_front_middle	MFG-s; MFG-i	Middle frontal gyrus (F2)
G_front_sup	SFG-m; SFG-l	Superior frontal gyrus (F1)
G_Ins_lg_and_S_cent_ins	LIG	Long insular gyrus and central sulcus of the insula
G_insular_short	SIG	Short insular gyri
G_occipital_sup	SOG-s	Superior occipital gyrus (O1)
G_oc-temp_lat-fusifor	OTG-s; OTG-i; FuG-its; FuG-cos	Lateral occipito-temporal gyrus (fusiform gyrus, O4-T4)
G_oc-temp_med-Lingual	LiG-pest; LiG-str	Lingual gyrus, ligual part of the medial occipito-temporal gyrus, (O5)
G_oc-temp_med-Parahip	PHG-l; PHG-cos	Parahippocampal gyrus, parahippocampal part of the medial occipito-temporal gyrus, (T5)
G_orbital	LOrG; MOrG	Orbital gyri
G_pariet_inf-Angular	AnG-s; AnG-i	Angular gyrus
G_pariet_inf-Supramar	SMG-s; SMG-i	Supramarginal gyrus
G_parietal_sup	SPL-s; SPL-i	Superior parietal lobule (lateral part of P1)
G_postcentral	PoG-cs; PoG-sl	Postcentral gyrus
G_precentral	PrG-prc; PrG-sl; PrG-il	Precentral gyrus
G_precuneus	Pcu-s; Pcu-i	Precuneus (medial part of P1)
G_rectus	GRe	Straight gyrus, Gyrus rectus
G_temp_sup-G_T_transv	HG	Anterior transverse temporal gyrus (of Heschl)
G_temp_sup-Lateral	STG-l; STG-i	Lateral aspect of the superior temporal gyrus
G_temp_sup-Plan_polar	PLP	Planum polare of the superior temporal gyrus
G_temporal_inf	ITG-its; ITG-l; ITG-mts	Inferior temporal gyrus (T3)
G_temporal_middle	MTG-s; MTG-i	Middle temporal gyrus (T2)
Cerebellum.Cortex	PV-V; PV-VI; He-VI; PV-Crus I; He-Crus I; He-Crus II; PV-VIIB; He-VIIB	Cerebellar cortex
Thalamus.Proper	DTA; ILc; LGd; DTLd; DTLv; DTM; ILr	Thalamus
Caudate	BCd; HCd; TCd	Caudate
Putamen	Pu	Putamen
Pallidum	GPi	Pallidum
Hippocampus	DG; CA1; CA2; CA3; CA4; S	Hippocampus
Amygdala	ATZ; BLA; BMA; CeA; COMA; LA	Amygdala

— 6 —

General discussion

In this thesis, we have explored several ways to use spatial gene expression data of the brain to inform studies into brain genetics. The methods and goals in the chapters are varied, but they have some strong connections.

All of the chapters in this thesis make use of the gene expression data of the Allen Human Brain Atlas. More specifically, they use similarities derived from this data. First, this can be in the form of gene-gene similarities. Genes have a strong similarity to other genes when they share a pattern of activity throughout the brain. In other words, genes are similar if they are transcribed in the same brain regions. Chapter 2 makes use of this type of similarity. Secondly, similarities can be calculated between brain regions. Annotated parts of the brain, such as the amygdala, are compared to other parts of the brain with respect to their transcriptional profiles. They are considered similar if the same genes are transcribed. Chapter 4 uses this region-region similarity. In Chapters 3 and 5, we use both gene-gene and region-region similarities.

In addition, all chapters have a focus on the integration of modalities, mainly of genetic variation and gene expression. In Chapters 4 and 5, we also include structural brain measurements. The genetic variant data and these structural brain measurements are acquired from study populations for specific brain disorders. The gene expression data, on the other hand, was measured in the healthy individuals of the Allen Human Brain Atlas data. This means the data is derived from disjoint samples, which makes integration quite challenging. In this case it is not possible to simply connect the data sets by concatenation, or by linking the identifiers of the individuals. We cannot equate a "row" in one data matrix to one in the other. Each chapter of this thesis has its own ways to deal with this issue.

In Chapter 2, we looked for molecular processes and brain regions involved in migraine, purely based on SNP data of migraine study participants. We did this in two ways. In the first method, we focussed on the small signals of sub-significant SNPs in gene clusters, and in the second on genes with high-confidence SNPs and their co-expression networks. The Allen Human Brain Atlas expression data was used to calculate the gene similarities required to build these gene clusters and co-expression networks. The two methods converged on a number of genomic functions and brain regions.

In Chapter 3 we visualised the Allen Brain Atlas gene expression data to get a visual overview of the patterns in the data. It shows the similarities between genes with respect to their activity in the brain, and similarities between brain regions with respect to their transcriptional profiles. So it focusses on the two types of similarities used throughout this thesis. The BrainScope portal helps to explore the Allen Human Brain Atlas data, but it also has a more general methodological contribution. Many recent studies have data that is both large (genome-wide for instance) and complex, because measurements were taken over time or across tissues. BrainScope shows an intuitive way to visualise this type of data.

In Chapters 4 and 5, we try to find more model-based ways to combine disease genomics with the brain transcriptome data. In Chapter 4, we use a structural equation modelling approach, which is more commonly used in psychometrics. Here, we encode the brain transcriptome data in the model structure to understand the relationship between genetic variation and variation in the volumes of brain areas in the context of Alzheimer's disease. In Chapter 5, we use a Bayesian method to incorporate the brain transcriptome data as prior information to study the relationship between variation in the genome and brain area volumes in the context of schizophrenia.

6.1 Generating and testing hypotheses

The gene expression data used in each of the chapters of this thesis is available in the Allen Human Brain Atlas. This data, as would an atlas of the earth, facilitates exploration. While questions about specific molecular pathways in specific areas of the brain can be answered using this resource, we performed mostly explorative studies. We were interested in which sets of genes or brain regions could be involved in migraine, Alzheimer's disease, or schizophrenia.

The results are far removed from clinical applications, and aim to point future experimental research in a better direction.

Explorative research was most famously advocated by John Tukey, as a necessary attitude and a way to find the right questions for us to answer (Tukey, 1980). In his view, we can distinguish three modes of data analysis, often performed as stages in a single study (Behrens & Yu, 2003). The first mode is purely explorative, and involves descriptive statistics and plots. In the second mode one can use probabilistic methods, but do so in a loose way. This "rough confirmatory" mode is an extension of the exploratory data analysis to see which models are promising. Many of the analyses in this thesis are of this type. In the third mode, the real confirmatory mode, hypotheses are tested in a formal way.

With exploratory data analysis, and data visualisations in particular, we can find unexpected patterns in data. It can also save us from making big mistakes in our analyses. It can be tempting to the analyst to directly jump from data collection into the final probabilistic analysis. The famous example of Anscombe's quartet (Anscombe, 1973) shows how this can go wrong. It has four bivariate datasets with identical summary statistics and ordinary least squares regression statistics. However, they are widely different in the relationships between the two variables, as can be seen by making scatter plots of them (Figure 6.1). After looking at the plots, one would not describe these data sets as identical, but rather proceed with controlling for outliers or choosing non-linear models for example.

In light of this, it is good to explore our data before fitting our models of interest. This presents a challenge, though. An exploratory phase should be allowed to influence the confirmatory phase, since we should have an open mind during exploration. But if the explorations influences the set of hypotheses that are tested, we introduce a bias. We will be more likely to test hypotheses that already seemed to be true, and avoid those for which we saw no evidence in exploration. The effective number of tests for which we would need to correct in the confirmatory phase is then not the number of tests that we actually perform, but all those that we could have considered before exploration. This was also well-known by Tukey, who urged to be very careful in real confirmatory studies, relying on experiments with randomisation and a very limited set of hypotheses (Tukey, 1980). The best way to go about this, is to collect new data for the confirmatory analyses.

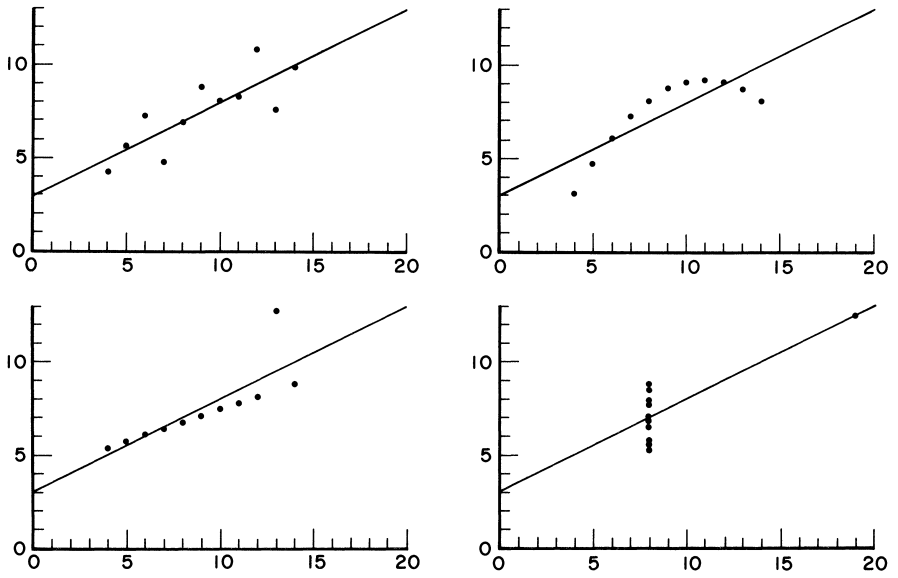


Figure 6.1: Scatterplots of four datasets ("Anscombe's quartet") with identical linear regression statistics. Reprinted from "Graphs in Statistical Analysis," by F. J. Anscombe, 1973, The American Statistician, 27(1), 19–20.

Some data sets, however, are one-of-a-kind. Although several data collections of gene expression in the human brain are now available (Mahfouz et al., 2017), these are so different in their set-up that they cannot be considered to be true replicates. To avoid hypothesis creation and testing on the same data, we could use different subsets of the data, analogous to how model parameters are estimated in a cross-validation. However, to fully understand the data, and for instance detect outliers, some exploration of the whole dataset is needed. Moreover, when the same dataset is used in multiple studies by the same researcher, the parts that are left out in one study can still influence the hypotheses posed in the next. In the chapters of this thesis we stayed in the rough confirmatory mode, and left the formal confirmation to future studies.

Probabilistic analyses, rough or confirmatory, can be performed as part of at least two distinct paradigms. Chapters 2 to 4 are frequentist in nature. Chapter 5 makes some use of Bayesian ideas. These paradigms may contain very similar methods, but they are philosophically quite distinct (Wakefield,

2013). In a frequentist setting, the statistics we calculate on our data are random characterisations of a true and fixed, but unknown, model. If we were to repeat our experiment, the estimated parameters would on average get closer and closer to the true parameters. Bayesian inference, on the other hand, considers the parameters themselves to be random. This follows from a different definition of probability. Arguably, probabilistic statements are often about personal knowledge or belief. Researchers have a mental model of the truth, and change their beliefs when confronted with data. In a Bayesian setting, there is a prior probability distribution for each parameter, and with data we can update these to obtain posterior distributions.

In Chapter 5, we used the Allen Human Brain Atlas data to propose models. In effect, we used this gene expression to specify which imaging genetic models had a high prior probability of being informative. Then we updated that information by fitting these models on data obtained from schizophrenia patients and healthy controls. This is a simple application of the Bayesian idea, but it shows a strong practical advantage. We can use the prior to incorporate information from data of a different type. The prior allows for the integration of data from different modalities, measured on different individuals. And, as stated, this was a common challenge in all chapters of this thesis.

Bayesian analysis also has its disadvantages. The subjective probability definition and, more specifically, the use of prior distributions is not welcomed by all statisticians (Gelman, 2008). The main reason is that personal bias, even if formalised, should not be part of the scientific process. Practically, Bayesian analyses can also be challenging, since they usually require a good understanding of probability distributions and complex algebra and computation. Finally, Bayesian inference is often less standardised than frequentist alternatives. For instance, multiple testing correction using family wise error rate or false discovery rate control has a strong tradition in frequentist inference. In a Bayesian setting we can use explicit models for the testing process (Wakefield, 2013), prior probabilities for null models in Bayes Factors (Westfall et al., 1997), or one could say that correct specification of parameter priors makes multiple testing correction unneeded (Wakefield, 2013).

6.2 The brain

One aspect of the exploratory nature of our analyses is that we consider all parts of the brain for which we have full data. In some cases, this limits the scope. In Chapter 5 for instance, we leave out all FreeSurfer measurements for ventricles, since they don't contain the tissues we have gene expression data for. This means we could miss some important signals. On the other hand, our analyses might include data on more brain regions than necessary. It can be wise to be more restrictive, by ruling out parts of the data beforehand. If we had focussed on the areas for which some involvements in the disease of interest were known, it would have lowered the multiple testing burden. In calculating gene-gene similarities across the brain, including irrelevant regions may add noise. In the end, this is a balance between keeping all options open, and making use of valuable prior information.

The chapters in this thesis are region-of-interest based, in a broad sense. That is, the whole brain is divided up in anatomically labelled regions, and these are the level of measurement for analysis. More agnostic methods use brain scan data on a voxel level (Bigos & Weinberger, 2010). Often these methods rely on a dimension reduction step in the analysis, to reduce the multiple testing burden and increase the interpretability of the results (J. Liu & Calhoun, 2014). Our choice to use region data was mostly determined by the fact that we combined datasets on the level of anatomical labels. Mapping from one anatomical atlas to another is a challenge in general, especially if they use different coordinate systems (Evans et al., 2012). The Allen Human Brain Atlas contains expression data that is measured on dissected samples of anatomically labelled areas. These samples do not cover the whole brain, but are "chunks" taken from the brain to represent the regions from which they were taken. As a result, the expression data is less suitable to be combined with other modalities on a voxel level.

The types of brain data considered in this thesis are limited to micro-array gene expression measurements and features derived from structural magnetic resonance imaging (structural MRI). Other types of brain measurements, such as those obtained by functional or diffusion weighted MRI provide more dynamic information, and diffusion tensor images give information on brain networks. Since we use region-region similarity information from the Allen Human Brain Atlas, combining these modalities would be interesting. It would

focus more on a dynamic connected brain, rather than a static "molecular brain".

6.3 The genetics

The genetics of brain disease are complex. Even when only single nucleotide polymorphisms (SNPs) are considered, they can affect brain phenotypes in a number of ways. SNPs found in GWAS are not necessarily causal. This could of course be due to linkage disequilibrium (LD) with causal SNPs, but also because they can be markers for structural variations that are ignored in GWAS (Brodie, Azaria & Ofran, 2016). The causal variations can affect protein function if they lie in coding regions, or abundance if they affect gene regulation or splicing. Rare variant information is becoming more widely available, but due to the costs of whole exome and whole genome sequencing, and the challenges in statistical analysis, many population studies still focus on common variants found with SNP arrays. Nevertheless, whole genome data can capture much more of the meaningful genetic variation (Telenti et al., 2016).

The genetic data used in this thesis is of two types. On the one hand, we consider the common genetic variants in disease populations. On the other hand, we look at the spatial gene expression in a small set of healthy individuals. In Chapters 2 and 5, the gene expression data is used to make groups (or modules) of genes. Because genes within a module are similar in their expression patterns across the brain, and co-expression is linked to similarities in function (as can be seen in Chapter 3), these modules are considered to be meaningful in a brain context. Note, though, that these are still modules of genes, not of genetic variants. To attribute variants to modules, we rely on a mapping. Chapter 2 links variants to genes if they lie within a 15 kbp window around those genes. However, variants often affect genes up to 2 Mbp away (Brodie et al., 2016). Chapter 5 addresses this by also considering known eQTL SNPs, that were found in brain data. Still the link between SNPs and genes is far from perfect. As a result, tests for SNPs in gene modules may be underpowered.

6.4 The future

The field of genetics changes particularly fast. Only 15 years ago, the first full human genome was presented (International Human Genome Sequencing Consortium, 2004). A year later, high throughput second generation sequencing methods became commercially available (van Dijk, Auger, Jaszczyszyn & Thermes, 2014). These are still the most used sequencing techniques in human studies, but a third generation of technologies was announced about a decade ago (Schadt, Turner & Kasarskis, 2010). These latest techniques have the main advantage of producing very long reads (van Dijk, Jaszczyszyn, Naquin & Thermes, 2018). Currently, the cheapest way to find the genetic variations that are associated with phenotypic differences is not to perform sequencing, but to use a SNP array. The costs of second generation sequencing have dropped sufficiently for some large scale studies to consider whole exome (D. J. Liu et al., 2014) or even whole genome sequencing (Telenti et al., 2016). As a result, rare variants can be associated with phenotypes on a population level. This does introduce challenges in statistical analysis. If all variations are considered in a univariate way, the multiple testing correction should be more stringent than the Bonferroni correction for one million independent tests currently used in GWAS (Pulit, de With & de Bakker, 2017). Alternatively, testing can be performed on a gene or pathway level, by first calculating combined statistics (Auer & Lettre, 2015). These still rely on a mapping to genes, which can be a challenging activity especially for variants located in the 98% of the genome that is non-coding (Mooney et al., 2014).

The third generation of sequencing techniques extends the possibilities and challenges. Much more so than second generation sequencing, it reveals large structural variations. With a better reconstruction of individual genomes, the identified variations become more unique for the sampled individuals. In an extreme case, one could perform de novo assembly of each genome (Chaisson, Wilson & Eichler, 2015). Now the genetic variation between individuals can no longer be characterised in a simple data matrix, by scoring the presence of single nucleotide variations or small indels. Instead, this variation is best described in a graph, where sequences are represented by nodes and individuals follow their own unique paths (Paten, Novak, Eizenga & Garrison, 2017). Most of the currently used statistical tools for association studies will have to be updated to work with these new representations, or to deal with the even

larger number of potential features that can be extracted from this data.

The genetic characterisation of samples increases in detail, but at the same time the sampling itself becomes more detailed as well. Single cell genome sequencing has a major impact on cancer studies, where sequence variations between the cells in a single tumour can show its development (Gawad, Koh & Quake, 2016). To measure gene expression, single cell and single nucleus RNA sequencing are on the rise (Svensson, Vento-Tormo & Teichmann, 2018). In the brain, this is often done with the goal to characterise cell types (Tasic, 2018). With single cell or single nucleus RNA sequencing of dissected brain regions (Bakken et al., 2018), region specific brain disorders can be studied on a cell-type level. The current samples cover only selected brain regions, but in the long term brain-wide genome-wide single cell expression data of the brain will most likely become available. As is the case with third generation sequencing, the added resolution will come at a cost. Interpretation might only be informative on a region-of-interest level, which would require a way to integrate the data over the brain cells; and in the statistical analyses the multiple testing burden may again increase dramatically. Perhaps studies will either have to focus on specific questions, or rely on dimension reduction techniques.

Despite these challenges, the increase in detail and the growing amount of data will inevitably help our understanding of the brain on a molecular level. Models need to integrate data of different modalities, from different study samples. In this thesis we have discussed some ways of doing this. Ultimately, in silico models describing both connectivity and molecular biology of the brain will have to combine the new high resolution data to understand the brain as a whole.

References

Abbott, S. & Fairbanks, D. J. (2016, oct). Experiments on Plant Hybrids by Gregor Mendel. *Genetics*, *204*(2), 407–422. doi: 10.1534/genetics.116.195198

Akbarian, S., Liu, C., Knowles, J. A., Vaccarino, F. M., Farnham, P. J., Crawford, G. E., ... Sestan, N. (2015). The PsychENCODE project. *Nature Neuroscience*, *18*(12), 1707–1712. doi: 10.1038/nn.4156

Anscombe, F. J. (1973, feb). Graphs in Statistical Analysis. *The American Statistician*, *27*(1), 17–21. doi: 10.1080/00031305.1973.10478966

Anttila, V., Stefansson, H., Kallela, M., Todt, U., Terwindt, G. M., Calafato, M. S., ... Others (2010). Genome-wide association study of migraine implicates a common susceptibility variant on 8q22. 1. *Nature genetics*, *42*(10), 869.

Anttila, V., Winsvold, B. S., Gormley, P., Kurth, T., Bettella, F., McMahon, G., ... Palotie, A. (2013, aug). Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nature Genetics*, *45*(8), 912–917. doi: 10.1038/ng.2676

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., ... Dermitzakis, E. T. (2015, may). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. doi: 10.1126/science.1262110

Atias, N., Istrail, S. & Sharan, R. (2013, dec). Pathway-based analysis of genomic variation data. *Current Opinion in Genetics and Development*, *23*(6), 622–626. doi: 10.1016/j.gde.2013.09.002

Auer, P. L. & Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Medicine*, *7*(1), 16. doi: 10.1186/s13073-015-0138-2

Bakken, T. E., Hodge, R. D., Miller, J. A., Yao, Z., Nguyen, T. N., Aevermann, B., ... Tasic, B. (2018). Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS ONE*, *13*(12), 1–24. doi: 10.1371/journal.pone.0209648

Bakken, T. E., Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., ... Lein, E. S. (2016). Comprehensive transcriptional map of primate brain development. *Nature*, *535*(7612), 367–375. doi: 10.1038/nature18637

Batmanghelich, N. K., Dalca, A. V., Quon, G., Sabuncu, M. R. & Golland, P. (2016). Probabilistic Modeling of Imaging, Genetics and Diagnosis. *IEEE transactions on medical imaging*, *0062*(c), 1–1. doi: 10.1109/TMI.2016.2527784

Batmanghelich, N. K., Saeedi, A., Cho, M., Estepar, R. S. J. & Golland, P. (2015). Generative Method to Discover Genetically Driven Image Biomarkers. In A. C. F. Colchester

& D. J. Hawkes (Eds.), *Information processing in medical imaging* (Vol. 511, pp. 30–42). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-319-19992-4_3

Bayés, Á., van de Lagemaat, L. N., Collins, M. O., Croning, M. D. R., Whittle, I. R., Choudhary, J. S., ... Grant, S. G. N. (2011, jan). Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, *14*(1), 19–21. doi: 10.1038/nn.2719

Beaton, D., Kriegsman, M., Dunlop, J., Filbey, F. M. & Abdi, H. (2016). Imaging Genetics with Partial Least Squares for Mixed-Data Types (MiMoPLS). In H. Abdi, V. Esposito Vinzi, G. Russolillo, G. Saporta & L. Trinchera (Eds.), *The multiple facets of partial least squares and related methods: Pls, paris, france, 2014* (Vol. 173, pp. 73–91). Cham: Springer International Publishing. doi: 10.1007/978-3-319-40643-5_6

Behrens, J. T. & Yu, C.-H. (2003, apr). Exploratory Data Analysis. In *Handbook of psychology* (pp. 12–40). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/0471264385.wei0202

Ben-David, E. & Shifman, S. (2012, jan). Networks of neuronal genes affected by common and rare variants in autism spectrum disorders. *PLoS genetics*, *8*(3), e1002556. doi: 10.1371/journal.pgen.1002556

Bernard, A., Lubbers, L. S., Tanis, K. Q., Luo, R., Podtelezhnikov, A. A., Finney, E. M., ... Lein, E. S. (2012, mar). Transcriptional Architecture of the Primate Neocortex. *Neuron*, *73*(6), 1083–1099. doi: 10.1016/j.neuron.2012.03.002

Bettens, K., Sleegers, K. & Van Broeckhoven, C. (2013, jan). Genetic insights in Alzheimer's disease. *The Lancet Neurology*, *12*(1), 92–104. doi: 10.1016/S1474-4422(12)70259-4

Bigos, K. L. & Weinberger, D. R. (2010, nov). Imaging genetics—days of future past. *NeuroImage*, *53*(3), 804–809. doi: 10.1016/j.neuroimage.2010.01.035

Blokland, G. A., del Re, E. C., Mesholam-Gately, R. I., Jovicich, J., Trampush, J. W., Keshavan, M. S., ... Petryshen, T. L. (2017, oct). The Genetics of Endophenotypes of Neurofunction to Understand Schizophrenia (GENUS) consortium: A collaborative cognitive and neuroimaging genetics project. *Schizophrenia Research*. doi: 10.1016/j.schres.2017.09.024

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bouhaddani, S. E., Houwing-Duistermaat, J., Salo, P., Perola, M., Jongbloed, G. & Uh, H.-W. (2016, dec). Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, *17*(S2), S11. doi: 10.1186/s12859-015-0854-z

Boyle, E. A., Li, Y. I. & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. doi: 10.1016/j.cell.2017.05.038

Brodie, A., Azaria, J. R. & Ofran, Y. (2016). How far from the SNP may the causative genes be? *Nucleic Acids Research*, *44*(13), 6046–6054. doi: 10.1093/nar/gkw500

Bunyavanich, S., Schadt, E. E., Himes, B. E., Lasky-Su, J., Qiu, W., Lazarus, R., ... Weiss, S. T. (2014, aug). Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC medical genomics*, *7*(1), 48. doi: 10.1186/1755-8794-7-48

Butt, A. M., Fern, R. F. & Matute, C. (2014, nov). Neurotransmitter signaling in white matter. *Glia*, *62*(11), 1762–1779. doi: 10.1002/glia.22674

Calhoun, V. D., Liu, J. & Adalı, T. (2009, mar). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, *45*(1), S163–S172. doi: 10.1016/j.neuroimage.2008.10.057

Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. (2015, nov). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, *16*(11), 627–640. doi: 10.1038/nrg3933

Chanda, P., Huang, H., Arking, D. E. & Bader, J. S. (2013, jan). Fast association tests for genes with FAST. *PloS one*, *8*(7), e68585. doi: 10.1371/journal.pone.0068585

Chasman, D. I., Schurks, M., Anttila, V., de Vries, B., Schminke, U., Launer, L. J., ... Kurth, T. (2011, jun). Genome-wide association study reveals three susceptibility loci for common migraine in the general population. *Nature genetics*, *43*(7), 695–698. doi: 10.1038/ng.856

Chekouo, T., Stingo, F. C., Guindani, M. & Do, K.-A. A. (2016). A Bayesian predictive model for imaging genetics with application to schizophrenia. *Annals of Applied Statistics*, *10*(3), 1547–1571. doi: 10.1214/16-AOAS948

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., ... Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 128. doi: 10.1186/1471-2105-14-128

Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(SUPPL. 2), 305–311. doi: 10.1093/nar/gkp427

Chestkov, I., Jestkova, E., Ershova, E., Golimbet, V., Lezheiko, T., Kolesina, N., ... Kostyuk, S. (2018, jul). Abundance of ribosomal RNA gene copies in the genomes of schizophrenia patients. *Schizophrenia Research*, *197*, 305–314. doi: 10.1016/j.schres.2018.01.001

Chou, I.-j. J., Kuo, C.-f. F., Huang, Y.-s. S., Grainge, M. J., Valdes, A. M., See, L.-c. C., ... Doherty, M. (2017). Familial aggregation and heritability of schizophrenia and co-aggregation of psychiatric illnesses in affected families. *Schizophrenia Bulletin*, *43*(5), 1070–1078. doi: 10.1093/schbul/sbw159

Consortium, C.-D. G. o. t. P. G., Smoller, J. W. S., Kendler, K. K., Craddock, N., Lee, P. H., Neale, B. M. N., ... O'Donovan, M. M. M. (2013, apr). Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *The Lancet*, *381*(9875), 1371–1379. doi: 10.1016/S0140-6736(12)62129-1

Darmanis, S., Sloan, S. a., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., ... Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, *112*(23), 201507125. doi: 10.1073/pnas.1507125112

De Fusco, M., Marconi, R., Silvestri, L., Atorino, L., Rampoldi, L., Morgante, L., ... Casari, G. (2003, feb). Haploinsufficiency of ATP1A2 encoding the Na+/K+ pump alpha2 subunit associated with familial hemiplegic migraine type 2. *Nature genetics*, *33*(2), 192–196. doi: 10.1038/ng1081

Debanne, D., Campanac, E., Bialowas, A., Carlier, E. & Alcaraz, G. (2011, apr). Axon physiology. *Physiological reviews*, *91*(2), 555–602. doi: 10.1152/physrev.00048.2009

de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Computational Biology*, *11*(4), 1–19. doi:

10.1371/journal.pcbi.1004219

Destrieux, C., Fischl, B., Dale, A. & Halgren, E. (2010, oct). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. doi: 10.1016/j.neuroimage.2010.06.010

de Vries, B., Anttila, V., Freilinger, T., Wessman, M., Kaunisto, M. A., Kallela, M., ... Dichgans, M. (2016). Systematic re-evaluation of genes from candidate gene association studies in migraine using a large genome-wide association data set. *Cephalalgia*, *36*(7), 604–614.

de Vries, B., Eising, E., Broos, L. A. M., Koelewijn, S. C., Todorov, B., Frants, R. R., ... van den Maagdenberg, A. M. J. M. (2014, mar). RNA expression profiling in brains of familial hemiplegic migraine type 1 knock-in mice. *Cephalalgia : an international journal of headache*, *34*(3), 174–182. doi: 10.1177/0333102413502736

Dichgans, M., Freilinger, T., Eckstein, G., Babini, E., Lorenz-Depiereux, B., Biskup, S., ... Strom, T. M. (2005, jul). Mutation in the neuronal voltage-gated sodium channel SCN1A in familial hemiplegic migraine. *Lancet (London, England)*, *366*(9483), 371–377. doi: 10.1016/S0140-6736(05)66786-4

Ding, S.-L., Royall, J. J., Sunkin, S. M., Ng, L., Facer, B. A., Lesnar, P., ... Lein, E. S. (2016, nov). Comprehensive cellular-resolution atlas of the adult human brain. *Journal of Comparative Neurology*, *524*(16), 3127–3481. doi: 10.1002/cne.24080

Distefano, C., Zhu, M. & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11. doi: 10.1.1.460.8553

Du, L., Huang, H., Yan, J., Kim, S., Risacher, S. L., Inlow, M., ... Shen, L. (2016). Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. *Bioinformatics*, *32*(10), 1544–1551. doi: 10.1093/bioinformatics/btw033

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., ... Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi: 10.1038/nature11247

Evans, A. C., Janke, A. L., Collins, D. L. & Baillet, S. (2012). Brain templates and atlases. *NeuroImage*, *62*(2), 911–922. doi: 10.1016/j.neuroimage.2012.01.024

Farlow, M. R., He, Y., Tekin, S., Xu, J., Lane, R. & Charles, H. C. (2004). Impact of APOE in mild cognitive impairment. *Neurology*, *63*(10), 1898–1901. doi: 10.1212/01.WNL.0000144279.21502.B7

Ferrari, M. D., Klever, R. R., Terwindt, G. M., Ayata, C. & van den Maagdenberg, A. M. J. M. (2015, jan). Migraine pathophysiology: lessons from mouse models and human genetics. *The Lancet. Neurology*, *14*(1), 65–80. doi: 10.1016/S1474-4422(14)70220-0

Fields, R. D. (2008, dec). Oligodendrocytes changing the rules: action potentials in glia and oligodendrocytes controlling action potentials. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, *14*(6), 540–543. doi: 10.1177/1073858408320294

Fisher, R. (1950). Statistical methods for research workers. Biological monographs and manuals. No. V. *Statistical methods for research workers. Biological mono-graphs and manuals. No. V.*(11th ed.).

Fisher, R. A. (1932). *The bearing of genetics on theories of evolution.*

Franke, B., Stein, J. L., Ripke, S., Anttila, V., Hibar, D. P., van Hulzen, K. J. E., ... Sullivan, P. F. (2016, feb). Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept. *Nature Neuroscience*, *19*(3), 420–431. doi: 10.1038/nn.4228

Freilinger, T., Anttila, V., de Vries, B., Malik, R., Kallela, M., Terwindt, G. M., ... van den Maagdenberg, A. M. J. M. (2012, jul). Genome-wide association analysis identifies susceptibility loci for migraine without aura. *Nature genetics*, *44*(7), 777–82. doi: 10.1038/ng.2307

Gaiteri, C., Mostafavi, S., Honey, C. J., De Jager, P. L. & Bennett, D. A. (2016, jun). Genetic variants in Alzheimer disease — molecular and brain network approaches. *Nature Reviews Neurology*, *12*(7), 413–427. doi: 10.1038/nrneurol.2016.84

Gawad, C., Koh, W. & Quake, S. R. (2016, mar). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, *17*(3), 175–188. doi: 10.1038/nrg.2015.16

Gehrmann, J., Matsumoto, Y. & Kreutzberg, G. W. (1995). Microglia: Intrinsic immuneffector cell of the brain. *Brain Research Reviews*, *20*(3), 269–287. doi: 10.1016/0165-0173(94)00015-H

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*(3), 445–450. doi: 10.1214/08-BA318

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S. L., ... Singleton, A. B. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in Human Brain. *PLoS Genetics*, *6*(5), 29. doi: 10.1371/journal.pgen.1000952

Gibson, G. (2012, feb). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, *13*(2), 135–145. doi: 10.1038/nrg3118

Goeman, J. J. & Solari, A. (2014, may). Multiple hypothesis testing in genomics. *Statistics in medicine*, *33*(11), 1946–78. doi: 10.1002/sim.6082

Grange, P., Bohland, J. W., Okaty, B. W., Sugino, K., Bokil, H., Nelson, S. B., ... Mitra, P. P. (2014, apr). Cell-type-based model explaining coexpression patterns of genes in the brain. *Proceedings of the National Academy of Sciences*, *111*(14), 5397–5402. doi: 10.1073/pnas.1312098111

Granziera, C., Daducci, A., Romascano, D., Roche, A., Helms, G., Krueger, G. & Hadjikhani, N. (2014, apr). Structural abnormalities in the thalamus of migraineurs with aura: A multiparametric study at 3 T. *Human Brain Mapping*, *35*(4), 1461–1468. doi: 10.1002/hbm.22266

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., ... Pasaniuc, B. (2016, feb). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), 245–252. doi: 10.1038/ng.3506

Guyuron, B., Yohannes, E., Miller, R., Chim, H., Reed, D. & Chance, M. R. (2014, nov). Electron Microscopic and Proteomic Comparison of Terminal Branches of the Trigeminal Nerve in Patients with and without Migraine Headaches. *Plastic and Reconstructive Surgery*, *134*(5), 796e–805e. doi: 10.1097/PRS.0000000000000696

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C. M., Hulshoff Pol, H. E. & Kahn, R. S. (2013). Brain volumes in schizophrenia: A meta-analysis in over 18 000

subjects. *Schizophrenia Bulletin*, *39*(5), 1129–1138. doi: 10.1093/schbul/sbs118

Harno, H., Hirvonen, T., Kaunisto, M., Aalto, H., Levo, H., Isotalo, E., ... Farkkila, M. (2003, dec). Subclinical vestibulocerebellar dysfunction in migraine with and without aura. *Neurology*, *61*(12), 1748–1752. doi: 10.1212/01.WNL.0000098882.82690.65

Hawrylycz, M. J., Lein, E. S., Guillozet-Bongaarts, A. L., Shen, E. H., Ng, L., Miller, J. A., ... Jones, A. R. (2012, sep). An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, *489*(7416), 391–399. doi: 10.1038/nature11405

Hawrylycz, M. J., Miller, J. A., Menon, V., Feng, D., Dolbeare, T., Guillozet-Bongaarts, A. L., ... Lein, E. S. (2015). Canonical genetic signatures of the adult human brain. *Nature Neuroscience*, *18*(12), 1832–1844. doi: 10.1038/nn.4171

Hawrylycz, M. J., Ng, L., Page, D., Morris, J., Lau, C., Faber, S., ... Jones, A. R. (2011, nov). Multi-scale correlation structure of gene expression in the brain. *Neural networks*, *24*(9), 933–942. doi: 10.1016/j.neunet.2011.06.012

Hawrylycz, M. J., Sunkin, S. & Ng, L. (2015, feb). Spatial mapping of multi-modal data in neuroscience. *Methods*, *73*, 1–3. doi: 10.1016/j.ymeth.2015.01.011

Headache Classification Committee of the International Headache Society (IHS). (2013, jul). The International Classification of Headache Disorders, 3rd edition (beta version). *Cephalalgia*, *33*(9), 629–808. doi: 10.1177/0333102413485658

Hershey, A., Horn, P., Kabbouche, M., O'Brien, H. & Powers, S. (2012, jan). Genomic Expression Patterns in Menstrual-Related Migraine in Adolescents. *Headache: The Journal of Head and Face Pain*, *52*(1), 68–79. doi: 10.1111/j.1526-4610.2011.02049.x

Hershey, A. D., Tang, Y., Powers, S. W., Kabbouche, M. A., Gilbert, D. L., Glauser, T. A. & Sharp, F. R. (2004, nov). Genomic Abnormalities in Patients With Migraine and Chronic Migraine: Preliminary Blood Gene Expression Suggests Platelet Abnormalities. *Headache: The Journal of Head and Face Pain*, *44*(10), 994–1004. doi: 10.1111/j.1526-4610.2004.04193.x

Hibar, D. P., Kohannim, O., Stein, J. L., Chiang, M.-C. & Thompson, P. M. (2011). Multilocus Genetic Analysis of Brain Images. *Frontiers in Genetics*, *2*(October), 1–11. doi: 10.3389/fgene.2011.00073

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., ... Medland, S. E. (2015, jan). Common genetic variants influence human subcortical brain structures. *Nature*, *520*(7546), 224–229. doi: 10.1038/nature14101

Hilker, R., Helenius, D., Fagerlund, B., Skytthe, A., Christensen, K., Werge, T. M., ... Glenthøj, B. (2017). Heritability of schizophrenia and schizophrenia spectrum based on the nationwide Danish Twin Register. *Biological Psychiatry*(9), 1–7. doi: 10.1016/j.biopsych.2017.08.017

Huang, D., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., ... Lempicki, R. A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, *8*(9), R183. doi: 10.1186/gb-2007-8-9-r183

Huang, D. W., Sherman, B. T. & Lempicki, R. a. (2009, jan). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, *37*(1), 1–13. doi: 10.1093/nar/gkn923

Huisman, S. M., van Lew, B., Mahfouz, A., Pezzotti, N., Höllt, T., Michielsen, L., ...

Lelieveldt, B. P. (2017, jan). BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic Acids Research*, *45*(10), gkx046. doi: 10.1093/nar/gkx046

Huisman, S. M. H., Mahfouz, A., Batmanghelich, N. K., Lelieveldt, B. P. F. & Reinders, M. J. T. (2018, dec). A structural equation model for imaging genetics using spatial transcriptomics. *Brain Informatics*, *5*(2), 13. doi: 10.1186/s40708-018-0091-0

International Human Genome Sequencing Consortium. (2004, oct). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. doi: 10.1038/nature03001

Ji, S. (2013, jan). Computational genetic neuroanatomy of the developing mouse brain: dimensionality reduction, visualization, and clustering. *BMC bioinformatics*, *14*, 222. doi: 10.1186/1471-2105-14-222

Kaimal, V., Bardes, E. E., Tabar, S. C., Jegga, A. G. & Aronow, B. J. (2010, jul). ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Research*, *38*(Web Server), W96–W102. doi: 10.1093/nar/gkq418

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., ... Sestan, N. (2011, oct). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483–489. doi: 10.1038/nature10523

Kao, P. Y., Leung, K. H., Chan, L. W., Yip, S. P. & Yap, M. K. (2017). Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions. *Biochimica et Biophysica Acta - General Subjects*, *1861*(2), 335–353. doi: 10.1016/j.bbagen.2016.11.030

Kavanagh, D. H., Tansey, K. E., O'Donovan, M. C. & Owen, M. J. (2015). Schizophrenia genetics: emerging themes for a complex disorder. *Molecular Psychiatry*, *20*(1), 72–76. doi: 10.1038/mp.2014.148

Kim, D. (2015). Methods of integrating data to uncover genotype – phenotype interactions. *Nature Publishing Group*, *16*(2), 85–97. doi: 10.1038/nrg3868

Ko, Y., Ament, S. A., Eddy, J. A., Caballero, J., Earls, J. C., Hood, L. & Price, N. D. (2013, feb). Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proceedings of the National Academy of Sciences*, *110*(8), 3095–100. doi: 10.1073/pnas.1222897110

Kruit, M. C. (2004, jan). Migraine as a Risk Factor for Subclinical Brain Lesions. *JAMA*, *291*(4), 427. doi: 10.1001/jama.291.4.427

Lambert, J.-C. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., ... Amouyel, P. (2013, oct). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, *45*(12), 1452–1458. doi: 10.1038/ng.2802

Lauritzen, M. (1994). Pathophysiology of the migraine aura. *Brain*, *117*(1), 199–210. doi: 10.1093/brain/117.1.199

Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., ... Duchesnay, É. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage*, *63*(1), 11–24. doi: 10.1016/j.neuroimage.2012.06.061

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., ... Jones,

A. R. (2007, jan). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, *445*(7124), 168–76. doi: 10.1038/nature05453

Leo, L., Gherardini, L., Barone, V., De Fusco, M., Pietrobon, D., Pizzorusso, T. & Casari, G. (2011, jun). Increased Susceptibility to Cortical Spreading Depression in the Mouse Model of Familial Hemiplegic Migraine Type 2. *PLoS Genetics*, *7*(6), e1002129. doi: 10.1371/journal.pgen.1002129

Li, D. & Shafto, P. (2011). Bayesian Hierarchical Cross-Clustering. *Aistats2011*, *15*, 443–451.

Li, M.-X. X., Gui, H.-S. S., Kwan, J. S. H. & Sham, P. C. (2011, mar). GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *The American Journal of Human Genetics*, *88*(3), 283–293. doi: 10.1016/j.ajhg.2011.01.019

Li, M.-X. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. (2012, may). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*, *131*(5), 747–756. doi: 10.1007/s00439-011-1118-2

Liu, C., Bousman, C. A., Pantelis, C., Skafidas, E., Zhang, D., Yue, W. & Everall, I. P. (2017). Pathway-wide association study identifies five shared pathways associated with schizophrenia in three ancestral distinct populations. *Translational psychiatry*, *7*(2), e1037. doi: 10.1038/tp.2017.8

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., ... Abecasis, G. R. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature Genetics*, *46*(2), 200–204. doi: 10.1038/ng.2852

Liu, J. & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in neuroinformatics*, *8*(March), 29. doi: 10.3389/fninf.2014.00029

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. doi: 10.1038/ng.2653

Ma, S. & Dai, Y. (2011, nov). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, *12*(6), 714–722. doi: 10.1093/bib/bbq090

Maaten, L. V. D. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

MacGregor, E. A. (2004, jun). Oestrogen and attacks of migraine with and without aura. *The Lancet Neurology*, *3*(6), 354–361. doi: 10.1016/S1474-4422(04)00768-9

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., ... McCarroll, S. A. (2015, may). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. doi: 10.1016/j.cell.2015.05.002

Mahfouz, A., Huisman, S. M. H., Lelieveldt, B. P. F. & Reinders, M. J. T. (2017, may). Brain transcriptome atlases: a computational perspective. *Brain Structure and Function*, *222*(4), 1557–1580. doi: 10.1007/s00429-016-1338-2

Mahfouz, A., van de Giessen, M., van der Maaten, L., Huisman, S., Reinders, M., Hawrylycz, M. J. & Lelieveldt, B. P. (2015, feb). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods*, *73*, 79–89. doi: 10.1016/j.ymeth.2014.10.004

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., ...

Stamatoyannopoulos, J. A. (2012, sep). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195. doi: 10.1126/science.1222794

McCarthy, M. J., Liang, S., Spadoni, A. D., Kelsoe, J. R. & Simmons, A. N. (2014). Whole brain expression of bipolar disorder associated genes: Structural and genetic analyses. *PLoS ONE*, *9*(6), e100204. doi: 10.1371/journal.pone.0100204

Medland, S. E., Jahanshad, N., Neale, B. M. & Thompson, P. M. (2014, jun). Whole-genome analyses of whole-brain data: working within an expanded search space. *Nature neuroscience*, *17*(6), 791–800. doi: 10.1038/nn.3718

Miller, J. A., Ding, S.-L. L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., ... Lein, E. S. (2014, apr). Transcriptional landscape of the prenatal human brain. *Nature*, *508*(7495), 199–206. doi: 10.1038/nature13185

Mooney, M. A., Nigg, J. T., McWeeney, S. K. & Wilmot, B. (2014, sep). Functional and genomic context in pathway analysis of GWAS data. *Trends in Genetics*, *30*(9), 390–400. doi: 10.1016/j.tig.2014.07.004

Morgan, T., Sturtevant, A., Muller, H. & Bridges, C. (1915). *The Mechanism of Mendelian Heredity*. New York: Henry Hold & Company.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., ... Beckett, L. (2005, nov). The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, *15*(4), 869–877. doi: 10.1016/j.nic.2005.09.008

Mulder, E. J., van Baal, C., Gaist, D., Kallela, M., Kaprio, J., Svensson, D. A., ... Palotie, A. (2003, oct). Genetic and Environmental Influences on Migraine: A Twin Study Across Six Countries. *Twin Research*, *6*(5), 422–431. doi: 10.1375/136905203770326420

Nagata, E., Hattori, H., Kato, M., Ogasawara, S., Suzuki, S., Shibata, M., ... Takagi, S. (2009, may). Identification of biomarkers associated with migraine with aura. *Neuroscience Research*, *64*(1), 104–110. doi: 10.1016/j.neures.2009.02.001

Neueder, A. & Bates, G. P. (2014). A common gene expression signature in Huntington's disease patient brain regions. *BMC medical genomics*, *7*, 60. doi: 10.1186/s12920-014-0060-2

Ng, B., White, C. C., Klein, H. U., Sieberts, S. K., McCabe, C., Patrick, E., ... De Jager, P. L. (2017). Brain xQTL Map: Integrating The Genetic Architecture Of The Human Brain Transcriptome And Epigenome. *bioRxiv*, 1–23. doi: 10.1101/142927

Ng, L., Bernard, A., Lau, C., Overly, C. C., Dong, H.-W., Kuan, C., ... Hawrylycz, M. J. (2009, mar). An anatomic gene expression atlas of the adult mouse brain. *Nature neuroscience*, *12*(3), 356–362. doi: 10.1038/nn.2281

Noseda, R. & Burstein, R. (2013, dec). Migraine pathophysiology: Anatomy of the trigeminovascular pathway and associated neurological symptoms, cortical spreading depression, sensitization, and modulation of pain. *Pain*, *154*, S44–S53. doi: 10.1016/j.pain.2013.07.021

Okada, N., Fukunaga, M., Yamashita, F., Koshiyama, D., Yamamori, H., Ohi, K., ... Hashimoto, R. (2016, oct). Abnormal asymmetries in subcortical brain volume in schizophrenia. *Molecular Psychiatry*, *21*(10), 1460–1466. doi: 10.1038/mp.2015.209

Ophoff, R. A., Terwindt, G. M., Vergouwe, M. N., van Eijk, R., Oefner, P. J., Hoffman, S. M., ... Frants, R. R. (1996, nov). Familial Hemiplegic Migraine and Episodic

Ataxia Type-2 Are Caused by Mutations in the Ca2+ Channel Gene CACNL1A4. *Cell*, *87*(3), 543–552. doi: 10.1016/S0092-8674(00)81373-2

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., ... Geschwind, D. H. (2013, nov). Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. *Cell*, *155*(5), 1008–1021. doi: 10.1016/j.cell.2013.10.031

Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. (2017, may). Genome graphs and the evolution of genome inference. *Genome Research*, *27*(5), 665–676. doi: 10.1101/gr.214155.116

Pearlson, G. D., Liu, J. & Calhoun, V. D. (2015). An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Frontiers in Genetics*, *6*(SEP), 1–13. doi: 10.3389/fgene.2015.00276

Pezzotti, N., Lelieveldt, B. P., Van Der Maaten, L., Höllt, T., Eisemann, E., Vilanova, A., ... Vilanova, A. (2017). Approximated and user steerable tSNE for progressive visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *23*(7), 1739–1752. doi: 10.1109/TVCG.2016.2570755

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... Pritchard, J. K. (2010, apr). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772. doi: 10.1038/nature08872

Porokhovnik, L. N., Passekov, V. P., Gorbachevskaya, N. L., Sorokin, A. B., Veiko, N. N. & Lyapunova, N. A. (2015, apr). Active ribosomal genes, translational homeostasis and oxidative stress in the pathogenesis of schizophrenia and autism. *Psychiatric Genetics*, *25*(2), 79–87. doi: 10.1097/YPG.0000000000000076

Pulit, S. L., de With, S. A. & de Bakker, P. I. (2017). Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations. *Genetic Epidemiology*, *41*(2), 145–151. doi: 10.1002/gepi.22032

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., ... Weale, M. E. (2014, aug). Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience*, *17*(10), 1418–1428. doi: 10.1038/nn.3801

Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C. Y., Purcell, S. M., Sklar, P., ... Daly, M. J. (2009, jun). Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLoS Genetics*, *5*(6), e1000534. doi: 10.1371/journal.pgen.1000534

Reyngoudt, H., Achten, E. & Paemeleire, K. (2012, aug). Magnetic resonance spectroscopy in migraine: What have we learned so far? *Cephalalgia*, *32*(11), 845–859. doi: 10.1177/0333102412452048

Richiardi, J., Altmann, A., Milazzo, A.-C., Chang, C., Chakravarty, M. M., Banaschewski,

T., ... Tahmasebi, A. (2015, jun). Correlated gene expression supports synchronous activity in brain networks. *Science*, *348*(6240), 1241–1244. doi: 10.1126/science.1255905

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., ... Consortium, R. E. (2015, feb). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. doi: 10.1038/nature14248

Rosseel, Y. (2012). lavaan : an R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–20.

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., ... Daly, M. J. (2011, jan). Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genetics*, *7*(1), e1001273. doi: 10.1371/journal.pgen.1001273

Sandor, P. S., Di Clemente, L., Coppola, G., Saenger, U., Fumal, A., Magis, D., ... Schoenen, J. (2005, feb). Efficacy of coenzyme Q10 in migraine prophylaxis: A randomized controlled trial. *Neurology*, *64*(4), 713–715. doi: 10.1212/01.WNL.0000151975.03598.ED

Sándor, P. S., Mascia, A., Seidel, L., De Pasqua, V. & Schoenen, J. (2001, may). Subclinical cerebellar impairment in the common types of migraine: A three-dimensional analysis of reaching movements. *Annals of Neurology*, *49*(5), 668–672. doi: 10.1002/ana.1019

Sangiorgi, S., Mochi, M., Riva, R., Cortelli, P., Monari, L., Pierangeli, G. & Montagna, P. (1994, feb). Abnormal Platelet Mitochondrial Function in Patients Affected by Migraine With and Without Aura. *Cephalalgia*, *14*(1), 21–23. doi: 10.1046/j.1468-2982.1994.1401021.x

Sauro, K. M. & Becker, W. J. (2009, oct). The Stress and Migraine Interaction. *Headache: The Journal of Head and Face Pain*, *49*(9), 1378–1386. doi: 10.1111/j.1526-4610.2009.01486.x

Schadt, E. E., Turner, S. & Kasarskis, A. (2010, oct). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–R240. doi: 10.1093/hmg/ddq416

Schoenen, J., Jacquy, J. & Lenaerts, M. (1998, feb). Effectiveness of high-dose riboflavin in migraine prophylaxis A randomized controlled trial. *Neurology*, *50*(2), 466–470. doi: 10.1212/WNL.50.2.466

Schu, M. & Lrp, P. Á. (2012). Genetics of migraine in the age of genome-wide association studies. *Journal of headache Pain*, *13*, 1–9. doi: 10.1007/s10194-011-0399-0

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L. M., Trojanowski, J. Q., ... Weiner, M. W. (2009). MRI of hippocampal volume loss in early Alzheimers disease in relation to ApoE genotype and biomarkers. *Brain*, *132*(4), 1067–1077. doi: 10.1093/brain/awp007

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. (2010, aug). Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLoS Genetics*, *6*(8), e1001058. doi: 10.1371/journal.pgen.1001058

Shekhar, K., Brodin, P., Davis, M. M. & Chakraborty, A. K. (2014). *Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)* (Vol. 111) (No. 1). doi: 10.1073/pnas.1321405111

Sjöstedt, E., Fagerberg, L., Hallström, B. M., Häggmark, A., Mitsios, N., Nilsson, P., ... Mulder, J. (2015). Defining the Human Brain Proteome Using Transcriptomics and Antibody-Based Profiling with a Focus on the Cerebral Cortex. *Plos One*, *10*(6), e0130028. doi: 10.1371/journal.pone.0130028

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., ... Kasprzyk, A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research*, *43*(W1), W589–98. doi: 10.1093/nar/gkv350

Sparaco, M., Feleppa, M., Lipton, R., Rapoport, A. & Bigal, M. (2006, apr). Mitochondrial Dysfunction and Migraine. *Cephalalgia*, *26*(4), 361–372. doi: 10.1111/j.1468-2982.2005.01059.x

Stein, J. L., Medland, S. E., Vasquez, A. A., Derrek, P., Senstad, R. E., Winkler, A. M., ... Dara, M. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nature Genetetics*, *44*(5), 552–561. doi: 10.1038/ng.2250.Identification

Stingo, F. C., Guindani, M., Vannucci, M. & Calhoun, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, *108*(503), 876–891. doi: 10.1080/01621459.2013.804409

Stys, P. K. (2011, aug). The axo-myelinic synapse. *Trends in Neurosciences*, *34*(8), 393–400. doi: 10.1016/j.tins.2011.06.004

Sun, Y. V. (2012, oct). Integration of biological networks and pathways with genetic association studies. *Human Genetics*, *131*(10), 1677–1686. doi: 10.1007/s00439-012-1198-7

Sunkin, S. M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T. L., Thompson, C. L., ... Dang, C. (2013, jan). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic acids research*, *41*(Database issue), D996–D1008. doi: 10.1093/nar/gks1042

Svensson, V., Vento-Tormo, R. & Teichmann, S. A. (2018, mar). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*, 599.

Tasic, B. (2018, jun). Single cell transcriptomics in neuroscience: cell classification and beyond. *Current Opinion in Neurobiology*, *50*, 242–249. doi: 10.1016/j.conb.2018.04.021

Telenti, A., Pierce, L. C. T., Biggs, W. H., di Iulio, J., Wong, E. H. M., Fabani, M. M., ... Venter, J. C. (2016, oct). Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences*, *113*(42), 11901–11906. doi: 10.1073/pnas.1613365113

Thompson, C. L., Ng, L., Menon, V., Martinez, S., Lee, C.-K. K., Glattfelder, K., ... Jones, A. R. (2014, jul). A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. *Neuron*, *83*(2), 309–323. doi: 10.1016/j.neuron.2014.05.033

Tolner, E. A., Houben, T., Terwindt, G. M., de Vries, B., Ferrari, M. D. & van den Maagdenberg, A. M. (2015, apr). From migraine genes to mechanisms. *PAIN*, *156*, S64–S74. doi: 10.1097/01.j.pain.0000460346.00213.16

Tottene, A., Conti, R., Fabbro, A., Vecchia, D., Shapovalova, M., Santello, M., ... Pietrobon, D. (2009, mar). Enhanced Excitatory Transmission at Cortical Synapses as the Basis for Facilitated Spreading Depression in CaV2.1 Knockin Migraine Mice.

Neuron, *61*(5), 762–773. doi: 10.1016/j.neuron.2009.01.027

Trabzuni, D., Ryten, M., Walker, R., Smith, C., Imran, S., Ramasamy, A., ... Hardy, J. (2011). Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *Journal of Neurochemistry*, *119*(2), 275–282. doi: 10.1111/j.1471-4159.2011.07432.x

Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician*, *34*(1), 23–25. doi: 10.1080/00031305.1980.10482706

Tzeng, J., Lu, H. H. & Li, W. H. (2008). Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, *9*, 179. doi: 10.1186/1471-2105-9-179

Van Erp, T. G., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., ... Turner, J. A. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Molecular Psychiatry*, *21*(4), 547–553. doi: 10.1038/mp.2015.63

van den Maagdenberg, A. M., Pietrobon, D., Pizzorusso, T., Kaja, S., Broos, L. A., Cesetti, T., ... Ferrari, M. D. (2004, mar). A Cacna1a Knockin Migraine Mouse Model with Increased Susceptibility to Cortical Spreading Depression. *Neuron*, *41*(5), 701–710. doi: 10.1016/S0896-6273(04)00085-6

van den Maagdenberg, A. M. J. M., Pizzorusso, T., Kaja, S., Terpolilli, N., Shapovalova, M., Hoebeek, F. E., ... Ferrari, M. D. (2010, jan). High cortical spreading depression susceptibility and migraine-associated symptoms in Ca v 2.1 S218L mice. *Annals of Neurology*, *67*(1), 85–98. doi: 10.1002/ana.21815

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. (2014, sep). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*(9), 418–426. doi: 10.1016/j.tig.2014.07.001

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. (2018, sep). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–681. doi: 10.1016/j.tig.2018.05.008

van Unen, V., Li, N., Molendijk, I., Temurhan, M., Höllt, T., van der Meulen-de Jong, A. E., ... Koning, F. (2016, may). Mass Cytometry of the Human Mucosal Immune System Identifies Tissue- and Disease-Associated Immune Subsets. *Immunity*, *44*(5), 1227–1239. doi: 10.1016/j.immuni.2016.04.014

Vied, C. M., Freudenberg, F., Wang, Y., Raposo, A. A. S. F., Feng, D. & Nowakowski, R. S. (2014). A multi-resource data integration approach: identification of candidate genes regulating cell proliferation during neocortical development. *Frontiers in Neuroscience*, *8*(August), 1–13. doi: 10.3389/fnins.2014.00257

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. doi: 10.1016/j.ajhg.2017.06.005

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., ... Geschwind, D. H. (2011, jun). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, *474*(7351), 380–384. doi: 10.1038/nature10110

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D. & Montana, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage*, *60*(1), 700–716. doi: 10.1016/j.neuroimage.2011.12.029

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. New York, NY: Springer New York. doi: 10.1007/978-1-4419-0925-1

Wang, Y., Zhang, X.-S. & Chen, L. (2018, apr). Integrating data- and model-driven strategies in systems biology. *BMC Systems Biology*, *12*(S4), 38. doi: 10.1186/s12918-018-0562-1

Westfall, P. H., Johnson, W. O. & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, *84*(2), 419–427. doi: 10.1093/biomet/84.2.419

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., ... State, M. W. (2013, nov). Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell*, *155*(5), 997–1007. doi: 10.1016/j.cell.2013.10.020

Wong, M. T., Chen, J., Narayanan, S., Lin, W., Anicete, R., Kiaang, H. T. K., ... Newell, E. W. (2015, jun). Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell Reports*, *11*(11), 1822–1833. doi: 10.1016/j.celrep.2015.05.022

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... Visscher, P. M. (2010, jul). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. doi: 10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. (2011, jan). GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. doi: 10.1016/j.ajhg.2010.11.011

Zhang, B. & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).

Zhang, Q., Burdette, J. E. & Wang, J.-P. P. (2014, dec). Integrative network analysis of TCGA data for ovarian cancer. *BMC systems biology*, *8*(1), 1338. doi: 10.1186/s12918-014-0136-9

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., ... Wu, J. Q. (2014, sep). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *Journal of Neuroscience*, *34*(36), 11929–11947. doi: 10.1523/JNEUROSCI.1860-14.2014

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P. & Wang, L. (2014, apr). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, *30*(7), 1006–1007. doi: 10.1093/bioinformatics/btt730

Summary

Medical studies are rarely easy, and it is especially challenging to understand brain disease. Brains are highly complex organs, and it is, for instance, hard to see the relationships between behavioural change in a person and the changes in the connections among the billions of cells in the brain that cause this behavioural change. Many brain related disorders, such as autism, schizophrenia, and Alzheimer's disease, have some genetic basis. They are influenced by small differences in people's genetic code, which are called variants. Genetic variants can cause differences in the activity or effectiveness of genes. And if genes are involved, knowing which genes these are, and what effect they have can help to find treatments for these diseases.

Sometimes a genetic variant has a very strong effect. In that case a disease occurs specifically in some families. The only thing we need to do then, is to look at the genetic differences within these families, and see which variants are present in the individuals that have the disease. However, many genetic variants don't cause a disease directly, but they have some impact on the chance of developing it. Often this chance is influenced by a large number of variants, all with a very small effect. To find a variant with a small effect, we need a large study sample. And if we look at a large study sample, with individuals that are not closely related, there are a lot of potential variants to look at. The human genome has around 20 000 genes, and millions of common variants that can affect these genes. Still, many studies have succeeded in finding common variants with small effects. To achieve this, these studies, called genome wide association studies (GWAS), include thousands or even millions of individuals.

Even with such large samples, much of the signal is lost in the noise. The statistical tests that are used to find the associations between genetic variants

and diseases suffer from a multiple testing problem. With each statistical test we do, we have some chance of making a mistake. To compensate for these mistakes, we make the tests more stringent. We set a higher bar for what we call statistically significant. And with a million tests, one for each genetic variant, this bar is set so high that we probably miss out on a lot of real associations.

In this thesis, we worked on a solution to this problem. Instead of doing the test per genetic variant, we tested per group of genes. Now we only have to correct our tests for the number of groups, rather than the number of variants. To do this, the variants have to be linked to genes, and the genes have to be split up in groups. These operations make sense from a biological standpoint, since variants have effects on genes, and genes work together with other genes in molecular pathways. A pathway is a set of genes with a shared function, where each gene is needed for one step in a cellular process. Not all pathways are well characterised, and not all pathways are important at all times and in all tissues. So the way we made our groups of genes, is to look at their activity (expression) in the brain. Genes that work together are often expressed together. So we looked at co-expression across the human brain to find the groups of genes.

Besides the statistical advantage of doing tests per group, this approach has some other advantages. Because genes work together, a mistake in one gene of a pathway can have the same effect as a mistake in another gene in that same pathway. For that reason, the interpretation of the results can be made easier by looking at pathways, or our groups that represent these pathways. Instead of pointing to a variant somewhere on the genome, we can identify a specific process that is important in the brain. A second advantage of our approach is that we can now say something about the brain areas that are important for the disease. The groups that we find, per definition, have a specific pattern of activity in the brain. If the genes in the group are active in, for instance, only the hippocampus, then the hippocampus may be of interest.

The methods that we propose can be used to interpret genome wide association studies. Rather than looking at single variants, we can now look at groups of genes with a shared function in a specific part of the brain. These results are exploratory. They will not lead directly to the development of a treatment, but they could help future researchers to design new studies. In this way, our methods make a small step in understanding human brain disease.

Samenvatting

Medische studies zijn zelden eenvoudig, en dat geldt in het bijzonder voor studies naar hersenziektes. De hersenen zijn een complex orgaan, en het is bijvoorbeeld lastig om het verband te zien tussen de ontwikkeling van verbindingen tussen de miljarden cellen in het brein en de gedragsverandering die daar het gevolg van is. Veel hersenaandoeningen, zoals autisme, schizofrenie en de ziekte van Alzheimer, worden beïnvloed door de genetica. De kleine verschillen tussen mensen in hun genetische code noemen we varianten. Genetische varianten kunnen een effect hebben op de activiteit of effectiviteit van genen. En als er genen betrokken zijn bij een ziekte, is het belangrijk om te weten welke genen dit zijn en wat hun functies zijn. Deze informatie kan dan gebruikt worden om medicijnen te ontwikkelen voor de behandeling van de ziekte.

Soms heeft een genetische variant een erg groot effect. In dat geval komt een ziekte specifiek in bepaalde families voor. Om er dan achter te komen wat de belangrijke varianten zijn, kunnen we kijken naar de genetische verschillen tussen leden van zo'n familie, en observeren welke varianten voorkomen in de familieleden die de ziekte hebben. Vaak is de situatie echter lastiger. Er is dan niet een enkele variant met een groot effect, maar een groot aantal varianten die elk een klein effect hebben op de kans om een ziekte te ontwikkelen. Om deze varianten met kleine effecten te vinden, hebben we een grote wetenschappelijke studie nodig, met veel deelnemers. En als we kijken naar een groot aantal mensen, die geen naaste familie van elkaar zijn, zijn er heel veel varianten om te beschouwen. Het menselijk genoom bevat ongeveer 20 000 genen en miljoenen varianten die een effect kunnen hebben op deze genen. Toch zijn studies erin geslaagd om veelvoorkomende varianten te vinden die een klein effect hebben. Om dit voor elkaar te krijgen, hebben deze *genome wide association studies*

(GWAS) duizenden of tegenwoordig zelfs miljoenen deelnemers.

Zelfs in zulke grote studies gaat veel van het signaal verloren in de ruis. Bij het statistisch testen voor de associaties tussen genetische varianten en ziektes treedt kanskapitalisatie op (het *multiple testing*-probleem). Bij iedere test die we doen, hebben we een kans om een toevallige fout te maken. Om te compenseren voor deze fouten, maken we onze testen strenger. De waargenomen associatie moet dan sterker zijn om nog als statistisch significant beschouwd te worden. Als we een miljoen testen doen, een voor elke variant, worden deze testen zo streng dat we waarschijnlijk een groot aantal echte associaties mislopen.

In dit proefschrift hebben we getracht dit probleem op te lossen. In plaats van te testen per variant, doen we dit per groep van genen. We hoeven de testen dan slechts te corrigeren voor het aantal groepen en niet voor het aantal varianten. Om dit te kunnen doen moeten we wel de varianten toeschrijven aan genen, en de genen opdelen in groepen. Biologisch zijn dit logische stappen, want varianten hebben een effect op genen en genen werken samen in moleculaire routes (*pathways*). Hierin werken genen (of eigenlijk de eiwitten waarvoor ze coderen) samen in een biologisch proces. Ieder gen zorgt voor een stapje in dit proces. Deze pathways zijn niet allemaal goed beschreven, en ze zijn niet altijd actief in alle weefsels van het lichaam. Daarom hebben wij groepen van genen gemaakt op basis van de activiteit van deze genen in de hersenen. Als genen samenwerken zijn ze vaak ook in dezelfde gebieden actief. We kunnen dus de correlatie tussen activiteiten van de genen in de menselijke hersenen gebruiken om informatieve groepen van genen te maken.

Naast het statistisch voordeel van deze aanpak, heeft het twee andere voordelen. In de eerste plaats kan de interpretatie eenvoudiger zijn. Omdat genen samenwerken, kan een variant in een gen hetzelfde effect hebben als een variant in een ander gen dat betrokken is bij hetzelfde proces. In plaats van dat we nu een variant aanwijzen, kunnen we iets zeggen over welk biologisch proces belangrijk is voor de ziekte. Het tweede voordeel van onze aanpak is dat we informatie krijgen over welke hersengebieden van belang zijn. De groepen die we hebben gedefinieerd hebben een specifiek patroon van activiteit in de hersenen. Als de genen in een geselecteerde groep bijvoorbeeld alleen actief zijn in de hippocampus, dan is de hippocampus wellicht van belang voor de ziekte.

De methoden die wij voorstellen kunnen dus gebruikt worden bij de inter-

pretatie van bepaalde genetische studies (GWAS). In plaats van naar enkele varianten te kijken, vinden we nu groepen van genen met gezamenlijke functies in specifieke delen van het brein. Deze resultaten zijn verkennend van aard. Ze zullen niet direct leiden tot de ontwikkeling van een medicijn, maar ze kunnen onderzoekers wel helpen bij het opzetten van nieuwe studies. Op deze manier leveren onze methodes een kleine bijdrage aan het begrip van hersenziektes bij de mens.

Acknowledgements

This thesis is not just a product created by me. Many people have contributed to its contents, either scientifically, just by keeping me sane, or both. In the first place, I would like to thank my promotors, Marcel Reinders and Boudewijn Lelieveldt. Marcel, you were always prepared to offer a critical view on my work, and have pushed me to keep going. Boudewijn, your enthusiasm for good visualisations and a clear message has been very valuable. Both of you have guided me through this process. I would also like to thank the members of my committee, Jelle Goeman, Patrick Groenen, Roeland van Ham, Peter-Bram 't Hoen, Danielle Posthuma, and Lodewyk Wessels, for reading my thesis and your comments. Asking critical questions is the basis of science.

During the PhD process, I have worked in and with a number of academic groups. In 2013, my roommates Erdogan, Sepideh, and Ahmed welcomed me into the world of bioinformatics. Ahmed, especially, guided me though a world of brains and RNAs and institutes, and then we spent many hours discussing work and life. Another colleague, Thies, quickly showed his appreciation for my interest in life outside work. Thank you, Thies, for your friendship and for being my paranymph. In our group in Delft, part bioinformaticians, part pattern recognisers, we had an ample supply of meetings and very intelligent people to learn from. I would like to thank everyone in the PRB group for our discussions. The same goes for the division of Image Processing (LKEB) in the LUMC, my part-time office. I never spent enough time getting to know you better. In particular, I would like to thank Baldur, for making BrainScope with me. Every time I need to impress someone with my work, I show them yours. Two other groups in the LUMC played an important part in my PhD. The first is the migraine lab, in Human Genetics. Else en Arn, it was a real pleasure writing a paper with you. The second is a new group with very familiar faces,

the Leiden Computational Biology Center. Finally, I spent a few months at the Department of Biomedical Informatics in Pittsburgh. Kayhan, thank you for our delightfully abstract discussions, and the warm welcome you gave me.

Some of the people I owe a lot to, I have never met. These are the participants of all the studies I used data from. Even though you will probably never see this, I hope you know we are grateful that you provide your data to people like me. During my PhD, I didn't collect a single bit of data myself, so I also owe my gratitude to the researchers that decided to record and share this information. For example, none of the chapters of this thesis could have done without the freely available data provided by the Allen Institute for Brain Sciences in Seattle.

Finally, I would like to thank those that perhaps understood little of what I was doing, but still helped me though this period. Marije, thank you for your support during all those years. Also, thank you, friends in Het Nonet, for giving me hundreds of calm, warm and musical Tuesday evenings; and the members of LSKO Collegium Musicum for doing the same on Mondays. Pimfandischa, thanks for being there in times of war and peace. Of course I would also like to thank my family. Hans and José, you have been my examples in life, always kind and welcoming, always in for a walk. Bregje, thank you for your kindness and support. I might not reply to your postcards, but I always decipher them with joy. Boi, thank you bro, and I am really sorry you don't get to wear the special monkey suit. And last and really not least, I would like to thank Anneke. You embrace all my moods, stressed out or chatty about the most boring subjects, and I am very happy that I can share these ends and beginnings with you.

List of publications

Mahfouz, A., van de Giessen, M., van der Maaten, L., Huisman, S., Reinders, M., Hawrylycz, M. J. & Lelieveldt, B. P. (2015, feb). Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods*, *73*, 79–89. doi: 10.1016/j.ymeth.2014.10.004

Eising, E., Huisman, S. M. H., Mahfouz, A., Vijfhuizen, L. S., Anttila, V., Winsvold, B. S., ... Reinders, M. J. T. (2016, apr). Gene co-expression analysis identifies brain regions and cell types involved in migraine pathophysiology: a GWAS-based study using the Allen Human Brain Atlas. *Human Genetics*, *135*(4), 425–439. doi: 10.1007/s00439-016-1638-x

Taskesen, E., Huisman, S. M. H., Mahfouz, A., Krijthe, J. H., de Ridder, J., van de Stolpe, A., ... Reinders, M. J. T. (2016, jul). Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Scientific Reports*, *6*, 24949. doi: 10.1038/srep24949

Huisman, S. M., van Lew, B., Mahfouz, A., Pezzotti, N., Höllt, T., Michielsen, L., ... Lelieveldt, B. P. (2017, jan). BrainScope: interactive visual exploration of the spatial and temporal human brain transcriptome. *Nucleic Acids Research*, *45*(10), gkx046. doi: 10.1093/nar/gkx046

Mahfouz, A., Huisman, S. M. H., Lelieveldt, B. P. F. & Reinders, M. J. T. (2017, may). Brain transcriptome atlases: a computational perspective. *Brain Structure and Function*, *222*(4), 1557–1580. doi: 10.1007/s00429-016-1338-2

Eising, E., Shyti, R., 't Hoen, P. A. C., Vijfhuizen, L. S., Huisman, S. M. H., Broos, L. A. M., ... van den Maagdenberg, A. M. J. M. (2017, may). Cortical Spreading Depression Causes Unique Dysregulation of Inflammatory Pathways in a Transgenic Mouse Model of Migraine. *Molecular Neurobiology*, *54*(4), 2986–2996. doi: 10.1007/s12035-015-9681-5

Huisman, S. M. H., Mahfouz, A., Batmanghelich, N. K., Lelieveldt, B. P. F. & Reinders, M. J. T. (2018, dec). A structural equation model for imaging genetics using spatial transcriptomics. *Brain Informatics*, *5*(2), 13. doi: 10.1186/s40708-018-0091-0