2-13-2014

PARALLEL INDEPENDENT COMPONENT ANALYSIS WITH REFERENCE FOR IMAGING GENETICS: A SEMI-BLIND MULTIVARIATE APPROACH

Jiayu Chen

Jiayu Chen

Candidate

Electrical and Computer Engineering

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the dissertation committee:

Vince D. Calhoun , Chairperson

Marios S. Pattichis

Jingyu Liu

Nora I. Perrone-Bizzozero

PARALLEL INDEPENDENT COMPONENT ANALYSIS WITH REFERENCE FOR IMAGING GENETICS: A SEMI-BLIND MULTIVARIATE APPROACH

by

JIAYU CHEN

B.S., Electronic and Communication Engineering,
Shanghai Jiao Tong University, 2002

M.S., Electrical and Computer Engineering,
University of New Mexico, 2006

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctorate of Philosophy
Engineering

The University of New Mexico
Albuquerque, New Mexico

Dec, 2013

DEDICATION

To

My dear parents

For their love, support and encouragement

ACKNOWLEDGEMENTS

My deepest gratitude is to my advisors, Dr. Vince D. Calhoun and Dr. Jingyu Liu, for their continuous support as well as generous guidance and encouragement over the past four years. I would also like to express my appreciation to my committee members, Dr. Marios S. Pattichis at Department of Electrical and Computer Engineering, and Dr. Nora I. Perrone-Bizzozero at Department of Neurosciences for their insightful suggestions and feedback on my research and dissertation. My extended gratitude goes to Dr. Godfrey D. Pearlson at Olin Neuropsychiatry Research Center and Dr. Jessica A. Turner at Department of Psychology and Neuroscience, Georgia State University for their detailed guidance on psychiatry and genetics. Special thanks are also given to members of Dr. Calhoun's medical image analysis lab (MIALab) for their efforts in data collection and preprocessing, as well as their support and friendship.

This work was performed at the University of New Mexico and the Mind Research Network, and supported by the National Institutes of Health, grant number: 5P20RR021938, R01EB005846, and 1R01MH094524-01A1.

PARALLEL INDEPENDENT COMPONENT ANALYSIS WITH REFERENCE FOR IMAGING GENETICS: A SEMI-BLIND MULTIVARIATE APPROACH

by

JIAYU CHEN

B.S., Electronic and Communication Engineering,
Shanghai Jiao Tong University, 2002

M.S., Electrical and Computer Engineering,
University of New Mexico, 2006

Ph.D., Engineering,
University of New Mexico, 2013

ABSTRACT

Imaging genetics is an emerging field dedicated to the study of genetic underpinnings of brain structure and function. Over the last decade, brain imaging techniques such as magnetic resonance imaging (MRI) have been increasingly applied to measure morphometry, task-based function and connectivity in living brains. Meanwhile, high-throughput genotyping employing genome-wide techniques has made it feasible to sample the entire genome of a substantial number of individuals. While there is growing interest in image-wide and genome-wide approaches which allow unbiased searches over a large range of variants, one of the most challenging problems is the correction for the huge number of statistical tests used in univariate models. In contrast, a reference-guided multivariate approach shows specific advantage for simultaneously assessing many variables for aggregate effects while leveraging prior information. It can improve the robustness of the results compared to a fully blind approach.

In this dissertation we present a semi-blind multivariate approach, parallel independent component analysis with reference (pICA-R), to better reveal relationships between hidden factors of particular attributes. First, a consistency-based order estimation approach is introduced to advance the application of ICA to genotype data. The pICA-R approach is then presented, where independent components are extracted from two modalities in parallel and inter-modality associations are subsequently optimized for pairs of components. In particular, prior information is incorporated to elicit components of particular interests, which helps identify factors carrying small amounts of variance in large complex datasets. The pICA-R approach is further extended to accommodate multiple references whose interrelationships are unknown, allowing the investigation of functional influence on neurobiological traits of potentially related genetic variants implicated in biology. Applied to a schizophrenia study, pICA-R reveals that a complex genetic factor involving multiple pathways underlies schizophrenia-related gray matter deficits in prefrontal and temporal regions. The extended multi-reference approach, when employed to study alcohol dependence, delineates a complex genetic architecture, where the CREB-BDNF pathway plays a key role in the genetic factor underlying a proportion of variation in cue-elicited brain activations, which plays a role in phenotypic symptoms of alcohol dependence. In summary, our work makes several important contributions to advance the application of ICA to imaging genetics studies, which holds the promise to improve our understating of genetics underlying brain structure and function in healthy and disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

In this dissertation we demonstrate the extended application of independent component analysis (ICA), a model for blind source separation, to the imaging genetics field. A consistency-based order estimation approach is designed for improved robustness and a semi-blind parallel ICA model is proposed to enable the analysis of large complex multi-modal datasets. Applications to real experimental data are also presented.

1. 1 Imaging Genetics

Imaging genetics explores the functional impact of genetic variations on neurological traits, making a valuable strategy for identifying biological mechanisms mediating the vulnerability to diseases.

Over the last decade, brain imaging techniques such as magnetic resonance imaging (MRI) (Giedd, 2004; Lawrie and Abukmeil, 1998; Paus et al., 2001) have been increasingly applied to study living brains. MRI can be used to measure morphometry, task-based function and structural and functional connectivity in the brain. Structural or functional imaging biomarkers are believed to be closer to the underlying biological mechanisms affected by genetic variants than behavioral or symptom-based measures (Rasch et al., 2010; Rose and Donohoe, 2013; Turner et al., 2006). Consequently, there is a growing interest in studying imaging measures. In the case of structural imaging, measurements can be obtained in different ways, ranging from single region-of-interest (ROI) methods, to image-wide approaches such as voxel based morphometry (VBM)

(Ashburner and Friston, 2005) and surface-based measures such as FreeSurfer (Fischl and Dale, 2000). In the case of functional imaging, blood-oxygen-level-dependent (BOLD) functional MRI (fMRI) has been widely used for studying the neural basis of spontaneous or task-related brain activities (Fox and Raichle, 2007; Logothetis, 2003).

High-throughput genotyping employing genome-wide techniques has made it feasible to sample the entire genome of a substantial number of individuals (Oliphant et al., 2002; Shen et al., 2005). Targeted candidate gene strategies examining a limited number of genetic variations have been successfully applied to the investigation of illnesses such as Fragile X syndrome (Lightbody and Reiss, 2009). Yet, the candidate gene approach is less applicable when the genetic basis of a disease is very complex and less understood. For instance, little success has been achieved in replicating evidence for causal genes in schizophrenia (SZ) (Duan et al., 2010) using a candidate gene approach. In contrast, recent work has lent support for a polygenic model (Gottesman and Shields, 1967; Purcell et al., 2009b) of the disorder, where an aggregate of common genetic variants were shown to collectively account for a substantial proportion of variation in risk. Indeed, besides SZ, a variety of complex mental disorders, including bipolar disorder, autism and addiction, are suggested as multifactorial and polygenic (Barnett and Smoller, 2009; Crabbe, 2002; Muhle et al., 2004). Given such evidence, an unbiased search of the entire genome bears more potential to delineate the underlying genetic architecture for complex disorders where a significant proportion of risk is likely ascribed to many genetic variants, each presenting a small effect size and failing to reach genome-wide significance individually.

1. 2 Motivation

While there is growing interest in image-wide and genome-wide approaches which allow unbiased searches over a large range of variants, novel mathematical and computational methods are desired to optimally combine two modalities. One of the most challenging problems is the correction for the huge number of statistical tests used in univariate models. The correction for multiple comparisons makes it highly difficult to identify a factor of small effect size with a practical sample size. In addition, univariate approaches are not well-suited to identify weak effects across multiple variables.

For this reason, multivariate approaches show specific advantage for simultaneously assessing many variables for an aggregate effect. In particular, ICA poses a promising candidate for modeling the linearly additive effect from multiple variables in the case of polygenicity. As a blind source separation approach, ICA tends to capture variants covarying with a same pattern into one single component, such that the data is decomposed into a linear combination of underlying components, whose associated loadings reflect the individual variations. In this way, genetic variants presenting small effect sizes on the same neurological or behavioral trait might be identified and the aggregate effect can be assessed.

1. 3 Specific Aims

Aim 1: Application of regular ICA to genotype data

The first aim of this project is to advance the application of regular ICA to genotype data. As a blind source separation approach, ICA has been widely used in signal and image processing (Comon, 1994; Hyvarinen et al., 2001). Meanwhile, to extend the application to genotype data, a major challenge lies in the order selection. This is due to the fact that genetic components in general accounting for small amounts of variance embedded in the genome, making it difficult to separate true signals from the background. To address this issue, an order selection approach is designed based on component stability.

Aim 2: Parallel ICA with reference

When applied to high-dimensional complex datasets, ICA suffers the curse of dimensionality. To address this issue, a semi-blind multivariate model, parallel ICA with reference (pICA-R), is proposed, such that prior knowledge is incorporated to guide the data decomposition and help elicit components related to particular attributes. Specifically, a closeness measure is imposed to extract independent components resembling the assigned reference**.** A formulization of the model is introduced in the dissertation, where the closeness is measured with L_2-norm Euclidian distance.

Aim 3: Parallel ICA with multiple references

A third aim is to extend pICA-R to accommodate multiple references for improved robustness. To achieve this, the reference input is designed as a matrix, with each row (or reference vector) representing a referential set spanning variants likely related while the interrelationships between different referential sets are to be investigated. The constrained component is dynamically selected for each referential set and multiple referential sets can constrain the same component.

Aim 4: Application of pICA-R to imaging genetics

The proposed approaches are applied to real experimental data to investigate the genetic underpinnings of abnormalities in brain structure and function involved in mental disorders. In this dissertation, we present results of schizophrenia (SZ) and alcoholism studies.

1. 4 Overview of Dissertation

The dissertation will be organized as follows:

Chapter 2 backgrounds the conducted research. The basic MRI technique for structural and functional brain imaging is explained. The concept of single nucleotide polymorphism (SNP) and a commonly used genotyping technique are also introduced. And a brief description is given to the principle and common implementations of ICA.

Chapter 3 introduces the consistency-based order estimation approach. The proposed approach successfully captures the order range where ICA extracts relatively more accurate components and loadings. We describe in detail the assumption and the mathematical model. Simulation is then presented to show the performance of the approach under different scenarios.

Chapter 4 introduces the pICA-R approach. The novel semi-blind multivariate model incorporates *a priori* knowledge to elicit components of specific attributes and assesses multiple variables for aggregate effects. A detailed description is given to the mathematical model and implementation. The approach has then been evaluated with simulated fMRI and SNP data. The results demonstrate the robustness of the approach and its applicability in real imaging genomics studies.

Chapter 5 introduces the extended parallel ICA with multiple references. When provided with multiple referential sets whose interrelationships are unknown, the extended approach successfully captures those associated with the same neurobiological trait through dynamic constraints for individual referential sets. The concept and formularization are described in detail. A comprehensive evaluation of the extended approach with simulated fMRI and SNP data is also presented.

Chapter 6 presents the investigation on scanning platform induced confounding effects in structural MRI (sMRI) studies. This is an issue difficult to avoid especially in imaging genetics as aggregated datasets are commonly employed to improve the sample size and statistical power. The initial exploration with a large sMRI dataset confirmed significant scanning effects from magnetic field strength, head coils and scanning sequences. A nonparametric correction was then designed and demonstrated with a second dataset to flexibly isolate scanning effects and refine the true effect of interest.

Chapter 7 demonstrates an application of pICA-R to the investigation of SZ. Voxelwise gray matter concentration data were analyzed in conjunction with genome-wide SNP data. Guided by a referential set derived from a susceptibility gene ANK3, pICA-R identified a significant sMRI-SNP association, revealing a complex genetic component underlying the SZ-related gray matter concentration reduction in frontal and temporal regions. The identified genetic component exhibited significant enrichment for SZ-relevance when independently assessed with the Psychiatric Genomic Consortium (PGC) data.

Chapter 8 presents an application of parallel ICA with multiple references to the investigation of alcohol dependence. Voxelwise cue-elicited brain activation data were analyzed in conjunction with genome-wide SNP data. A number of referential sets were derived from susceptibility genes implicated in previous studies. When assessed simultaneously, three referential sets derived from the CREB-BDNF pathway were identified as contributing to the same SNP component, which was significantly associated with a brain network reflecting hyperactivation in precuneus, superior parietal lobule and thalamus for more severe alcohol dependence. The identified genetic factor involved a number of neural signaling and development pathways implicated in a previous meta-analysis, confirming a complex and polygenic nature of the disorder.

Chapter 9 summarizes and concludes the project and provides some possibilities for future work.

CHAPTER 2 BACKGROUND

2. 1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a medical imaging technique developed to study internal structures and functions of the body. MRI images atomic nuclei by depicting their nuclear magnetic resonance properties and has a primary advantage over other imaging techniques in that it provides very high spatial resolution. In addition, MRI does not require ionizing radiation and yields good contrasts between different types of soft tissues, making it especially suitable for imaging living human brains.

In principle, an MRI image is a map that depicts the spatial distribution of a specific property of magnetized nuclear spins. This property might differ among tissues where the nuclear spins reside, allowing the morphology to be captured. The property might vary with the status of the body, which enables a dynamic imaging of responses to external stimuli. For human bodies, owing to the abundance of water, MRI signals are generally derived from the behavior of hydrogen atoms.

For a single hydrogen atom, the proton spins about itself due to thermal energy. This spin motion generates an electric current, which will induce a magnetic moment when situated within a magnetic field. Meanwhile, the spin also results in an angular momentum as the proton has an odd-numbered atomic mass. Under normal conditions, the spins are randomly oriented and the induced magnetization moments tend to cancel with each other, leading to a very small net magnetization.

When an external magnetic field is applied, the spinning proton will initiate a gyroscopic motion, known as precession, where the spin axis rotates around a central axis, as shown in Figure 2.1. The precession axis is aligned to the applied magnetic field and can possibly take two states: parallel (low-energy state) or antiparallel (high-energy state) to the magnetic field.

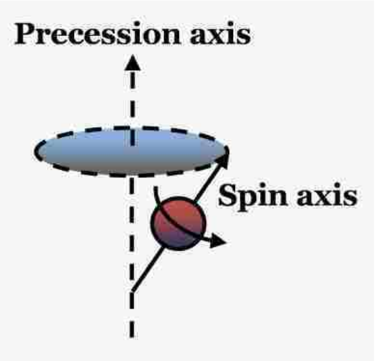

Figure 2.1: Precession.

The low-energy state is more stable and spins are more likely to assume this state under normal conditions. Meanwhile, through absorbing electromagnetic energy, spins can jump into the high-energy state, known as excitation. The corresponding resonance frequency is defined as the Larmor frequency (*v*), shown in (2.1). B_0 denotes the applied magnetic field and γ denotes the gyromagnetic ratio, which is the ratio of the magnetic moment over the angular momentum vector. For hydrogen, the Larmor frequency is within the radiofrequency (RF) band, around 42.58MHz/Tesla.

$$\nu = \frac{\gamma}{2\pi} B_0 \tag{2.1}$$

In the situation of receiving an excitation pulse, the precessing spins will be tipped from the longitudinal direction towards the transverse plane perpendicular to the precession axis. The net magnetization thus no longer aligns with B_0, but exhibits an angle, known as the flip angle. When the excitation is turned off, the tipped precession realigns back to the magnetic field, a phenomenon called relaxation. Figure 2.2 illustrates two primary relaxations: longitudinal and transverse relaxation. During longitudinal relaxation, excited spins return to the low-energy state, resulting in an increase of the longitudinal magnetization. The emitted energy is absorbed by the lattice of nuclei, also known as spin-lattice relaxation. The time constant associated with this recovery is called T_1 and the process is also called T_1-recovery. During transverse relaxation, the spins gradually lose phase coherence due to spin-spin interactions or field strength inhomogeneity, resulting in a decrease of the transverse magnetization. T_2 constant is used to characterize the decay induced by spin-spin interaction, while the overall decay is characterized by T_2^*.

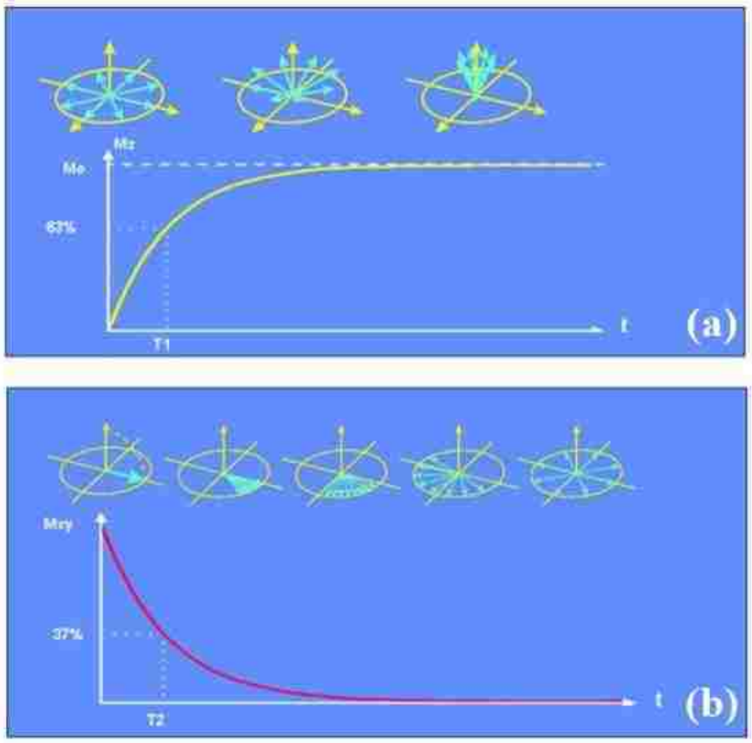

Figure 2.2: Schematic illustration of spin relaxation. (a) longitudinal/T1 relaxation; (b) transverse/T2 relaxation (Hashemi et al., 2003).

The varying net magnetization can be detected by receiver coils and is a measurable MR signal. To encode spatial information, a gradient magnetic field is introduced. In this way, the magnetic field strength varies systematically over space, resulting in different resonance frequencies. Thus a RF pulse can be applied to selectively excite a specific slice, known as slice selection. After the longitudinal net magnetization of one slice is tipped into the transvers plane, a frequency encoding and phase encoding gradient can be further applied to acquire the *k*-space MR signals. Finally the MR images can be easily reconstructed from the *k*-space signals using Fourier transform.

MRI can be versatilely configured to emphasize contrasts reflecting different tissue characteristics. As shown in Figure 2.3a, the recovery of longitudinal magnetization varies between tissues with different T_1 constants. This allows different levels of T_1-contrast to be obtained through adjusting the repetition time (TR, time interval between successive excitation pulses). T_1-weighted images are the most commonly used to study anatomical brain structures (sMRI), given a T_1 constant of ~900ms for gray matter (GM), ~600ms for white matter (WM) and 4200ms for cerebrospinal fluid (CSF). On the other hand, T_2-contrast reflects the difference in T_2 constants among the tissues and can be adjusted through the echo time (TE, time interval between the excitation and data acquisition), as shown in Figure 2.3b. T_2*-contrast is commonly used in BOLD fMRI, where brain functions are approximated by the associated changes in blood flow. The technique relies on the fact that when a brain region is in use or activated, the blood oxygen level increases and this oxygenation increases the T_2^* constant. Thus the visibility of blood vessels would reflect the regional brain activities in a T_2^*-contrast image.

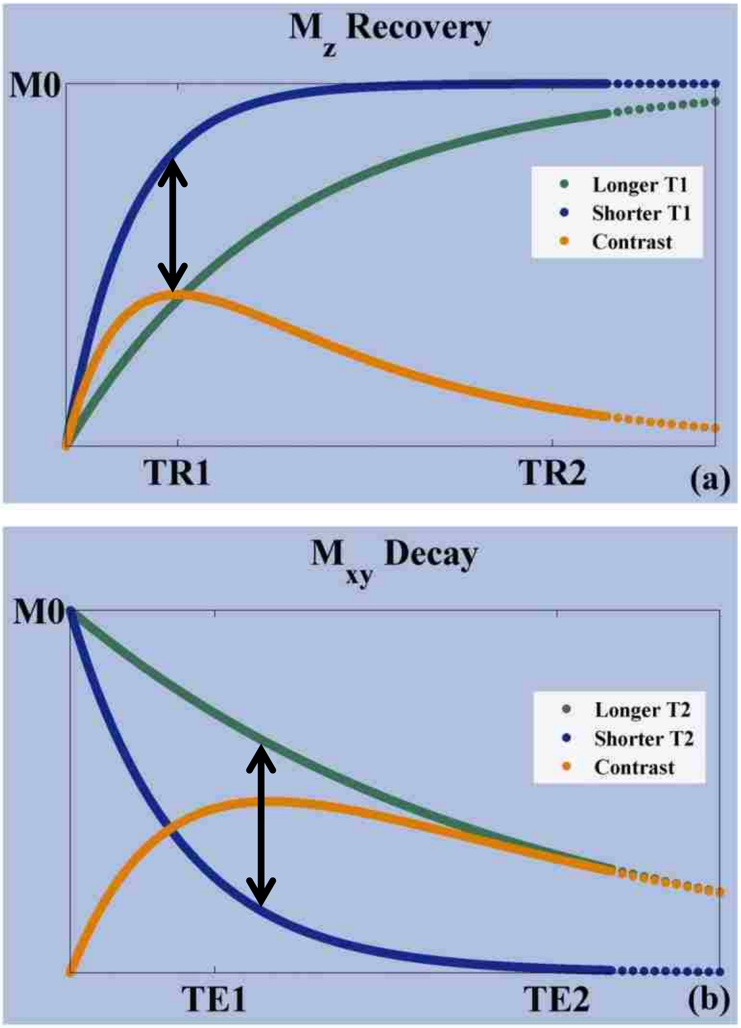

Figure 2.3: Schematic illustration of (a) T_1- and (b) T_2- contrasts.

2. 2 Single Nucleotide Polymorphism

DNA sequences may differ among members of a species at a single nucleotide, a type of genetic variation known as single nucleotide polymorphism (SNP), as shown in Figure 2.4. The alternate forms of base pairs are called alleles, of which the more frequently observed in one population is assigned as the major allele while the other as the minor allele. SNPs are highly abundant across the whole genome, with an occurrence rate estimated to be 1 out of every 300 bases for SNPs whose minor allele frequencies are higher than 1%, resulting in at least 10 million common SNPs out of 3 billion bases for the human genome (Gibbs et al., 2003).

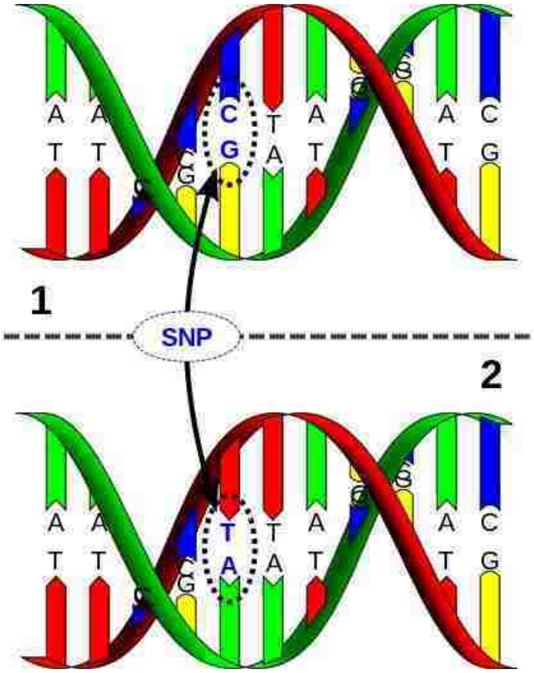

Figure 2.4: Illustration of a SNP with C/T polymorphism (use without permission from Wikipedia).

SNPs fall into both coding and non-coding regions and have different consequences at the phenotypic level. Polymorphisms in coding regions of genes may affect the structures or functions of the encoded proteins, which further contribute to diseases. For instance, The APOE (apolipoprotein E, chromosome 19) ε4 allele is a confirmed susceptibility factor for late-onset Alzheimer's disease (Farrer et al., 1997). On the other hand, the majority of SNPs reside in non-coding regions. Although their direct impacts on phenotypes are not known, these SNPs may still affect transcription factor binding, messenger RNA degradation, etc.

Consequently, SNPs pose as promising markers to locate genes that predispose individuals to diseases. The investigation can be performed through genotyping a collection of SNPs to identify those exhibiting allele frequency differences between control and patient groups. The number of SNPs that needs to be genotyped depends on the disease to be investigated. While genotyping a few SNPs in candidate genes may be adequate for monogenic disorders, explorations of complex and heterogeneous diseases whose genetic structures are less understood favor whole-genome SNP mapping. This is only made available upon the advance of the high-throughput genotyping technology.

One of the commonly used genotyping platforms is the BeadArray (Oliphant et al., 2002; Shen et al., 2005) from Illumina, based on which a whole-genome genotyping assay (Infinium) is developed (Gunderson et al., 2006). In the BeadArray technology, each array is assembled on an optical fiber bundle fused into a hexagonally packed matrix, as shown in Figure 2.5a. The fiber bundle is then exposed to the bead pool, such that the individual beads with distinct oligonucleotide probes can be assembled into the array. Figure 2.5b illustrates a scanning electron micrograph of an assembled array. Finally the arrays are arranged into a matrix to increase throughput.

The Infinium SNP genotyping assay consists of four steps: amplification, hybridization, SNP scoring and detection. Figure 2.6 illustrates the flowchart. Starting from a sample of 750ng, the DNA is whole-genome amplified 1000-2000× in an unbiased isothermal reaction (step 1-2). The amplified DNA then undergoes fragmentation, isopropanol precipitation and resuspension (step 3-4). Subsequently, the prepared sample is mounted to the BeadArray, where the designed probes capture the target loci through hybridization (step 5-6). Finally, the products are fluorescently stained (step 7), and the fluorescence intensities can be detected through BeadArray Reader (step 8). The data will then be analyzed to generate genotype calling using automated software.

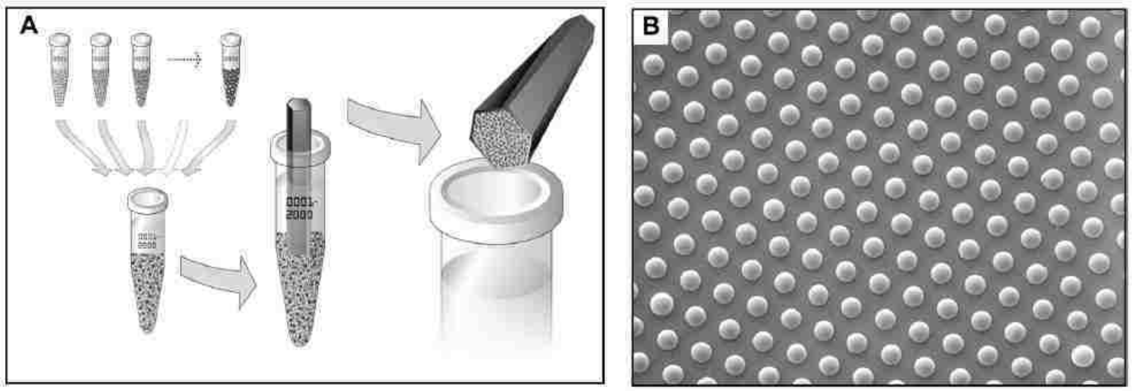

Figure 2.5: Assembly of an optical fiber array; (A) An etched fiber optic bundle is exposed to the bead pool, allowing individual beads to assemble into the microwells at the bundle's end; (B) Scanning electron micrograph of an assembled array containing 3-µm diameter silica beads ((Oliphant et al., 2002), Courtesy of Todd Dickinson).

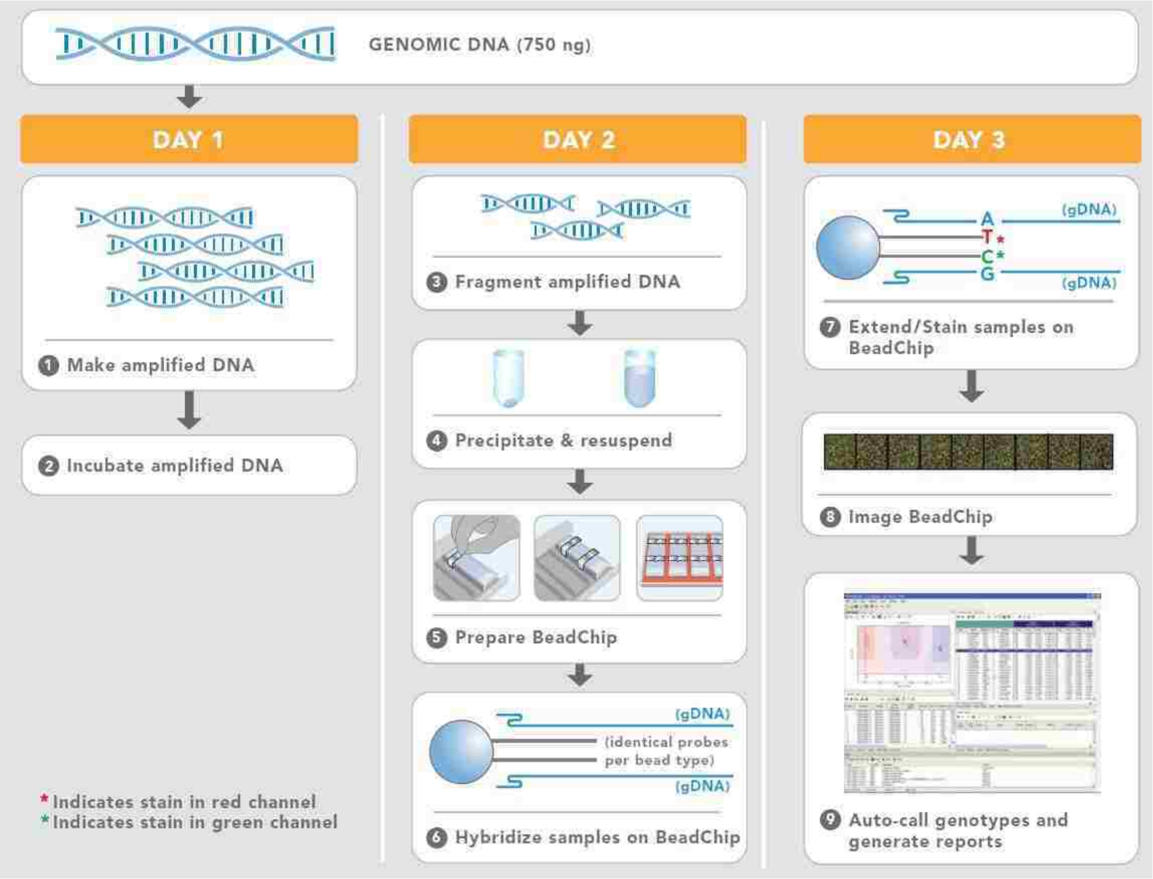
GENOMIC DNA (750 ng)
DAY 1
DAY 2
DAY 3
1 Make amplified DNA
2 Incubate amplified DNA
3 Fragment amplified DNA
4 Precipitate & resuspend
5 Prepare BeadChip
(gDNA)
(identical probes per bead type)
(gDNA)
6 Hybridize samples on BeadChip
A
T*
C*
G
(gDNA)
(gDNA)
7 Extend/Stain samples on BeadChip
8 Image BeadChip
9 Auto-call genotypes and generate reports
* Indicates stain in red channel
* Indicates stain in green channel

Figure 2.6: Illumina Infinium assay protocol (Gunderson et al., 2006).

2. 3 Independent Component Analysis

Independent component analysis (ICA) is a blind source separation method which has been widely used in many fields such as signal and image processing (Comon, 1994; Hyvarinen et al., 2001). In an ICA model, the observed data are treated as a linear combination of unknown independent sources, and the aim is to decompose the observed data and extract the sources through maximizing the independence among them. ICA becomes a popular technique for biomedical signal analysis, given that the measured signals are commonly mixtures of various underlying sources including both features of interest and noise/background signals.

A typical ICA model is shown below in (2.2), where the observed data are formed by a linear combination of the underlying sources, which are assumed to be not observable, statistically independent and non-Gaussian. **X** denotes the observed data with the dimension of sample (*M*) × feature (*N*) (typically, $N \geq M$). **S** denotes the unknown source matrix, where each row represents an independent component. L is the number of independent components (ICs). **A** denotes the unknown mixing matrix, with each column representing the loading coefficients associated with one independent component.

$$X_{M\times N} = A_{M\times L}S_{L\times N}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}, \ A = [a_1, \ \cdots \ a_L], \ S = \begin{bmatrix} s_1 \\ \vdots \\ s_L \end{bmatrix}$$

$$Y_{L\times N} = W_{L\times M}X_{M\times N} \tag{2.2}$$

In ICA, the underlying sources are extracted through estimating an unmixing matrix **W** such that **Y** is a good approximation to **S**, as shown in (2.2). The estimation of the unmixing matrix **W** is usually an iterative process where **W** is updated based on a specific objective function to optimize the independence among components. Several ICA algorithms have been implemented for the estimation of **W** where various independence metrics are employed, including Infomax, fastICA, JADE, EVD, and AMUSE (Bell and Sejnowski, 1995a; Cardoso and Soloumiac, 1993; Georgiev and Cichocki, 2001; Hyvarinen and Oja, 1997; Tong et al., 1990).

Among these ICA algorithms, Infomax (Amari, 1998; Bell and Sejnowski, 1995a) has been suggested as yielding reliable results for brain imaging data (Correa et al., 2007). More recently, the application of ICA was extended to genotype data (Chen et al., 2012c; Dawy et al., 2005; Liu et al., 2009) and showed great promise due to its multivariate nature, which helped identify components representing combined effects from multiple SNPs and associated with a given phenotype.

CHAPTER 3 CONSISTENCY-BASED ICA ORDER SELECTION

3. 1 Introduction

A commonly used ICA algorithms is Infomax (Amari, 1998; Bell and Sejnowski, 1995a), which has been suggested as yielding reliable results for brain imaging data (Correa et al., 2007). Infomax requires selecting the order, or the component number, before data decomposition. Information-theoretic criteria, such as Akaike information criterion (AIC) and minimal description length (MDL), have been employed (Akaike, 1973; Calhoun et al., 2001; Rissanen, 1978; Wax and Kailath, 1985) to solve this problem. In particular, a modified MDL criterion was specifically developed for ICA applied to functional magnetic resonance imaging (fMRI) data (Li et al., 2007).

The order selection is much more challenging for genotype data compared with MRI data, since in general genetic components account for small amounts of variance embedded in the genome (except for those accounting for the population structure), making it difficult to separate signal of interest from non-related information. In addition, a principal component analysis (PCA) data reduction is usually applied before Infomax-ICA to select out the same number of principal components accounting for the most variance of the data. This PCA reduction obviously does not guarantee the inclusion of information related to a genetic component carrying small variance. While using variance to identify the true component number works less effectively for genotype data, we observed that using consistency leads to relatively more accurate results. Thus, instead of using the information-theoretic criteria, we propose to select the order based on consistency for genotype data.

3. 2 Method

The proposed order selection procedure consists of three steps: ICA runs, consistency map construction and order selection.

ICA Runs

We apply Infomax-ICA to a given dataset $\mathbf{X}_{\mathrm{M\times N}}$ with different orders (denoted as *l*), as shown in (3.1). $\mathbf{S}^{l}$ and $\mathbf{A}^{l}$ respectively represent the components and loadings extracted by ICA with an order of *l*. The maximal tested order is denoted as *L*.

$$X^{l}{}_{M\times N} = A^{l}_{M\times l} \times S^{l}_{l\times N}\ ;\ (l = 2,3, \dots , L) \tag{3.1}$$

Consistency Map Construction

Given the ICA results from different tested orders, two consistency maps are constructed, one for components (**CS**) and the other for loadings (**CA**). The consistency evaluates the overall components' or loadings' similarity measured by correlations within a range of tested orders. Specifically, for the k^{th} component extracted in an ICA run with order *l* (denoted as $\mathbf{S}^{l}(k)$), we identify the most similar component extracted in the following ICA run with order *l+1* (denoted as $\mathbf{S}^{l+1}(k')$), and then record the absolute value of their correlation as an element **CS**(*k*,*l*) in the component consistency map, as shown in (3.2). This procedure is repeated for each component extracted in each ICA run, and thus the component consistency map, **CS**, is constructed as the upper triangular part of an *L*×*L* matrix. In a similar way, we construct the loading consistency map **CA**. Within the consistency matrices **CS** and **CA**, each column of the upper triangle reflects the overall consistency across all components or loadings extracted in one ICA run, while each row depicts the consistency evolution of one specific component or one set of loadings across all the tested orders.

$$CS(k,l) = abs\left[corr\left(S^{l}(k)\,,S^{l+1}(k')\right)\right] \tag{3.2}$$

Order Selection

In this step, we locate the desired order which leads to, relatively speaking, the most accurate components and loadings. Three strategies can be applied: overall consistency, reference-blind consistency, and reference-specific consistency.

A. Selection based on the overall consistency (overall)

Within the component consistency map, we focus on its upper triangle and calculate the mean of each column to obtain the overall component consistency $\mathbf{CS_{ova}}$ for each tested order *n*, as shown in (3.3). It is expected that the overall consistency remains stable with low orders and starts to decrease quickly when the increasing order results in a components over-splitting situation. Thus, the turning point provides a good guidance on the order selection. To avoid catching local oscillations, we search for a component order range, R_S, covering 10 consecutive tested orders, where the overall consistency exhibits the largest descending gradient (*G*). The above procedure is repeated for the loading consistency map and results in an order range R_A. Finally, to balance both component and loading consistencies, the median value of the overlapped range between R_S and R_A is selected as the final order, denoted as l_{sel}.

$$CS_{ova}(l) = \frac{1}{l}\sum_{k=1}^{l} CS(k,l)$$

$$G(\tilde{l}) = CS_{ova}(l) - CS_{ova}(l+9),\ \ \tilde{l} = \{l, \dots, l+9\}$$

$$R_s = \{\tilde{l}|max[G(\tilde{l})]\}$$

$$l_{sel} \in (R_s \cap R_A) \tag{3.3}$$

B. Selection based on the consistency of a reference

Given a component of interest, $\mathbf{S_r}$, as a reference, we select out from each ICA run one counterpart component $\mathbf{S_c}^{\boldsymbol{l}}$ that exhibits the most similar pattern to the reference. Then to evaluate the reference's consistency across tested orders, we apply a sliding window covering 10 consecutive orders and calculate the overall consistency CS_c (average of all pairwise correlations) among counterpart components within that window, as shown in (3.4). To avoid overfitting, among the windows exhibiting relatively high consistencies (>$CS_{c,th}$, chosen empirically), we select the leftmost to be the component order range, denoted as R_S. The above procedure is also repeated for the loadings, resulting in the order range R_A. Finally, to balance component and loading consistencies, the median value of the overlapped range between R_S and R_A is selected as the final order l_{sel}. Depending on the purpose of the study, the reference selection can be guided by the consistency map or phenotypical information, as described below:

$$CS_c(\tilde{l}) = mean\{abs[corr_{pairwise}(S_c^l, \dots, S_c^{l+9})]\}$$

$$CS_{c,th} = 0.9 \cdot median[CS_{c(top10)}]$$

$$R_s = min\{\tilde{l}|CS_c(\tilde{l}) > CS_{c,th}\}$$

$$l_{sel} \in (R_s \cap R_A) \tag{3.4}$$

Reference selected based on the consistency map (reference-blind): In the consistency map, a segment in a single row exhibiting consecutively high correlations indicates a high regional stability. The corresponding component is likely to be true and can serve as a good reference.

Reference selected based on phenotypical information (reference-specific): The selection of reference can also be guided by phenotypical information such as diagnoses labels. For instance, in a schizophrenia study, we can select a component whose loadings differentiate patients from controls as a reference.

3. 3 Simulation

We simulated a primary SNP dataset consisting of 200 samples and 5,000 SNP loci. 8 components were simulated using PLink (Purcell et al., 2007b), each involving 150 causal loci and a different case-control pattern. The causal loci exhibited different levels of effect sizes, ranging from 1.77 to 18.86 with a median of 2.20. Furthermore, we investigated the robustness of the procedure under different conditions, including effect size of causal loci, number of samples, number of SNP loci and number of true components.

ICA results derived from different orders were compared with the ground truth, and the average accuracies were calculated as a function of the tested order. Specifically, the component accuracy was evaluated by sensitivity, which is the ratio of correctly identified causal loci over the known true loci. The loading accuracy was reported as the absolute value of the correlation between the simulated case-control pattern and the extracted loadings. Based on the resulting accuracy, we examined whether the selected order would lead to the optimal results.

In the primary test, we performed ICA runs with orders ranging from 2 to 100 and then constructed the component and loading consistency maps, as shown in Figure 3.1, where the color map indicates the strength of correlation. All three selection strategies were tested. Using the overall consistency, the order was selected to be 19. Using the 8th component extracted with the order 17 as a reference (reference-blind), the order was selected to be 18. Using the case-control pattern of the first simulated component as a reference (reference-specific), the order was selected to be 21. The selected orders are marked in Figure 3.1 and 3.2, where Figure 3.1 shows the positions and consistency values of the selected orders in the two consistency maps, and Figure 3.2 provides a summary of the performance evaluation across tested orders, indicating that the selected orders lead to the optimal results.

3. 4 Results

The performances of the proposed procedure on datasets with different conditions are summarized in Figure 3.3-3.5, where the selected orders are marked and compared with other tested orders in terms of the resulting accuracies. It can be seen that we are mainly identifying the leftmost sliding window exhibiting an optimal accuracy. In general, the selected orders lead to relatively accurate components and loadings regardless of the ICA performances.

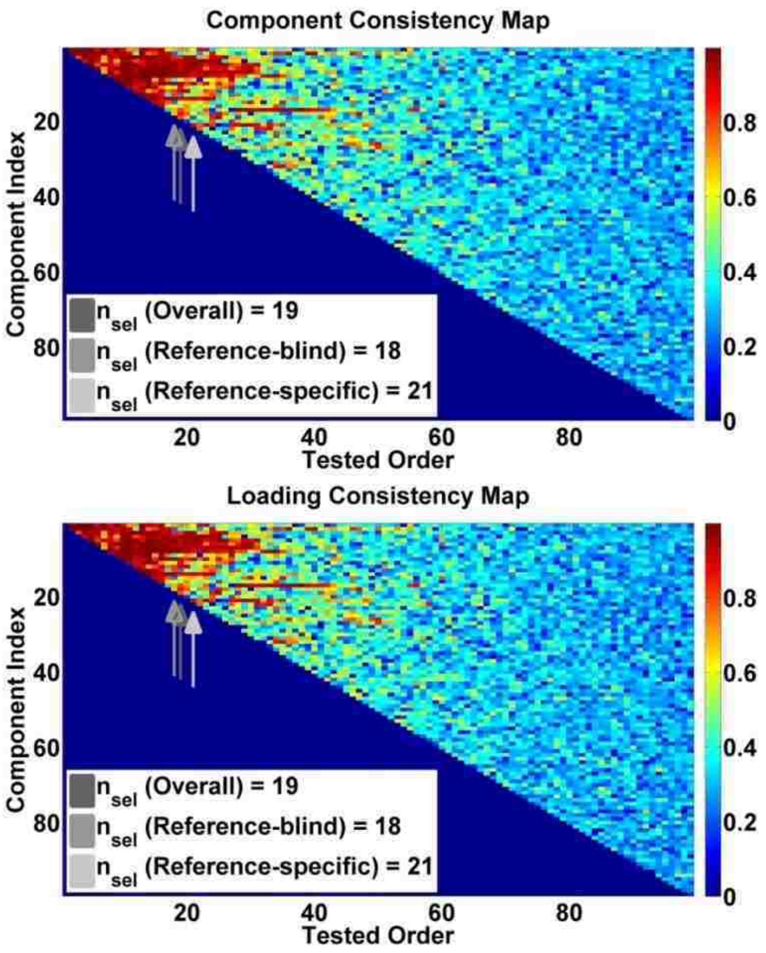

Figure 3.1: Component and loading consistency maps.

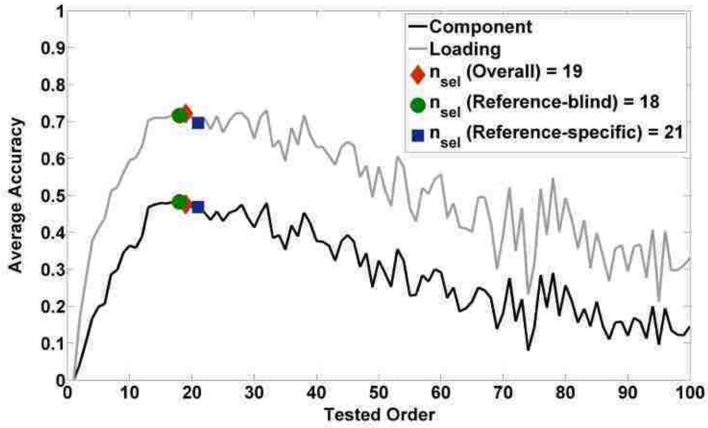

Figure 3.2: Performance evaluation of the primary test (200 samples, 5K SNP loci, 8 true components, median effect size of 2.20).

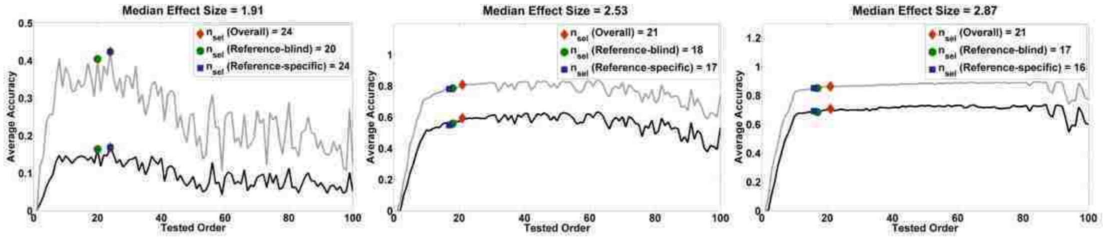

Figure 3.3: Performance evaluations on datasets with causal loci of different effect sizes (200 samples, 5K SNP loci, 8 true components). Black and gray lines represent component and loading accuracies respectively.

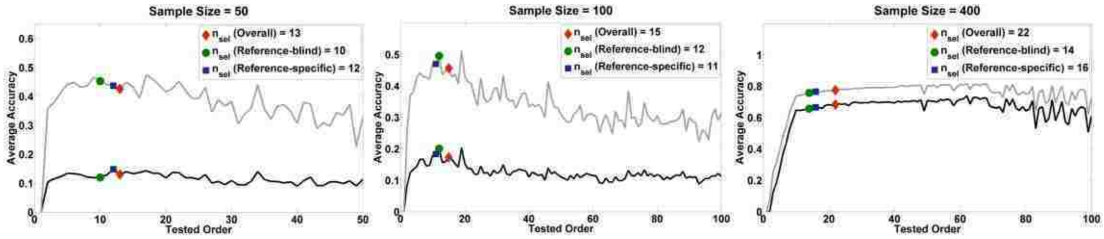

Figure 3.4: Performance evaluations on datasets with different sample sizes (5K SNP loci, 8 true components, median effect size of 1.99). Black and gray lines represent component and loading accuracies respectively.

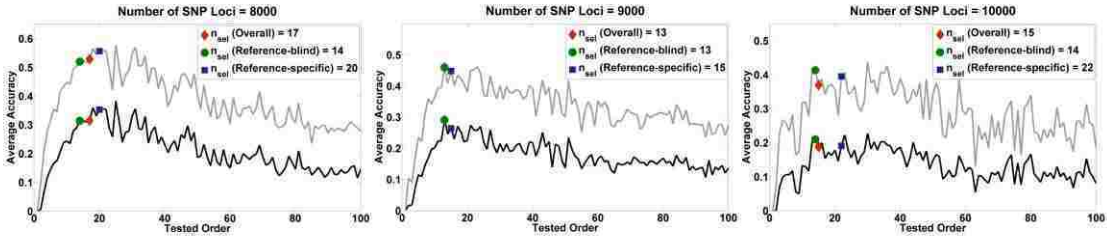

Figure 3.5: Performance evaluations on datasets with different numbers of SNP loci (200 samples, 8 true components, median effect size of 2.04). Black and gray lines represent component and loading accuracies respectively.

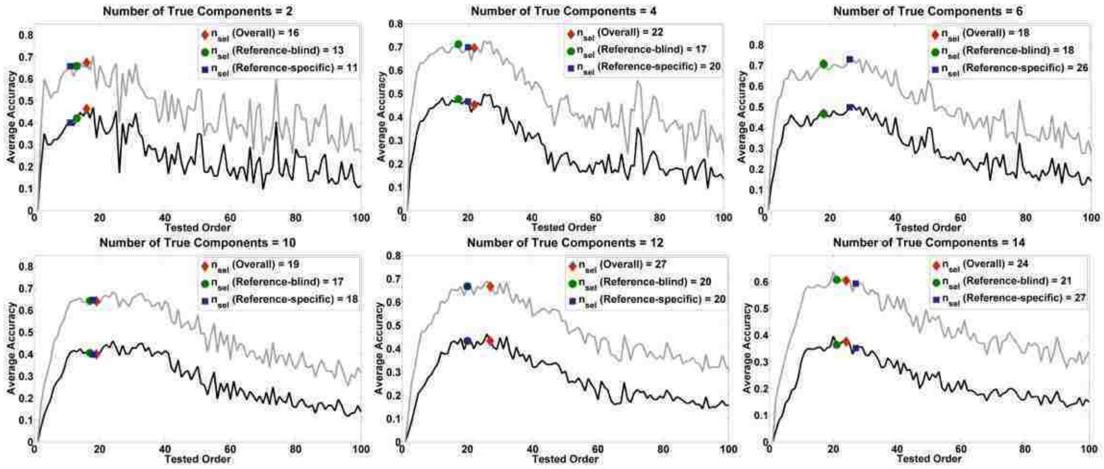

Figure 3.6: Performance evaluations on datasets with different numbers of true components (200 samples, 5K SNP loci, median effect size of 1.95). Black and gray lines represent component and loading accuracies respectively.

3. 5 Discussion

The proposed order selection procedure employs consistency as a criterion to locate the optimal order that results in relatively accurate components and loadings. Given its robustness, we expect that ICA can consistently extract a true component within a range of varying orders. This consistent region can be captured with different strategies, either through evaluating the overall consistency across all components or evaluating the consistency of a specific component across different orders, which can be selected based on regional stability or phenotypical information. Simulations demonstrate robust performances of all three strategies under different conditions.

Effect size of causal loci, number of samples and number of SNP loci: These varying conditions result in components accounting for different amounts of variance of the data. With a larger effect size, more samples or less input SNP loci, the simulated components account for more variance of the data than those with a smaller effect size, less samples or more input loci. When the components carry an adequate amount of variance, they can be accurately identified by ICA. In cases where ICA performs well, the order selection procedure accurately pinpoints the optimal order providing the best results. In cases where components are extracted with low accuracies, the proposed procedure still captures the range where relatively accurate components and loadings can be obtained, as shown in Figure 3.3-3.5.

Number of true components: We also simulated datasets with different numbers of true components, ranging from 2 to 14. Figure 3.6 summarizes the performance on these datasets. Overall, the proposed procedure exhibits robust performance where the selected order consistently leads to reasonable results regardless of varying numbers of true components. In addition, this evaluation clearly shows that, when a genetic component accounts for a small amount of variance, a true component number does not guarantee optimal results, since the component may be neglected in the PCA reduction applied before Infomax-ICA.

Among the three order selection strategies, the "overall" and the "reference-blind" methods are completely data-driven, while the "reference-specific" method involves phenotypical information. To investigate whether the selection of phenotypical information would affect the performance of the "reference-specific" method, we simulated components with different case-control patterns, yet always used the pattern of the first component to guide the reference selection. The simulation results indicate that the selected orders result in optimal average accuracies of all components and loadings regardless of the choice of phenotype. Thus we conclude that the reference selection can be guided by any phenotypical information and the performance of the procedure is not sensitive to this selection.

In summary, we design a procedure to select the ICA order based on consistency. The goal is to locate an order which allows ICA to extract relatively accurate, consistent components and loadings, while the components and background signal carry comparable variations. Three strategies have been implemented based on Infomax-ICA to achieve this goal. Simulation results indicate robust performances of all three strategies under different conditions and it is noteworthy that the procedure is able to select a reasonable order even when ICA operates less efficiently. While it awaits further evaluation with different ICA algorithms, we believe that there will be many applications for this procedure, not limited to genotype data, but any data with very low signal-to-noise ratio. Although the procedure proposed here is not mathematically ‘hard’ or ‘novel’, it will bring in great practical benefit for many researches.

CHAPTER 4 PARALLEL INDEPENDENT COMPONENT ANALYSIS WITH REFERENCE

4. 1 Introduction

Novel mathematical and computational methods are desired in imaging genetics to optimally combine the image-wide and genome-wide approaches which allow unbiased searches over a large range of variants. One of the most challenging problems is the correction for the huge number of statistical tests in univariate models. The correction makes it highly difficult to identify a factor of small effect size with a practical sample size. In addition, univariate approaches are not well-suited to identify weak effects across multiple variables. For this reason, multivariate approaches show specific advantage for simultaneously assessing many variables for an aggregate effect. To better identify aggregate effects across many variables, a number of models have been derived, including principal component regression (PCReg) (Wang and Abbott, 2008), sparse reduced-rank regression (sRRR) (Vounou et al., 2010) and parallel independent component analysis (pICA) (Liu et al., 2009).

PCReg, sRRR, and pICA are designed to deal with datasets of high dimensionality and yield interpretable results. However these approaches are not able to take prior information into account. Such information can be useful to enable a guided yet flexible approach and can improve the robustness of the results compared to a fully blind approach. For instance, some genes known to participate in a biological pathway critical to a disease may help identify a set of genes contributing in a coordinated way to a larger network. As observed in pilot studies, incorporation of prior information may be especially helpful in analyzing genomic data, where a component usually accounts for a small amount of variance in the data and is more difficult to identify (Chen et al., 2012c; Liu et al., 2012). Thus, we propose parallel independent component analysis with reference (pICA-R), which extends pICA to incorporate prior information to provide a reference to guide analyses. While pICA is designed based on regular (blind) ICA to enhance correlation between two modalities, pICA-R further takes advantage of *a priori* knowledge to guide the analysis and pinpoint a particular component of interest embedded in a large complex dataset. In this chapter, we compare pICA-R with other multivariate models through simulated data and evaluate the models under several scenarios.

4. 2 Method

Parallel independent component analysis with reference (pICA-R) is formulated by incorporating a reference constraint into parallel independent component analysis (pICA) (Liu et al., 2009) to guide the component extraction towards *a priori* knowledge. Typical pICA builds on regular infomax (Amari et al., 1996; Bell and Sejnowski, 1995b) to extract independent components in parallel for each modality, followed by a conditional enhancement of the inter-modality correlations. In comparison, pICA-R imposes an additional constraint upon the infomax framework to minimize the distance between a certain component and the reference. The mathematical model is shown below, and Figure 4.1 illustrates the flow of the approach.

Given a dataset **X** with dimension of sample (i.e., subjects) × feature (i.e., voxels [d=1], SNPs [d=2]), (4.1) illustrates the mathematical model of data decomposition, where

$$X_d = A_d S_d \rightarrow S_d = W_d X_d \,, A_d = W_d^{-1}, \quad d = 1, 2 \tag{4.1}$$

$$\begin{aligned} Y_d &= \frac{1}{1+e^{-U_d}}, U_d = W_d X_d + W_{d0} \\ F_1 &= max\{H(Y_1)\} = max\{-E[ln\, f_{y_1}(Y_1)]\} \\ F_2 &= max\{\lambda H(Y_2) + (1-\lambda)[-dist^2(\tilde{r}, |\tilde{S}_{2k}|)]\} \\ &= max\left\{\lambda(-E[ln\, f_{y_2}(Y_2)]) + (1-\lambda)\left(-\left\| |W_{2k}\tilde{X}_2| - \tilde{r} \right\|_2^2\right)\right\} \end{aligned} \tag{4.2}$$

$$F_3 = max\{\textstyle\sum_{i,j} Corr^2(A_{1i}, A_{2j})\} = max\left\{\textstyle\sum_{i,j} \frac{Cov^2(A_{1i}, A_{2j})}{Var(A_{1i})Var(A_{2j})}\right\} \tag{4.3}$$

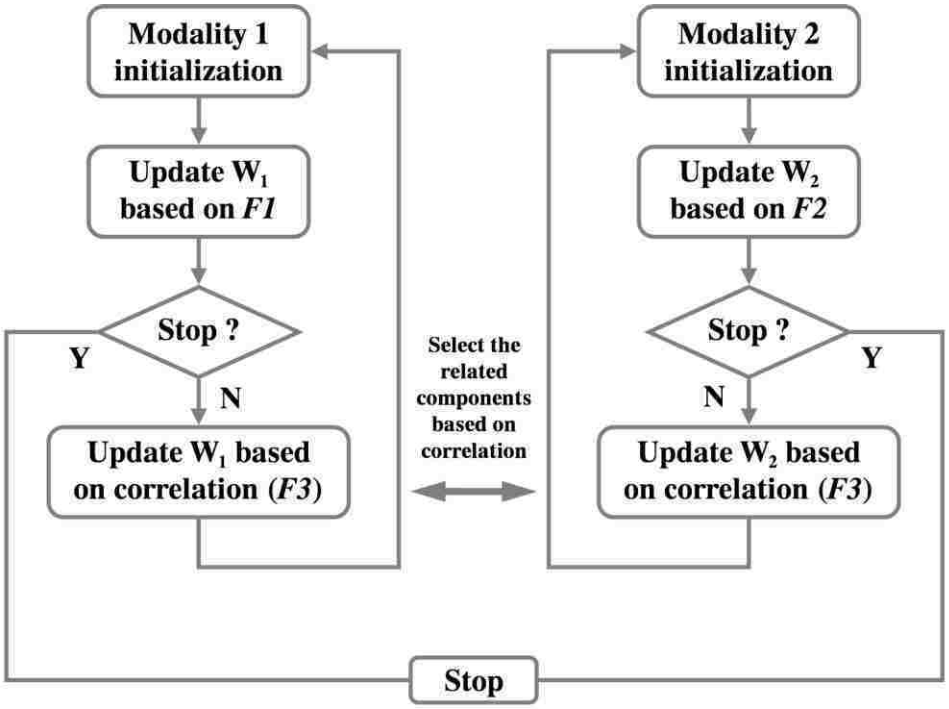
Modality 1 initialization
Update W₁ based on *F1*
Stop ?
Y
N
Update W₁ based on correlation (*F3*)
Select the related components based on correlation
Modality 2 initialization
Update W₂ based on *F2*
Stop ?
Y
N
Update W₂ based on correlation (*F3*)
Stop

Figure 4.1: Flow chart of pICA-R. W_1 and W2 denote the unmixing matrices of the two modalities, respectively. F_1, F_2 and F_3 represent the objective functions based on which unmixing matrices are updated.

the observed dataset **X** is decomposed into a linear combination of the underlying independent components, or sources. **S** is the component matrix, **A** is the loading or mixing matrix (estimated as the pseudo inverse of **W**), **W** is the unmixing matrix, and the subscript d runs from 1 to 2, denoting the data modality. Specifically, pICA-R iteratively solves the unmixing matrices $\mathbf{W_1}$ and $\mathbf{W_2}$ simultaneously for the two modalities, gradually maximizing the objective functions F_1, F_2 and F_3 in the manner described in Figure 4.1. In particular, F_1 is the objective function of the regular infomax (Bell and Sejnowski, 1995b) for modality 1, where independence among components is achieved by maximizing the entropy (*H)*, as shown in (4.2). f_y(Y) is the probability density function of **Y** and $\mathbf{W_0}$ is the bias vector. In contrast, F_2 is the objective function for modality 2, where an additional closeness metric is imposed to extract maximally independent components, one of which also closely resembles the reference **r**. The inter-modality correlation function F_3 shown in (4.3) is designed to maximize the correlations computed over the columns of the loading matrices $\mathbf{A_1}$ and $\mathbf{A_2}$, capturing connections between pairs of inter-modality components.

pICA-R incorporates an additional constraint to the unmixing matrix of modality 2 ($\mathbf{W_2}$), detaching itself from regular blind pICA. The objective function F_2 is shown in (4.2) and Figure 4.2 illustrates how the constraint is applied. In this application modality 2 is the genomic data. The reference **r** is a binary vector with the same number of loci as the genomic data, where the selected reference loci are set to “1” and the rest are “0”s. This binary reference effectively serves as a mask such that the closeness between the component and reference vector is measured on the reference loci only. This design considers that for a given reference a group of loci are presumably of interest and set to 1,

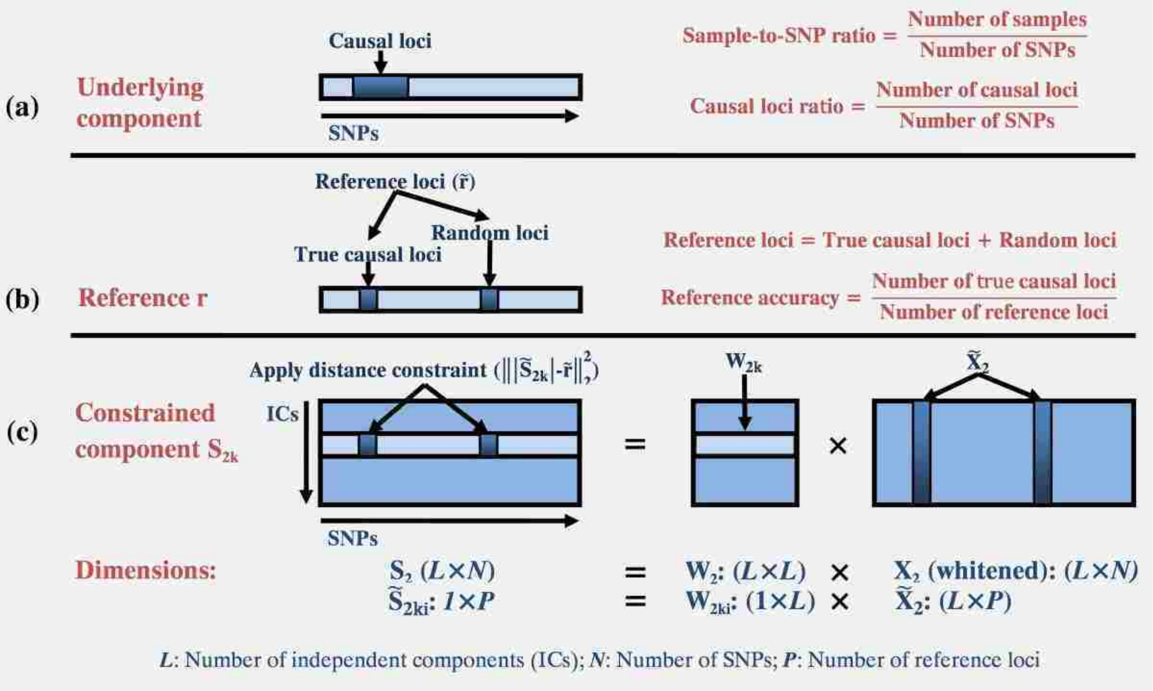
(a)
Underlying component
Causal loci
SNPs
Sample-to-SNP ratio = Number of samples / Number of SNPs
Causal loci ratio = Number of causal loci / Number of SNPs
(b)
Reference r
Reference loci (r̃)
Random loci
True causal loci
Reference loci = True causal loci + Random loci
Reference accuracy = Number of true causal loci / Number of reference loci
(c)
Constrained component S2k
Apply distance constraint (‖|S̃2k|-r̃‖²₂)
ICs
SNPs
W2k
X̃2
=
×
Dimensions:
S2 (L×N) = W2: (L×L) × X2 (whitened): (L×N)
S̃2ki: 1×P = W2ki: (1×L) × X̃2: (L×P)
L: Number of independent components (ICs); N: Number of SNPs; P: Number of reference loci

Figure 4.2: Illustration of the applied distance constraint: (a) the underlying component with highlighted causal loci (black region); (b) the generated reference, where r is the reference vector with selected reference loci set to 1 (gray region) and other loci set to 0. $\tilde{r}$ denotes a subvector consisting of all the reference loci; (c) the closeness is optimized specifically for the selected reference loci of one component. W_2 is the unmixing matrix of modality 2, X_2 is the data matrix and S_2 is the component matrix. $\tilde{S}_{2i}$ denotes a subvector of S_{2i} (the i^{th} row of S_2), W_{2i} denotes the i^{th} row of W_2 and $\tilde{X}_2$ denotes a submatrix of X_2.

however status of the remaining loci is to-be-determined instead of not interesting. Therefore, we choose to optimize the closeness specifically for the selected reference loci while allowing the remaining loci to show their own importance driven by data. This is equivalent to minimizing $\left\| |\tilde{\mathbf{S}}_{2\mathbf{k}}| - \tilde{\mathbf{r}} \right\|_2^2$ in F_2, where $\tilde{\mathbf{r}}$ denotes a subvector of $\mathbf{r}$, $\tilde{\mathbf{S}}_{2\mathbf{k}}$ denotes a subvector of $\mathbf{S}_{2\mathbf{k}}$ (the k^{th} row of $\mathbf{S}_2$), $\mathbf{W}_{2\mathbf{k}}$ denotes the k^{th} row of $\mathbf{W}_2$ and $\tilde{\mathbf{X}}_2$ denotes a submatrix of $\mathbf{X}_2$, as illustrated in Figure 4.2. $\|\cdot\|_2$ represents the L_2-norm Euclidian distance, and λ is a weighting parameter. It should be noted that we apply the constraint only to one modality in this work, which provides a simple proof-of-concept and also fits the proposed application in imaging genetics. The constraint can be extended to both modalities if necessary.

With the distance measure incorporated as an additional metric, the estimation of the unmixing matrix $\mathbf{W_2}$ turns into a multi-objective optimization problem. It is well understood that objective functions of individual objectives can be linearly weighted and combined into a single aggregate objective function (F_2), resulting in a less-complex problem. And by choosing different values for the weighting parameter λ, we can explore different points of the Pareto front (Klamroth and Tind, 2007), which justifies a strategy of choosing the weight via simulation. In addition, we have adopted several strategies to avoid overfitting. First, the constrained component (i.e., $\mathbf{S_{2k}}$ in F_2) is selected dynamically based on the data. Specifically, in each iteration, we examine the distances between the reference and all the components, and then select only the closest component to be constrained. Second, to avoid over-emphasizing the distance metric, we adaptively adjust the constraint weight λ. Starting with a heuristic weight, we monitor the overall independence ($log|det(\mathbf{W_2})|$) and the distance measure after each iteration, then adjust λ accordingly to ensure the balance between the two objectives in the objective function.

The three objective functions (F_1, F_2 and F_3) are optimized using gradient maximization. Specifically, for F_1 and F_2, $\mathbf{W_1}$ and $\mathbf{W_2}$ are updated by the natural gradient learning rule (Amari, 1998), and for F_3, $\mathbf{A_1}$ and $\mathbf{A_2}$ are updated by the steepest descent learning rule (Liu et al., 2009), as shown in Equations (4.4)-(4.6). α_1, α_2, α_{c1} and α_{c2} denote the leaning rates.

$$\Delta W_1 = \alpha_1 \cdot [I + (1 - 2Y_1)U_1^T] \times W_1 \tag{4.4}$$

$$\Delta W_2 = \alpha_2 \cdot \lambda \cdot [I + (1 - 2Y_2)U_2^T] \times W_2$$

$$\Delta W_{2k} = -\alpha_2 \cdot (1-\lambda) \cdot 2\left[\left(\left|W_{2k}\tilde{X}_2\right| - \tilde{r}\right) \times \left(C \cdot \tilde{X}_2^T\right) \times W_2^T W_2\right]$$

$$C \;=\; \underbrace{\left[sign\left([W_{2k}\tilde{X}_2]^T\right), \cdots\cdots, sign\left([W_{2k}\tilde{X}_2]^T\right)\right]}_{L\ columns} \tag{4.5}$$

$$\Delta A_{1i} = \alpha_{c1} \cdot \frac{2Corr\left(A_{1i},\, A_{2j}\right)}{Std(A_{1i})Std\left(A_{2j}\right)} \cdot \left\{\left(A_{2j} - \overline{A_{2j}}\right) + \frac{Cov\left(A_{1i},\, A_{2j}\right)\left(\overline{A_{1i}} - A_{1i}\right)}{Var(A_{1i})}\right\}$$

$$\Delta A_{2j} = \alpha_{c2} \cdot \frac{2Corr\left(A_{2j},\, A_{1i}\right)}{Std\left(A_{2j}\right)Std(A_{1i})} \cdot \left\{\left(A_{1i} - \overline{A_{1i}}\right) + \frac{Cov\left(A_{1i},\, A_{2j}\right)\left(\overline{A_{2j}} - A_{2j}\right)}{Var\left(A_{2j}\right)}\right\} \tag{4.6}$$

4. 3 Simulation

The proposed pICA-R approach was evaluated using simulated functional MRI (fMRI) and SNP data for its capability to extract factors of interest, particularly in the genetic modality. The fMRI data consisted of 200 samples (i.e., subjects) and 10K voxels. Eight non-overlapping brain networks were simulated using the SimTB toolbox ((Erhardt et al., 2011), http://mialab.mrn.org/software). The SNP data were simulated to investigate the performances of pICA-R when components accounted for different amounts of variance in the data, which was achieved through adjusting sample-to-SNP ratios, causal loci ratios, and effect sizes of causal loci. The sample-to-SNP ratio compared the sample size (or number of subjects) with the total number of SNP loci (or SNP dimensionality); the causal loci ratio compared the number of causal loci with the SNP dimensionality; the effect size of causal loci was measured by percentage of variance explained in disease status. Specifically, the SNP data consisted of 200 simulated samples (subjects), each with equal SNP dimensionality, which ranged from 10K to 500K. Eight non-overlapping SNP components were simulated using PLink

(Purcell et al., 2007a), each involving 150 causal loci associated with a randomly generated case-control pattern. The resulting sample-to-SNP ratio ranged from 0.02 (200/10K) to 4.00×10^{-4} (200/500K), and the causal loci ratio ranged from 0.015 (150/10K) to 3.00×10^{-4} (150/500K). The effect size of individual causal loci ranged from 0.003 to 0.21. None of the SNP components shared common causal loci. No high linkage disequilibrium (LD) was observed among causal loci (maximum correlation < 0.39). We further designed a mixing matrix for the fMRI data where randomly selected columns were correlated to particular case-control patterns of the SNP components. The simulated brain networks were then combined into one fMRI observation matrix through this mixing matrix. Random Gaussian noise was superimposed afterwards. We did not adjust the number of components in the simulations as the ability to recover the independent hidden factors is not significantly affected by how many components are embedded, provided that the number of components can be correctly approximated. We used second-level (subject × feature) fMRI data in this simulation, however we would expect comparable performances when pICA-R is applied to structural grey matter images, given that both are feature-based maps and structure-function associations have been observed at the feature level in an ICA framework (Calhoun et al., 2006; Segall et al., 2012).

We then applied pICA-R to the simulated datasets and compared its performance with those of ICA (regular infomax), ICA with reference (ICA-R) (Lin et al., 2010) and pICA. Default settings were used for infomax, ICA-R and pICA. Since infomax, pICA and pICA-R require selection of the component number, we set this to 8, the true component number for the simulated data, for the fMRI modality in all tests. For the SNP modality, due to different data properties, the true component number may not yield reliable results (Chen et al., 2012b). Therefore, in the tests with infomax and pICA, we examined component numbers ranging from 5 to 50 (in steps of 5), and selected the one yielding optimal results. The number of components was selected to be 50 in all pICA-R tests, given our observation that the proposed pICA-R tends to be robust to over-estimation.

The performance was evaluated based on accuracies of the genetic components and loadings, as well as the inter-modality connections. The SNP component accuracy was assessed by a sensitivity measure, the ratio of correctly identified causal loci (among the top 150 loci) to the built-in true causal loci. The genetic loading accuracy was reported as the absolute value of the correlation between the simulated case-control pattern and the extracted loadings. We also calculated the correlations between loadings of the two components (SNP and fMRI) that most resembled the ground truth of the two modalities, respectively, to assess the accuracy of the inter-modality connections.

Particularly, for the two semi-blind methods (pICA-R and ICA-R), we investigated how their performances would be affected by the reference accuracies (ratio of true causal loci in the reference, as illustrated in Figure 4.2). Previous work indicated that a 20-loci reference of accuracy 1 was required for ICA-R to reliably extract factors of interest when the sample-to-SNP ratio was 0.02 (Liu et al., 2012). Guided by this, we first tested a reference of accuracy 1, spanning 20 randomly selected true causal loci. We then tested a 40-loci reference of accuracy 0.5, primarily to investigate how the performances would be affected by adding random loci. Then accuracies were adjusted from 0.1 to 0.5 for the 40-loci references to investigate the influence. The performance was evaluated in terms of sensitivity (as described above) and reference-imposed false discovery rate

(FDR), which was to assess the overfitting by evaluating how many random referential loci were falsely elevated as causal.

4. 4 Results

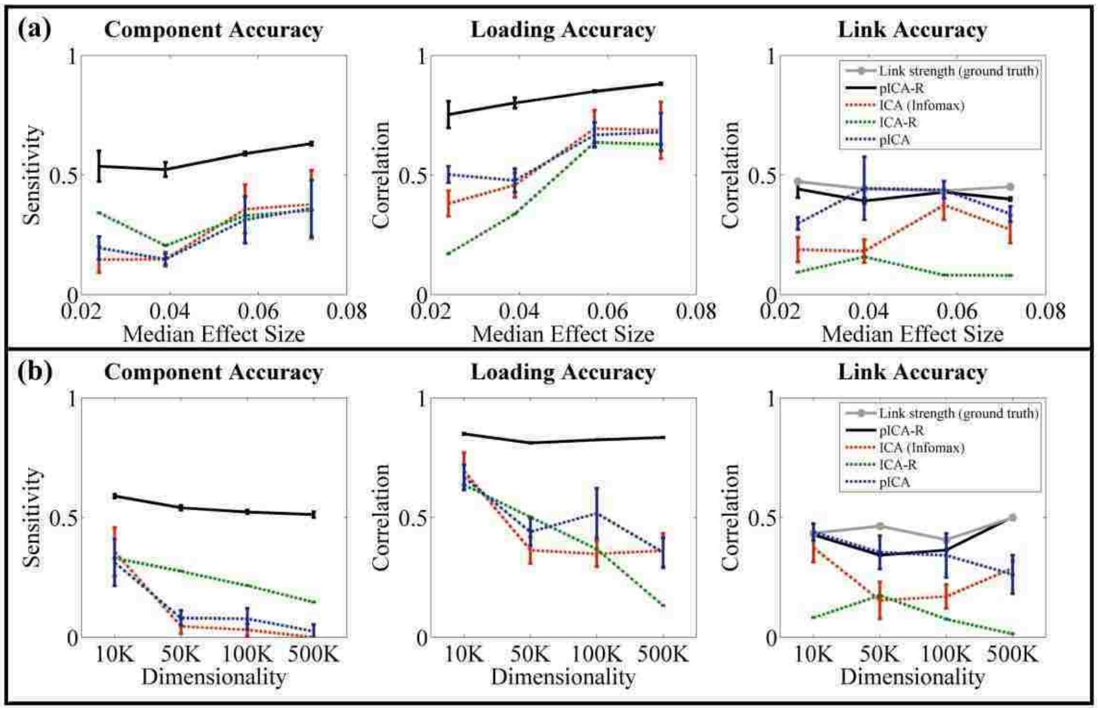

Figure 4.3: Performance comparisons among pICA-R, ICA (infomax), ICA-R and pICA: (a) on simulated datasets with different effect sizes when the sample-to-SNP ratio was controlled at 0.02 and causal loci ratio at 0.015; (b) on simulated datasets with SNP dimensionality ranging from 10K to 500K, resulting in sample-to-SNP ratios ranging from 0.02 to 4×10^{-4} and causal loci ratios from 0.015 to 3×10^{-4}, the median effect sizes were 0.057, 0.055, 0.050 and 0.050 respectively. For pICA-R and ICA-R, results were obtained with a 20-loci reference of accuracy 1. The error bars reflect mean ± SD based on 100 runs.

As expected, fMRI components were accurately identified (component and loading accuracies higher than 0.9) in all tests, given that each component carried a considerable amount of variance in the data. Regarding the SNP modality, with a 20-loci reference of accuracy 1, pICA-R exhibited consistently better performance than the other algorithms in identifying SNP components with different levels of sample-to-SNP ratio, causal loci ratio and effect size. Figure 4.3a and 4.3b summarize the simulation results, where the error bar reflects mean ± SD based on 100 runs. It can be seen that accuracies of SNP components, associated loadings and connections between SNP and fMRI measured by sensitivity or correlation were all improved compared with infomax, ICA-R and pICA. Also it is noted that pICA-R was able to identify the component with a sensitivity above 0.5 given a median effect size as low as 0.024 while the sample-to-SNP ratio was controlled at 0.02 and the causal loci ratio at 0.015. While the median effect size was controlled around 0.05, pICA-R in general exhibited robust performances within the tested ranges of sample-to-SNP ratio and causal loci ratio. We also conducted a simulation at the low sample-to-SNP ratio (200/500K) with an increased causal loci ratio (1000/500K), a scenario similar to real data SZ application, and found that pICA-R exhibited a comparable sensitivity (0.53) using a 20-loci reference of accuracy 1 (not shown). Therefore, we assume that a reference spanning at least 20 true causal loci is suitable for the real data application provided that the causal loci ratio is above 3.00×10^{-4}.

The reference accuracy is crucial for identifying the correct component, as illustrated in Figure 4.4. As expected, pICA-R showed increased sensitivities with references of higher accuracies. It is also noted that a 40-loci reference of accuracy 0.5 yielded a sensitivity around 0.5, comparable to that obtained with a 20-loci reference of accuracy 1. Most importantly, the results indicated that when the sample-to-SNP ratio was lower than 0.004 (200/50K) and the causal loci ratio lower than 0.003 (150/50K), pICA-R started to benefit in sensitivity compared to ICA and pICA with a reference accuracy as low as 0.2. In contrast to sensitivity, the performance in reference-imposed FDR was less affected by the reference accuracy and remained below 0.05. Overall, pICA-R exhibited improvements in both sensitivity and reference-imposed FDR compared to ICA-R.

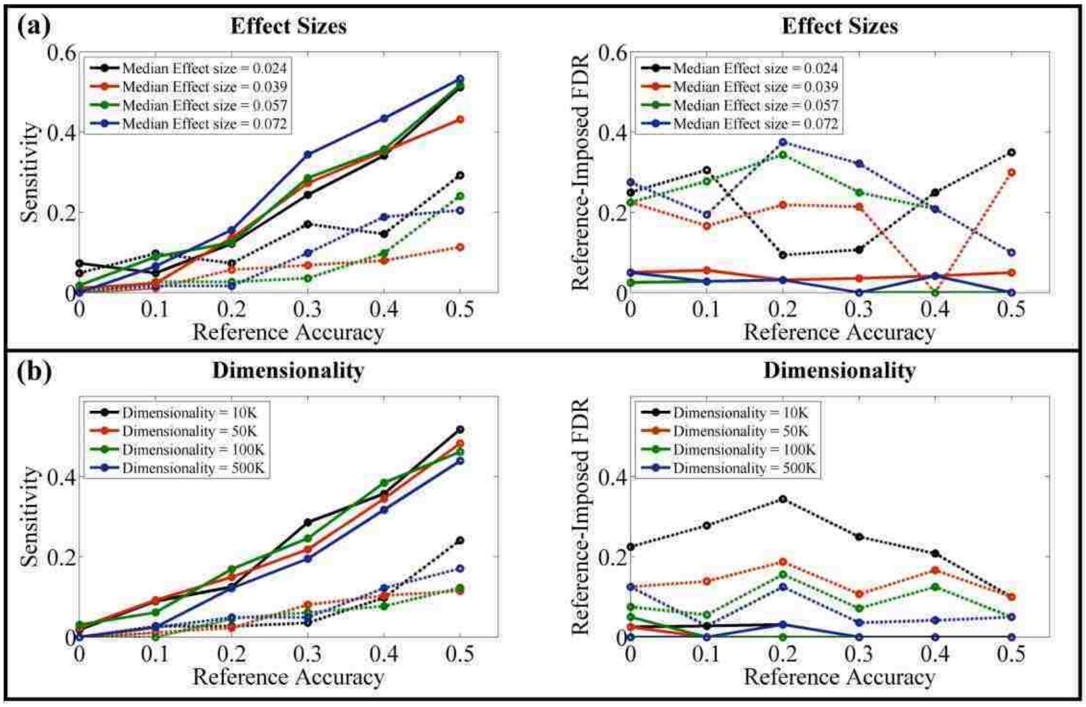

Figure 4.4: Performance comparisons between pICA-R and ICA-R, with 40-loci references of different accuracies: (a) on simulated datasets with different effect sizes when the sample-to-SNP ratio was controlled at 0.02 and causal loci ratio at 0.015; (b) on simulated datasets with SNP dimensionality ranging from 10K to 500K, resulting in sample-to-SNP ratios ranging from 0.02 to 4×10^{-4} and causal loci ratios from 0.015 to 3×10^{-4}, the median effect sizes were 0.057, 0.055, 0.050 and 0.050 respectively. The solid and dotted lines reflect results of pICA-R and ICA-R, respectively.

4. 5 Discussion

The simulation results demonstrate that the approach helps capture factors of interest more accurately. As illustrated in Figure 4.3a and 4.3b, pICA-R show consistently better results for component accuracy, component loadings and inter-modality link compared to regular ICA, ICA-R and pICA, and the improvement becomes more pronounced with lower sample-to-SNP ratio and causal loci ratio, or smaller effect size. It can be seen that the proposed approach yields a sensitivity above 0.5 at a low sample-to-SNP ratio of 4.00×10^{-4} (200/500K) and a causal loci ratio of 3.00×10^{-4} (150/500K), while the median effect size is around 0.05. This observation encourages the application of pICA-R to genomic data with comparable sample-to-SNP and causal loci ratios, where a million or so loci may be involved given an increased yet affordable sample size and hundreds of causal loci.

On the other hand, it needs to be emphasized that reference accuracy plays an important role in the performance of pICA-R. As clearly shown in Figure 4.4, when random loci are incorrectly selected as references, pICA-R exhibits reduced sensitivity. However, at relatively low sample-to-SNP ratios (below 200/50K), even with accuracies as low as 0.2, pICA-R still benefits in sensitivities compared to blind ICA and pICA, indicating a big tolerance of false inputs. Meanwhile, the reference-imposed FDR remains below 0.05, and decreases to 0 with accuracies greater than 0.3. This effective control on reference-imposed FDR is believed to result from a well maintained balance between independence and closeness metric such that the latter never dominates to excessively elevate the referential random loci. Based on the simulation results, a general conclusion can be drawn that a relatively accurate reference is recommended for pICA-R. Compared to a large number of reference loci with low confidence, a small set of reliable reference loci would lead to a better performance. Retrospectively, through investigating the sensitivity and reference-imposed FDR as functions of reference accuracy, we can empirically infer the quality of a reference. The simulation shows that, if more than 10% of the reference SNPs show up in the most significant (i.e., top component weights) findings, the reference accuracy is most likely higher than 0.2 and, the reference benefits the performance. In contrast, a low ratio of reference loci in the most significant findings usually indicates the distance metric being de-emphasized due to low reference accuracy.

In pICA-R, reference SNPs are predicted to contribute simultaneously to only a single component. Therefore, it may be inappropriate to directly combine multiple presumed susceptibility loci identified in univariate analyses, which may then result in a reference containing true SNP hits from multiple components. In this case, the reference is essentially of low accuracy as pICA-R is currently designed to optimize the distance between the reference and one constrained component and the true hits from other components cannot be recognized. Given a low-accuracy reference, minimizing distance will contradict with maximizing independence, which can be captured by the online monitoring of the overall independence. pICA-R will then adaptively adjust the constraint weight to de-emphasize the distance metric to assure the integrity of independent components (as reflected in simulations, Figure 4.4). When the distance metric is significantly de-emphasized, pICA-R effectively converges with results from blind pICA.

While it is true that reference accuracy plays an important role in pICA-R performance, this should not compromise the applicability of the model. First, we implement a binary reference, thus users only need to determine whether the loci are relevant or not to the trait of interest instead of specifying the accurate effect sizes. Second, the model is highly robust to inaccurate reference SNPs. As demonstrated in simulations, pICA-R outperforms blind methods with the accuracy as low as 0.2 when the sample-to-SNP ratio is lower than 4.00×10^{-3} (Figure 4.3 and Figure 4.4). Last but not least, while the choice of reference SNPs is informed by evidence, this is not necessarily limited to association studies. Independent molecular, cellular or system biological knowledge can also guide the selection. Even when informed by association studies, an enormous sample is not a necessity. Replication across studies can help increase confidence in the selection. For example, an association is more likely to be true and poses a good candidate for the reference if consistently observed in several independent studies of small sample sizes. Overall, we believe that the large amount of available data and information learned from previous studies are sufficient to generate testable references for a particular research interest, which can be leveraged by our pICA-R method to increase, broaden or deepen our knowledge at large.

A primary strategy to generate a testable reference is through dissecting the natural LD blocks for individual genes. This is because genome-wide association study (GWAS) is based on the premise that a causal variant is located on a haplotype, and thus a marker allele in LD with the causal variant should show (by proxy) an association with the trait of interest (Stranger et al., 2011). Therefore, SNPs in one LD cluster are more likely to contribute simultaneously to one single component and serve as good candidates for reference. Given the current formulization that one single component is constrained in pICA-R, this primary strategy for reference generation can only test one LD referential set at a time. To improve the robustness, we will extend pICA-R to accommodate multiple referential sets where the interrelationships are unknown. The related contents are presented in the next chapter.

CHAPTER 5 PARALLEL INDEPENDENT COMPONENT ANALYSIS WITH MULTIPLE REFERENCES

5. 1 Introduction

A key factor that affects the performance of pICA-R is the reference accuracy. As shown in the previous chapter, degradation is expected in component, loading and linkage accuracies when the reference accuracy is below 0.2. This raises an issue on how an applicable reference can be effectively derived in real applications where the reference accuracy cannot be reliably estimated. One primary strategy is to derive a referential SNP set based on the natural LD clusters of individual genes. A marker allele in LD with the causal variant should show (by proxy) an association with the trait of interest (Stranger et al., 2011). Compared to random loci, LD loci are more likely to covary with a same underlying pattern (e.g. the trait of interest) and be captured in one component in pICA-R which inherits the linearly additive model from ICA. Based on this assumption, a referential SNP set derived from LD clusters is considered more homogeneous and exhibiting relatively high reference accuracy, which facilitates the pICA-R analysis.

While a single accurate referential SNP set is shown to greatly improve the performance compared to blind approaches, a capability to accommodate multiple referential SNP sets is desired in light of the real applications of pICA-R. For instance, one challenging issue in imaging genetics studies is that, a complex genetic structure is envisaged for plenty of neurological disorders which are polygenic and heterogeneous.

Polygenicity means that a number of genetic variants with possibly small individual effect sizes are involved in the etiology of a disease. Heterogeneity is often reflected in phenotypic or endophenotypic complexity which might indicate contributions from distinct pathological mechanisms. Therefore, computationally assessing multiple referential SNP sets for potential convergence of functional influences on neurobiological traits should help better delineate the genetic architecture underlying the complex disorders.

5. 2 Method

Given the formularization of pICA-R, it is inappropriate to directly combine multiple referential SNP sets, as this imposes an assumption that all the referential SNPs are expected to contribute to the same component, which is not necessarily the case. Instead, the extended approach, parallel ICA with multiple references, is designed to investigate how each of the referential sets is represented in the observed data such that those potentially related will be naturally grasped when they guide the algorithm to constrain the same component.

Recall that in pICA-R, the reference is designed as a binary vector where candidate causal loci are set to “1” and others set to “0”. When multiple referential sets are involved, the reference vector is expanded into a reference matrix. Each row represents a referential set which comprises a group of loci likely related. The interrelationship among different reference vectors is to be investigated. Figure 5.1 and Equations (5.1)-(5.3) illustrate the extended model, where the notations are consistent with those used for pICA-R in the previous chapter. The observed dataset **X** (sample × feature) is decomposed into a linear combination of the underlying independent components, or sources. **S**, **A** and **W** denote the component, mixing and unmixing matrices, respectively. The subscript d runs from 1 to 2, denoting the data modality. The unmixing matrices $\mathbf{W_1}$ and $\mathbf{W_2}$ are iteratively updated to maximize the objective functions F_1, F_2 and F_3. Similar to single-reference pICA-R, F_1 represents the objective function of the regular infomax (Bell and Sejnowski, 1995b) for modality 1, and the inter-modality association function F_3 is designed to maximize the correlations computed over the columns of the loading matrices $\mathbf{A_1}$ and $\mathbf{A_2}$. In contrast, F_2, the objective function for modality 2, is modified so that multiple components might be constrained to closely resemble the reference matrix **r**. Specifically, when each row of **r** represents a referential set, the algorithm sequentially works with each reference vector $\mathbf{r_i}$ to determine the closest component and then apply the constraint, as shown in (5.2). The subscript *i* denotes the row index in matrix **r** and *I* denotes the total number of rows. Figure 5.1illustrates the measurement of closeness. For a particular reference vector $\mathbf{r_i}$, the referential loci are represented by $\mathbf{\tilde{r}_i}$, a subvector of $\mathbf{r_i}$. The algorithm calculates the Euclidean distance between each of the components and the reference for the referential loci ($\left\| \left| \mathbf{\tilde{S}}_{2k_i} \right| - \mathbf{\tilde{r}}_i \right\|_2^2$), based on which the closest component is selected to be constrained, such that the distance is further minimized. $\mathbf{\tilde{S}_{2k_i}}$ represents a subvector of the constrained component $\mathbf{S_{2k_i}}$ (i.e. the k_i^{th} row of $\mathbf{S_2}$); $\mathbf{W_{2k_i}}$ denotes the k_i^{th} row of $\mathbf{W_2}$ and $\mathbf{\tilde{X}_2}$ denotes a submatrix of $\mathbf{X_2}$; $\|\cdot\|_2$ represents the L_2-norm Euclidian distance, and λ is the weighting parameter.

$$X_d = A_d S_d \rightarrow S_d = W_d X_d \,, A_d = W_d^{-1}, \quad d = 1, 2 \qquad (5.1)$$

$$\begin{aligned} &Y_d = \frac{1}{1+e^{-U_d}}, U_d = W_d X_d + W_{d0} \\ &F_1 = max\{H(Y_1)\} = max\{-E[ln\, f_{y_1}(Y_1)]\} \\ &F_2 = max\{\lambda H(Y_2) + (1-\lambda)[-\textstyle\sum_{i=1}^{I} dist^2(\tilde{r}_i, |\tilde{S}_{2k_i}|)]\} \\ &\quad = max\left\{\lambda\left(-E\left[ln\, f_{y_2}(Y_2)\right]\right) + (1-\lambda)\left(-\textstyle\sum_{i=1}^{I}\left\|\left|W_{2k_i}\tilde{X}_2\right| - \tilde{r}_i\right\|_2^2\right)\right\} \end{aligned} \tag{5.2}$$

$$F_3 = max\{\textstyle\sum_{i,j} Corr^2(A_{1i}, A_{2j})\} = max\left\{\sum_{i,j}\frac{Cov^2(A_{1i},A_{2j})}{Var(A_{1i})Var(A_{2j})}\right\} \tag{5.3}$$

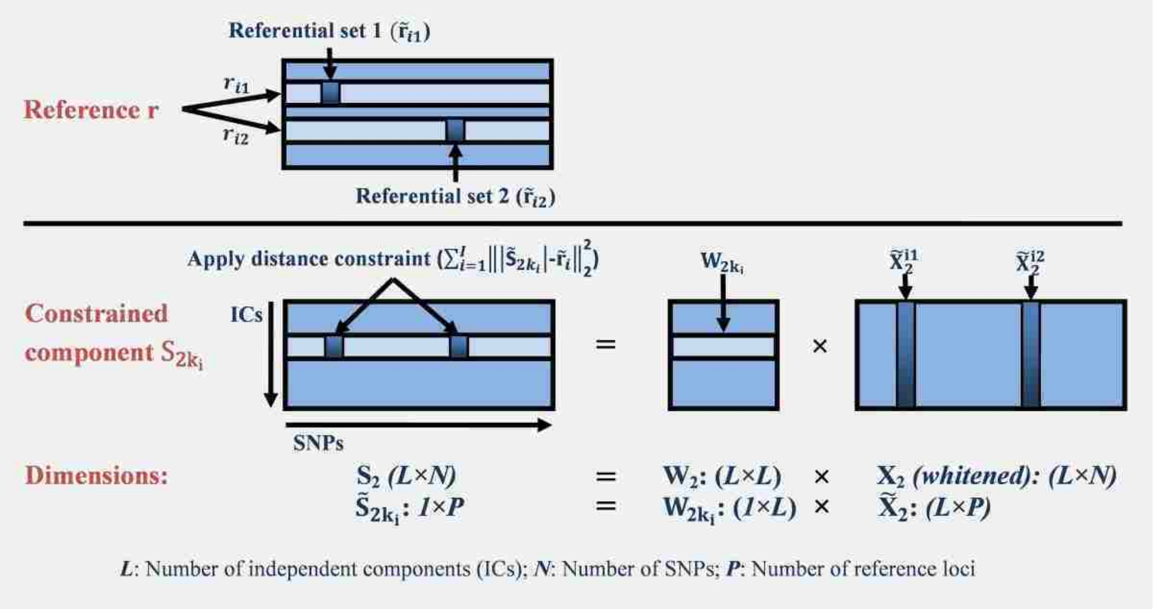
Referential set 1 ($\tilde{r}_{i1}$)
Reference r
r_{i1}
r_{i2}
Referential set 2 ($\tilde{r}_{i2}$)
Apply distance constraint ($\sum_{i=1}^{I}\||\tilde{S}_{2k_i}|-\tilde{r}_i\|_2^2$)
W_{2k_i}
$\tilde{X}_2^{i1}$
$\tilde{X}_2^{i2}$
Constrained component S_{2k_i}
ICs
=
×
SNPs
Dimensions:
S_2 $(L×N)$ = W_2: $(L×L)$ × X_2 (whitened): $(L×N)$
$\tilde{S}_{2k_i}$: $1×P$ = W_{2k_i}: $(1×L)$ × $\tilde{X}_2$: $(L×P)$
L: Number of independent components (ICs); N: Number of SNPs; P: Number of reference loci

Figure 5.1: Parallel ICA with multiple reference sets. Each row (e.g., $\mathbf{r_{i1}}$ and $\mathbf{r_{i2}}$,) represents a set of referential loci residing in one single LD cluster. In this demonstration, $\mathbf{r_{i1}}$ and $\mathbf{r_{i2}}$ constrain the same component ($\mathbf{k_{i1}} = \mathbf{k_{i2}}$) represented by $\mathbf{S_{2k_i}}$.

Again, the three objective functions (F_1, F_2 and F_3) are optimized using gradient maximization. For F_1 and F_3, the updating rules are same as those for single-reference pICA-R, and we refer readers to Equations (4.4) and (4.6). For F_2, two situations will be discussed. (a) When reference vectors $\mathbf{r_{i1}}$ and $\mathbf{r_{i2}}$ select out different constrained components $\mathbf{S_{2k_{i1}}}$ and $\mathbf{S_{2k_{i2}}}$, the corresponding rows of the unmixing matrix $\mathbf{W_{2k_{i1}}}$ and $\mathbf{W_{2k_{i2}}}$will be updated separately, as shown in (5.4). (b) When the two reference vectors select out the same constrained component $\mathbf{S_{2k_i}}$, the corresponding row of the unmixing matrix $\mathbf{W_{2k_i}}$ will be updated sequentially for $\mathbf{r_{i1}}$ and $\mathbf{r_{i2}}$. As shown in (5.5) and Figure 5.1, this is essentially equivalent to $\mathbf{W_{2k_i}}$ being updated for a single-reference vector $\mathbf{r_i}$ harboring both $\mathbf{r_{i1}}$ and $\mathbf{r_{i2}}$, and is expected to yield comparable performances with single-reference pICA-R. Such, pICA-R is extended to accommodate multiple referential sets without making any assumption on whether or not the loci highlighted in different referential sets are independent from each other, and the constraint of each referential set is applied based on how it is represented in the data.

$$\Delta W_2 = \alpha_2 \cdot \lambda \cdot [I + (1 - 2Y_2)U_2^T] \times W_2$$

$$\Delta W_{2k_{i1}} = -\alpha_2 \cdot (1 - \lambda) \cdot 2\left[\left(\left|W_{2k_{i1}}\tilde{X}_2^{i1}\right| - \tilde{r}_i\right) \times \left(C_{k_{i1}} \cdot \tilde{X}_2^{i1^T}\right) \times W_2^T W_2\right]$$

$$C_{k_{i1}} = \underbrace{\left[sign\left(\left[W_{2k_{i1}}\tilde{X}_2^{i1}\right]^T\right), \cdots\cdots, sign\left(\left[W_{2k_{i1}}\tilde{X}_2^{i1}\right]^T\right)\right]}_{L\ columns}$$

$$\Delta W_{2k_{i2}} = -\alpha_2 \cdot (1 - \lambda) \cdot 2\left[\left(\left|W_{2k_{i2}}\tilde{X}_2^{i2}\right| - \tilde{r}_i\right) \times \left(C_{k_{i2}} \cdot \tilde{X}_2^{i2^T}\right) \times W_2^T W_2\right]$$

$$C_{k_{i2}} = \underbrace{\left[sign\left(\left[W_{2k_{i2}}\tilde{X}_2^{i2}\right]^T\right), \cdots\cdots, sign\left(\left[W_{2k_{i2}}\tilde{X}_2^{i2}\right]^T\right)\right]}_{L\ columns}$$

$$(\ k_{i1} \neq k_{i2}) \tag{5.4}$$

$$\Delta W_{2k_i} = -\alpha_2 \cdot (1 - \lambda) \cdot 2\left\{\left(\left|W_{2k_i}\tilde{X}_2\right| - \tilde{r}_i\right) \times \left(C_{k_i} \cdot \tilde{X}_2^T\right) \times W_2^T W_2\right\}$$

$$= -\alpha_2{}'\left\{\left(\left|W_{2k_i}\left[\tilde{X}_2^{i1}, \tilde{X}_2^{i2}\right]\right| - [\tilde{r}_{i1}, \tilde{r}_{i2}]\right) \times \left(\begin{bmatrix} C_{k_{i1}} \\ C_{k_{i2}} \end{bmatrix} \cdot \begin{bmatrix} \tilde{X}_2^{i1^T} \\ \tilde{X}_2^{i2^T} \end{bmatrix}\right) \times W_2^T W_2\right\}$$

$$= -\alpha_2{}'\left\{\left[\left|W_{2k_i}\tilde{X}_2^{i1}\right| - \tilde{r}_{i1}, \left|W_{2k_i}\tilde{X}_2^{i2}\right| - \tilde{r}_{i2}\right] \times \left(\begin{bmatrix} C_{k_{i1}} \\ C_{k_{i2}} \end{bmatrix} \cdot \begin{bmatrix} \tilde{X}_2^{i1^T} \\ \tilde{X}_2^{i2^T} \end{bmatrix}\right) \times W_2^T W_2\right\}$$

$$= -\alpha_2{}' \left\{ \left(\left| W_{2k_i} \tilde{X}_2^{i1} \right| - \tilde{r}_{i1} \right) \times \left(C_{k_{i1}} \cdot \tilde{X}_2^{i1^T} \right) + \left(\left| W_{2k_i} \tilde{X}_2^{i2} \right| - \tilde{r}_{i2} \right) \times \left(C_{k_{i2}} \cdot \tilde{X}_2^{i2^T} \right) \right\} \times$$

$$W_2^T W_2$$

$$= \Delta W_{2k_i}|_{r_{i1}} + \Delta W_{2k_i}|_{r_{i2}}$$

$$C_{k_i} = \underbrace{\left[sign\left(\left[W_{2k_i} \tilde{X}_2 \right]^T \right), \cdots\cdots, sign\left(\left[W_{2k_i} \tilde{X}_2 \right]^T \right) \right]}_{L\ columns} \tag{5.5}$$

5. 3 Simulation

Similar to single-reference pICA-R, the extended multi-reference approach was firstly evaluated with simulated fMRI and SNP data for its effectiveness. As we already compared pICA-R with other competing approaches (see chapter 4), the key point for the extended approach then lied in whether it is able to correctly capture referential sets contributing to the same component. Therefore, the simulation was focused on investigating the true and false positive rates with respect to this issue. Again the tolerance of reference accuracy was comprehensively assessed. For proof-of-concept, we conducted the simulations with two referential sets imposed. However, the algorithm is able to deal with more.

The simulated data consisted of 200 samples (i.e., subjects). Eight independent vectors were randomly generated from normal distributions to form a mixing matrix for the fMRI data. Subsequently, eight case-control patterns were correspondingly generated through linear transformations of the fMRI mixing vectors with random Gaussian noises superimposed, such that associations were built between fMRI and diagnosis. The case-control patterns would then be used in PLink (Purcell et al., 2007b) for simulating SNP data.

The simulated fMRI data had a feature dimension of 40K voxels. Eight non-overlapping brain networks were simulated using the SimTB toolbox ((Erhardt et al., 2011), http://mialab.mrn.org/software) to serve as the fMRI components. The fMRI observation matrix was then obtained as the product of the mixing and component matrices with random Gaussian noises superimposed onto each sample. Similar to single-reference simulations, the SNP data were simulated via PLink. The performances of the approach was investigated with components accounting for different amounts of variance in the data, which was achieved through adjusting the SNP dimensionality and causal loci effect size. Eight non-overlapping SNP components were simulated using PLink (Purcell et al., 2007a), each involving 150 causal loci. The SNP dimensionality ranged from 50K to 500K, resulting in the sample-to-SNP ratio from 0.004 (200/50K) to 0.0004 (200/500K), and the causal loci ratio from 0.003 (150/50K) to 0.0003 (150/500K). These eight SNP components consisted of 4 pairs, where each pair comprised two components whose causal loci were linked to the same case-control pattern. Note that PLink does not generate SNPs in LD. Thus through linking two components to the same case-control pattern, we obtained two groups of SNPs associated with the same diagnosis and fMRI loadings. A two-sample t-test showed that the correlations among SNPs linked to the same diagnosis were not significantly different from the correlations among randomly generated SNPs ($p = 0.35$). Evaluated with explained variances, the effect sizes of individual causal loci ranged from 0.0037 to 0.1926.

To evaluate the performance of the extended approach, we first investigated out of 100 runs, what would be the ratio for the algorithm to accurately detect the referential sets contributing to the same SNP component and neurobiological trait, denoted as linked reference matching ratio in the following text. Specifically, a reference matrix was generated, with each vector harboring a referential set derived from one of the two groups of causal loci that were linked to the same case-control pattern and fMRI loading. The evaluation started with two accurate referential sets, each spanning 20 true causal loci. Then referential sets spanning 40 loci of accuracies ranging from 0 to 0.5 were tested to investigate the performance boundary. Equivalently, we also evaluated the ratio for the algorithm to falsely constrain the same SNP component for two isolated referential sets, denoted as isolated reference mismatching ratio. For this purpose, the reference matrix was generated to harbor two referential sets derived from two groups of causal loci that were linked to distinct case-control patterns. Again, the same range of reference accuracies was evaluated, as for the linked reference matching ratio.

The model requires selection of the component number. In the simulations, the fMRI component number was set to 8, the true component number for the simulated data, for all tests. For the SNP modality, the number of components was in general selected to be 50 given our observation that the semi-blind pICA-R model tends to be robust to over-estimation. However, when isolated reference mismatching ratio was evaluated with two isolated referential sets, the SNP component number was set to 100. The reason is the following. When ICA was applied to the genotype data, referenced PCA was commonly employed for data reduction instead of regular PCA, where a case-control pattern was employed such that top phenotype-related PCs, instead of those explaining the largest amounts of variance, were selected. Obviously, referenced PCA is not robust to multiple phenotypic references, for which the top related PCs might be different. To address this issue, in our simulation we ranked together the projected eigenvalues when multiple phenotypic references were used and then selected out top related PCs. This strategy favors an objective selection based on how phenotypes are represented in the data. It should be noted that this is majorly for simulations to test the isolated reference mismatching ratio and is not expected to be a common practice in real applications, where isolated referential sets potentially contributing to different phenotypes should be tested in two runs.

Besides the linked reference matching ratio and isolated reference mismatching ratio, accuracies of genetic components, loadings, and inter-modality connections, as well as the reference-imposed FDR were also evaluated for the extended approach. The SNP component accuracy was assessed with the ratio of correctly identified causal loci to the built-in true causal loci. The genetic loading accuracy was reported as the absolute value of the correlation between the simulated case-control pattern and the extracted loadings. The correlation between loadings of the two components (SNP and fMRI) most resembling the ground truth of the two modalities, respectively, was calculated and compared with the built-in correlation to reflect the linkage accuracy. In addition, when two isolated referential sets were tested, the performance of a combined run was compared with that of two separated runs. The former conducted the analysis with the extended approach where two isolated referential sets were assessed simultaneously, and the latter conducted two separate analyses with single-reference pICA-R. Due to the computation burden, we did this combined versus separate comparison only for one dataset with a SNP dimensionality of 50K and median effect size of 0.059.

5. 4 Results

Overall, the extended approach successfully captured the reference structure when two referential sets were assessed simultaneously.

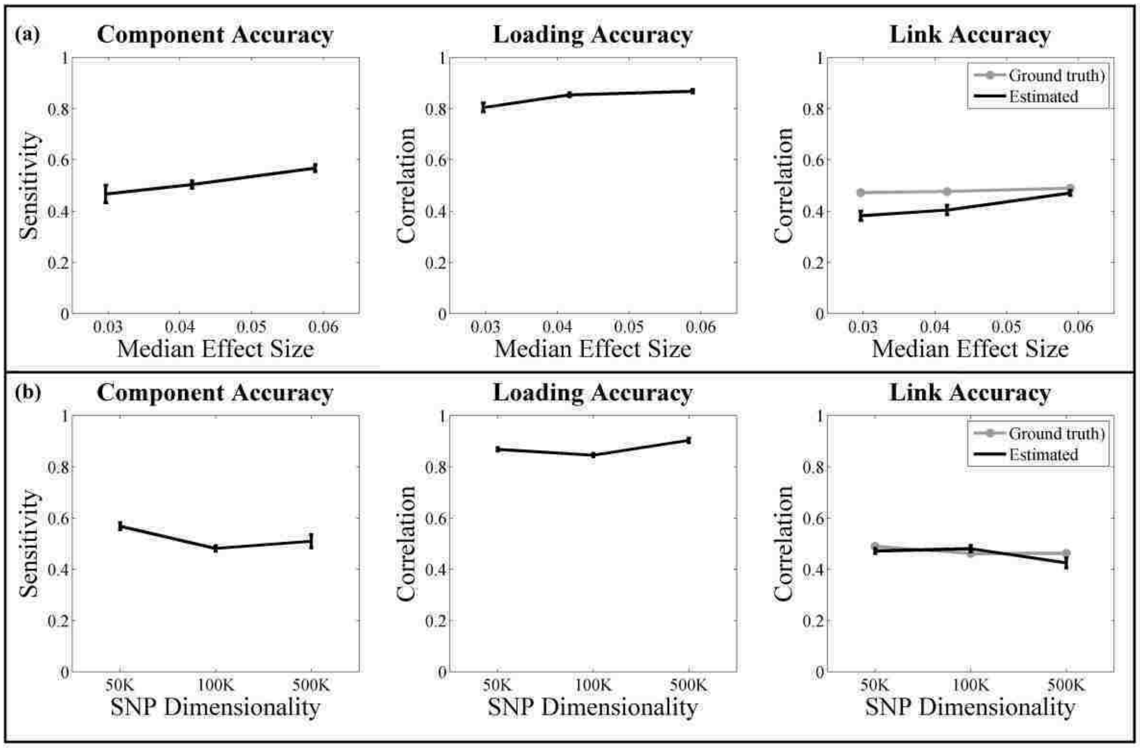

Figure 5.2: Performance of parallel ICA with multiple references: (a) on simulated datasets with different effect sizes when the sample-to-SNP ratio was controlled at 0.004 and causal loci ratio at 0.003; (b) on simulated datasets with SNP dimensionality ranging from 50K to 500K, resulting in sample-to-SNP ratios from 0.004 to 0.0004 and causal loci ratios from 0.003 to 0.0003, the median effect sizes were 0.059, 0.057 and 0.060, respectively. Results were obtained with 20-loci references of accuracy 1. The error bars reflect mean ± SD based on 100 runs.

Given referential sets of 20 true causal loci (accuracy = 1), the linked reference matching ratio was 1 based on 100 runs, regardless of the causal loci effect size or SNP dimensionality. As demonstrated in (5.5), this is equivalent to one component being constrained by a single reference vector harboring both referential sets. Figure 5.2 shows the corresponding performances of component, loading and link accuracies, measured with sensitivity and correlations, respectively. In agreement with single-reference pICA-R, the extended approach showed robust performances under different scenarios. Given a sample-to-SNP ratio of 0.004 (200/50K), the component accuracy remained close to 0.5 when the median causal loci effect size was decreased to 0.03. Meanwhile, consistent performances were observed when the sample-to-SNP ratio decreased from 0.004 to 0.0004 (200/500K).

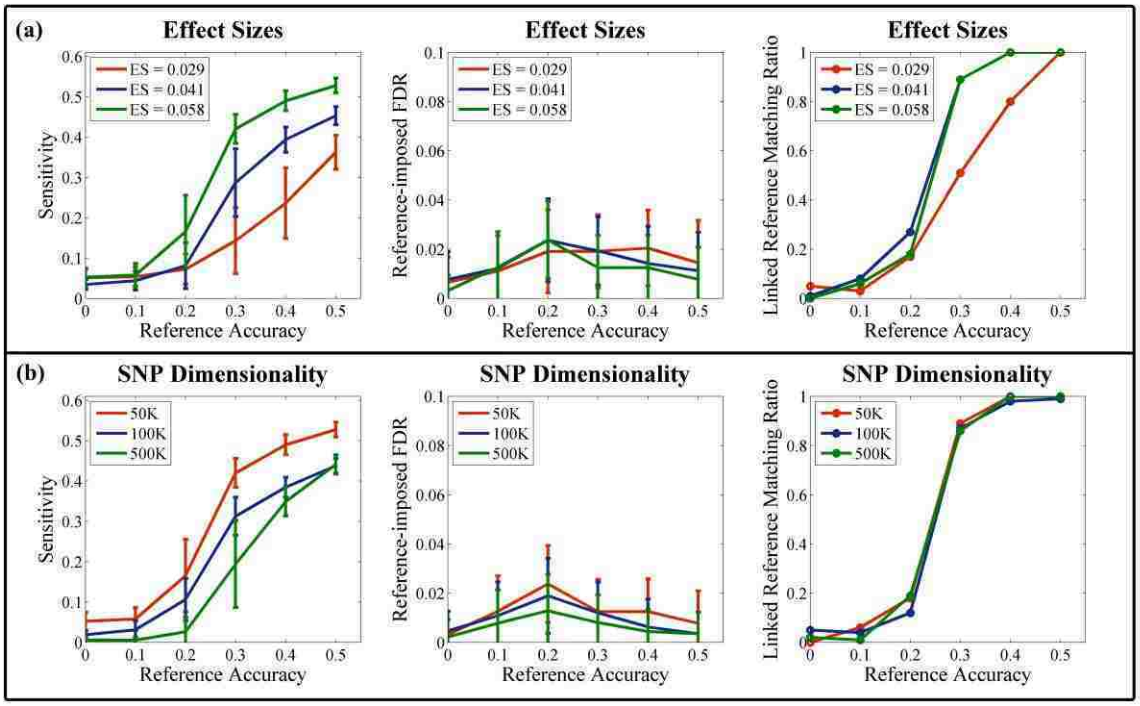

Figure 5.3: Performance comparisons with 40-loci references of different accuracies: (a) on simulated datasets with different effect sizes when the sample-to-SNP ratio was controlled at 0.004 and causal loci ratio at 0.003; (b) on simulated datasets with SNP dimensionality ranging from 50K to 500K, resulting in sample-to-SNP ratios ranging from 0.004 to 0.0004 and causal loci ratios from 0.003 to 0.0003, the median effect sizes were 0.059, 0.057, and 0.060, respectively. The error bars reflect mean ± SD based on 100 runs.

As expected, linked reference matching ratio was significantly affected by reference accuracies. Given 40-loci referential sets, a reference accuracy above 0.2 was required for the algorithm to effectively capture the reference structure, as shown in Figure 5.3. When the reference accuracy was below 0.2, the linked reference matching ratio was lower than 0.2, indicating that for more than 80 out of 100 runs, the algorithm did not constrain the same SNP component for the two referential sets. Meanwhile, a dramatic improvement was in general observed at the reference accuracy of 0.3 for the linked reference matching ratio, which reached 0.9. When the reference accuracy was further increased to 0.5, a linked reference matching ratio of 1 was obtained regardless of the causal loci effect size or SNP dimensionality, consistent with the previous observation for referential sets spanning 20 true causal loci. With respect to component accuracies, the performances of the extend approach were comparable to those of single-reference pICA-R when the reference structure was correctly identified given relatively higher accuracies (> 0.3). On the other hand, larger performance deviations were observed for lower reference accuracies, which was possibly due to component splitting when the algorithm could not identify the reference structure such that two distinct SNP components were constrained for the two referential sets. Again the reference-imposed component FDR was not affected by the reference accuracy, remaining below 0.05 for all the tested scenarios.

When two isolated referential sets were assessed simultaneously, the isolated reference mismatching ratio was below 0.05 for all the tested reference accuracies, as shown in Figure 5.4. This suggests that there is a low chance for the algorithm to mistakenly constrain the same component when the two referential sets are essentially not related. It was also noted that, when two isolated referential sets were assessed simultaneously, it yielded (a) comparable component accuracies (Figure 5.4) to those obtained from the single-reference analysis for referential set A; (b) degraded accuracies for referential set B. Meanwhile, no significant difference in reference-imposed component FDR was observed between combined and separate runs.

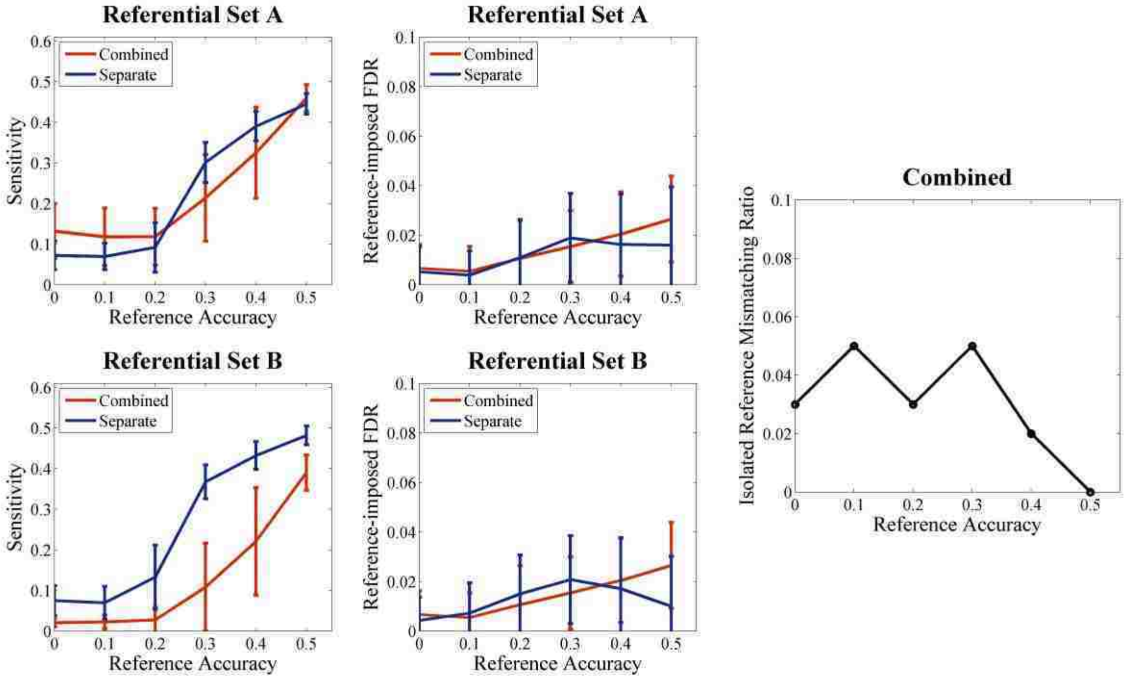

Figure 5.4: Performance comparisons with 40-loci references of different accuracies when two isolated reference sets were assessed in a combined run (parallel ICA with multiple references) or two separate runs (single-reference pICA-R), respectively. The tested dataset has a sample-to-SNP ratio of 0.004 and a median effect size of 0.059. The error bars reflect mean ± SD based on 100 runs. The isolated reference mismatching ratio was reported for the combined runs.

5. 5 Discussion

The simulation results demonstrate that the extended approach helps capture the embedded reference structure in a nonparametric manner. As illustrated in Figure 5.2, when provided with accurate referential sets associated with the same trait of interest, the algorithm correctly recognizes that they contribute to the same component and applies the constraints. The resulting component, loading and link accuracies are comparable to those previously observed in single-reference analyses. In particular, reliable detection of underlying reference structure is observed for a low sample-to-SNP ratio (200/500K) given a median effect size around 0.06, confirming the feasibility of applying the extended approach to real imaging genomics studies to analyze a million or so loci provided that hundreds of subjects are available.

Meanwhile, it is noticed that reference accuracy plays an important role in the performance of linked reference matching ratio for parallel ICA with multiple references. As shown in Figure 5.3, for reference accuracies below 0.2, the linked reference matching ratio is lower than 0.2. This is not surprising, as for the proposed data-driven approach, when random loci are incorrectly selected to be referential, the distance between the reference vector and the component is smeared and no longer distinguishes itself from those randomly observed. On the other hand, it's encouraging to observe a dramatic improvement of linked reference matching ratio to around 0.9 at the reference accuracy of 0.3, which is likely to achieve when using the strategy of deriving a single referential set based on LD loci.

Besides reference accuracy, the performance of linked reference matching ratio is also affected by causal loci effect size. Obviously, degradation is observed for data with lower causal loci effect sizes (0.029), for which the linked reference matching ratio only reaches 0.5 at a reference accuracy of 0.3, as shown in Figure 5.3. Interestingly, the performance is less vulnerable to the increase of SNP dimensionality. This might be attributable to the design of the model. Recall that the Euclidean distance is calculated between the component and the reference vector specifically for those referential loci. Thus, an increased SNP dimensionality simply results in an increased number of to-be-investigated loci, which might not significantly affect the estimated distance metric. Instead, decrease in effect sizes is expected to increase the distance between the reference vector and the true component, such that other components might by chance be closer to the reference vector and selected for constraint, resulting in in a low linked reference matching ratio.

While linked reference matching ratio is more sensitive to reference accuracy and data properties, isolated reference mismatching ratio is less affected and reliably below 0.05 when two isolated referential sets are assessed simultaneously, as shown in Figure 5.4. Note that when the reference accuracy is 0, the two isolated referential sets essentially consist of random loci, which is equivalent to the situation when linked reference matching ratio is evaluated for two referential sets of accuracy 0. In both cases, the chance is below 5% for the algorithm to constrain the same component for the two tested referential sets, as consistently observed in Figure 5.3 and 5.4. This also applies to the situations when the reference accuracy is increased for the two isolated referential sets, as the added true causal loci of one set are essentially recognized as random loci by the other set when they are indeed not related. Therefore, the linked reference matching ratio is expected to remain below 0.05, regardless of reference accuracy.

When two referential sets associated with different phenotypes are assessed in one run, the resulting performances might be affected by the data reduction if referenced PCA is employed. In our simulation, we adopted an objective selection such that more related information would be included for the phenotype better represented in the observed data.

As shown in Figure 5.4, for referential set A, comparable component accuracies are observed between combined and separate runs, while degradation is observed for referential set B. It's noted that the median causal loci effect sizes for these two components are 0.059 and 0.057, respectively. Thus, more information related to component A is included through referenced PCA, which results in higher sensitivity in the combined runs.

Overall, parallel ICA with multiple references is able to assess multiple referential sets simultaneously while the interrelationships are not known. Compared to single-reference pICA-R, the extended approach is more flexible in dynamically identifying the constrained component for individual referential set and allows some extent of heterogeneity in the reference. Simulation results demonstrate high linked reference matching ratio and low isolated reference mismatching ratio, confirming the validity of Euclidean distance as a metric for the assessment of reference structure. Meanwhile, some cautions need to be exercised when conducting an analysis. First, an accurate reference is favored. Compared to single-reference pICA-R, the extended approach is slightly more sensitive to reference accuracy. This is because when two referential sets of low accuracies (e.g. accuracy <= 0.2) are employed, component splitting is likely to happen if they fail to constrain the same component while some of the referential loci are indeed related. A practical strategy is to derive individual referential sets based on LD blocks of genes. In general SNPs in LD are more likely associated with the same trait of interest, and hence contribute to the same component. Second, when referenced PCA is employed for data reduction, it is not recommended to test a variety of hypotheses in one single run. As different behavioral manifestations may involve different biological mechanisms, when one of the phenotypes is selected as a PCA reference, the reduced data better represents mechanisms underlying this specific phenotype than those underlying other phenotypes. Consequently, there is little chance to identify components accurately representing those unattended mechanisms even if the references are accurate. Instead, the extended approach is more suitable for assessing the architecture of genes which, although previously implicated in the same biological mechanism, still await investigations on their homogeneous or heterogeneous functional influences on neurobiological conditions.

CHAPTER 6 EXPLORATION OF SCANNING EFFECTS IN MULTI-SITE STRUCTURAL MRI STUDIES

6. 1 Introduction

Structural magnetic resonance imaging (sMRI) is increasingly used to study morphology of living brain given its non-invasive nature. Based on high resolution T_1-weighted images, measures can be derived to quantify brain structure for further analysis such as assessment of neurological diseases (Fornito et al., 2009; Glahn et al., 2008). A variety of computational methods have been developed to deal with the anatomical complexity, one of those commonly used is voxel-based morphometry (VBM) (Ashburner and Friston, 2000, 2005). VBM involves tissue segmentation, spatial normalization and smoothing procedures, followed by voxelwise univariate statistical tests on feature changes across subjects. This method has been successfully applied to characterize structural abnormalities in a variety of diseases such as schizophrenia (SZ) and Alzheimer's disease (Ferreira et al., 2011; Giuliani et al., 2005; Honea et al., 2005) as well as track the structural changes as a response of environmental factors (Maguire et al., 2000).

Most sMRI studies are performed at a single site; however pooling of multi-site data is becoming more common, especially imaging genetics studies. This is due to a desire for a large sample size to provide sufficient statistical power for the investigation of subgroups, or a rare condition, or a factor of relatively small effect size. In contrast to the desire for data pooling, collaborative sMRI studies face some big challenges, one of which is that inconsistent collection platforms can introduce systematic differences to distort the image information and confound the true effect of interest. In addition, even in single-site studies a scanner will likely undergo hardware exchanges or software upgrades over time, making it difficult to keep the status consistent over the period of a longer study. Previous work has revealed scanning effects resulting from a number of factors, including static magnetic field inhomogeneity, imaging gradient nonlinearity and difference in subject positioning (Focke et al., 2011; Jovicich et al., 2006; Littmann et al., 2006; Vovk et al., 2007). On the other hand, it is also noted that these scanning effects may be orthogonal to and not necessarily interfere with the true effects of interest (Segall et al., 2009; Stonnington et al., 2008), or that a specific statistical modeling will ameliorate the scanning effects (Fennema-Notestine et al., 2007), which encourages more efforts towards collaborative studies.

In the present study, we aim to explore how various scanning parameters influence the sMRI image pattern and whether a correction is applicable. Through studying gray matter concentration (GMC) images collected from a large cohort of 1,460 healthy subjects, we expect to locate pivotal scanning parameters, which may be calibrated to avoid significant systematic differences, or be selectively included as covariates in post-hoc analyses. Specifically, ICA (Amari, 1998; Bell and Sejnowski, 1995a) is applied to decompose the data into a linear combination of underlying sources, which are then investigated for associations with various scanning parameters to assess the influence (also called source-based morphometry (SBM) (Xu et al., 2009)). A correction procedure has also been designed and tested in a second study (110 SZ patients versus 124 healthy controls), which allows evaluating the effectiveness in reducing scanning effects and in particular, refining true effects of interest.

6. 2 Materials and Methods

BIG Data

Subjects: The exploratory study included 1,460 subjects from the ongoing Brain Imaging Genetics (BIG) study, being conducted at the Radboud University Nijmegen (Nijmegen, the Netherlands). The regional medical ethics committee approved the study and all subjects provided written informed consents. The cohort included in this study consisted of 617 males (age: 23.35±4.22) and 843 females (age: 22.69±3.66) whose MRI scans were pooled from various studies conducted since 2003. All subjects are healthy, highly educated adults of Caucasian origin, and free of neurological or psychiatric history according to self-reports.

Neuroimaging: Structural images were acquired at the Donders Centre for Cognitive Neuroimaging (Nijmegen, the Netherlands). Table 6.1 provides a summary of the scanning settings. Subjects were scanned using different scanners, i.e. 1.5T Siemens Avanto and Sonata, as well as 3.0T Siemens Trio and TIM Trio. Transmitting and receiving coils also differed across subjects. A standard coronal T_1-weighted three-dimensional magnetization prepared rapid gradient echo (MP-RAGE) sequence was employed, while some variations were observed in repetition, inversion and echo time, as well as pixel bandwidth and flip angle. The use of parallel imaging (GRAPPA) with an acceleration factor of 2 was also included.

Table 6.1: Summary of BIG scanning parameters.

Scanning parameter	Variations across subjects
StationName	avanto (572), sonata (175), trio (56), triotim (657)
SequenceName	*tfl3d1 (16), *tfl3d1_ns (1101), spc3d1rr282ns (6), tfl3d1 (2), tfl3d1_ns (335)
SliceThickness	0.8 (2), 0.84 (2), 0.87 (1), 0.91 (1), 1 (1453), 1.1 (1)
RepetitionTime	1660 (3), 1960 (16), 2250 (645), 2300 (690), 2730 (100), 3200 (6)
EchoTime	2.02 (3), 2.86 (1), 2.92 (28), 2.94 (1), 2.95 (572), 2.96 (193), 2.99 (15), 3.03 (403), 3.04 (1), 3.08 (1), 3.11 (1), 3.13 (1), 3.55 (2), 3.68 (162), 3.93 (54), 4.01 (6), 4.43 (9), 4.58 (4), 5.59 (3)
InversionTime	1000 (100), 1100 (704), 750 (3), 850 (645), 900 (2), null (6)
NumberOfAverages	1 (1457), 2 (3)
MagneticFieldStrength	1.494 (112), 1.5 (635), 2.89362 (56), 3 (657)
NumberOfPhaseEncodingSteps	176 (3), 196 (6), 253 (5), 255 (658), 256 (786), 320 (2)
PercentPhaseFieldOfView	100 (1441), 68.75 (3), 81.25 (16)
PixelBandwidth	130 (714), 140 (735), 240 (2), 260 (3), 751 (6)
TransmittingCoil	body (1259), cp_head (53), txrx_head (148)
AcquisitionMatrix	[0 256 176 0] (3),[0 256 208 0] (16), [0 256 256 0] (1433), [0 256 258 0] (6), [0 320 320 0] (2)
FlipAngle	120 (6), 15 (645), 7 (100), 8 (707), 9 (2)
PixelSpacing	[0.5 0.5] (17), [0.8 0.8] (2), [1.0 1.0] (1440), [1.1 1.1] (1)
LAccelFactPE	1 (787), 2 (662), 3 (3), 4 (5), null (3)
tcoilID/ReceivingCoil	32ch_head (281), 8ch_head (667), body (1), cp_head (430), headmatrix (78), null (3)

Note: Each scanning setting is followed by the number of subjects that have been scanned using this setting. Scanning parameters with small variances across the subjects (displayed in gray) are excluded from the subsequent analysis.

MCIC Data

Subjects: The second dataset included 234 subjects from the Mind Clinical Imaging Consortium (MCIC) study (Gollub et al., in press), a collaborative effort of four research teams from University of New Mexico-Mind Research Network, Massachusetts General Hospital, University of Minnesota and University of Iowa. The institutional review board at each site approved the study and all subjects provided written informed consents. The cohort consisted of 110 SZ patients and 124 healthy controls. All healthy subjects were screened to ensure that they were free of any medical, neurological or psychiatric illnesses, including any history of substance abuse. The inclusion criteria for patients were based on a diagnosis of schizophrenia, schizophreniform or schizoaffective disorder confirmed by the Structured Clinical Interview for DSM-IV-TR disorders (SCID, (First et al., 1997)) or the Comprehensive Assessment of Symptoms and History (CASH, (Andreasen et al., 1992)). Table 6.2 provides the demographic information.

Neuroimaging: The brain images were coronal T_1-weighted MRIs collected at multiple sites. Subjects were scanned using different scanners, i.e. 1.5T Siemens Sonata and GE Signa, as well as 3.0T Siemens Trio. Closely matched acquisition sequences were used. Compared to the BIG data, the MCIC data have relatively limited scanning information, as summarized in Table 6.3 (Segall et al., 2009). Slight collinearity was observed between SZ diagnosis and scanner (r^2 = 0.043), as well as slice thickness (r^2 = 0.039).

Table 6.2: Demographic information of MCIC subjects.

Demographics		**SZ (110)**	**HC (124)**
Sex	Male	82	75
	Female	28	49
Age	Mean ± SD	35 ± 11	32 ± 11
	Range	18 - 60	18 - 58
Race/Ethnicity	Caucasian	83	110
	African American	17	4
	Asian	5	5
	American Indian	1	1
	Unreported	4	4
Collection site	Harvard	28	23
	Iowa	32	60
	Minnesota	29	18
	New Mexico	21	23

Table 6.3: Summary of MCIC scanning parameters.

Site	**Scanner**	**Field strength (T)**	**TR/TE (ms)**	**Slice thickness (mm)**	**Bandwidth**	**Voxel dimensions (mm)**	**Sequence**
M021 (51)	Siemens Sonata	1.5	12/4.76	1.5	110	0.625 × 0.625 × 1.5	Gradient Echo
M522 (92)	GE Signa	1.5	20/6	1.6	122	0.6641 × 0.6641 × 1.6	Gradient Echo
M554 (47)	Siemens Trio	3	2530/3.81	1.5	110	0.625 × 0.625 × 1.5	MP-RAGE
M871 (44)	Siemens Sonata	1.5	12/4.76	1.5	110	0.625 × 0.625 × 1.5	Gradient Echo

Note: Each site is followed by the number of subjects that have been scanned at that station.

Preprocessing

The T_1-weighted sMRI data were preprocessed in Statistical Parametric Mapping 5 (SPM5, http://www.fil.ion.ucl.ac.uk/spm) using unified segmentation (Ashburner and Friston, 2005), in which image registration, bias correction and tissue classification are performed using a single integrated algorithm. Brains were segmented into gray matter, white matter and cerebrospinal fluid and non-linearly transformed into the ICBM152 standard space (without Jacobian modulation). The resulting GMC images were re-sliced to 2 × 2 × 2 mm, with a field-of-view of 91 × 109 × 91 voxels. In the subsequent quality check, we excluded 4 subjects from the BIG data whose GMC images were four standard deviations away from the average GMC image across all subjects. No outliers were found for the MCIC data. A mask was then generated (mean GMC > 0) to include only the segmented gray matter voxels, resulting in a total of 298,707 voxels for the BIG data and 292,998 voxels for the MCIC data. Finally, a voxelwise linear regression analysis was performed to remove the effects of age and sex.

Analysis

The unsmoothed GMC images corrected for age and sex were investigated for associations with scanning parameters in SBM model, as illustrated in Figure 6.1. SBM analysis consists of data decomposition using ICA (infomax) (Amari, 1998; Bell and Sejnowski, 1995b) and association tests between loadings and scanning parameters. First, spatial ICA is applied to decompose the group GMC images into a linear combination of independent components (ICs), or sources, as illustrated in Figure 6.1.1 and (6.1). **X**, **A** and **S** denote the observed data, loading matrix and component matrix, respectively. Each row of **S** represents an underlying component and each column of **A** represents the loadings associated with one single component. The subscripts *m*, *n* and *l* denote the number of samples, voxels and components, respectively. After data decomposition, each of the extracted loadings (i.e. each column of **A**) is then assessed for association with each continuous or categorical scanning parameter using regression or ANOVA, respectively. The pairwise association tests result in a series of p-values. Given the dependence among scanning parameters, we choose to estimate the threshold for significant associations (p_{th}) based on the p-value distribution. Specifically, we plot the p-values ($-log_{10}(p)$) in a descending order and then perform linear fits to the two segments of the curve, as shown in Figure 6.1.3. The intersection *A* of the fitted lines *L1* and *L2* is then determined. Subsequently, we connect the origin and the intersection *A* to obtain the line *L3*, which is extended to intersect the p-value curve at the point *B*. The p-value corresponding to *B* is then selected as the threshold p-value. Compared with the commonly used FDR control, our approach is conservative and yields a more stringent threshold p-value.

$$X_{M\times N} = A_{M\times L}S_{L\times N}$$
$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix},\ A = [a_1, \quad \cdots \quad a_L],\ S = \begin{bmatrix} s_1 \\ \vdots \\ s_L \end{bmatrix}$$
$$Y_{L\times N} = W_{L\times M}X_{M\times N} \tag{6.1}$$

Based on the estimated threshold p-value (p_{th}), the scanning parameters significantly affecting ($p <= p_{th}$) the image values and the IC loadings can be identified. To be more cautious, the identified components are also examined for associations with traits of interest. If a component is identified as scanning-related while not exhibiting any significant effect of interest, a correction can then be performed to eliminate the scanning-related IC to improve data integrity. For instance, if the k^{th} component is to be corrected, we simply subtract the reconstructed $\mathbf{X}_k$ from the original dataset $\mathbf{X}$ to eliminate the variance induced by that factor, as illustrated in (6.2). The corrected data are denoted as $\mathbf{X}_c$.

$$X_c = X - X_k = \left(\sum_{i=1}^{l} a_i s_i\right) - (a_k s_k) \tag{6.2}$$

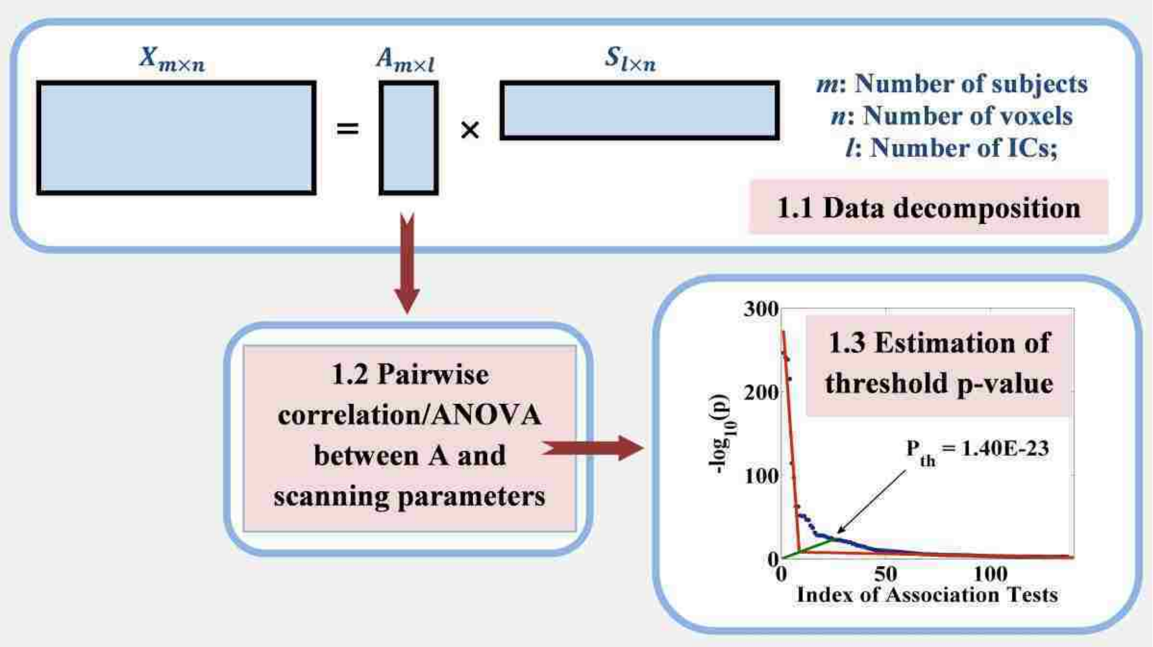

Figure 6.1: A flow chart of the SBM model.

The above procedure was applied to the BIG data to explore the pivotal parameters significantly affecting the image pattern. We first excluded scanning parameters with few variations across the subjects (displayed in gray in Table 6.1). Then we performed a pruning process to further exclude highly collinear ($r^2 > 0.85$) parameters, after which 8 parameters were retained. In particular, the inversion time (TI) was modeled as a function of the field strength (i.e. inversion time per field strength), as the direct effect of TI depends on the relaxation times which are different per field strength. In addition, for each subject, we calculated the signal-to-noise ratio (SNR) of the image, which is proportional to the ratio of the average T_1-contrast in the gray and white matter regions over the standard deviation of the T_1-contrast in air regions (Henkelman, 1985), as described in (6.3). The gray and white matter regions are defined as voxels exhibiting gray or white matter concentrations greater than 0.5, and the air regions are voxels with a relatively low signal intensity, as shown in (6.3). The SNR parameter was also investigated for its effect on the image pattern.

$$SNR = 0.655 \cdot \frac{mean\left(T_1 IMG_{gray/white}\right)}{std(T_1 IMG_{air})}$$

$$Air = \left\{voxels | T_1 IMG < 0.2 \cdot mean\left(T_1 IMG_{gray/white}\right)\right\} \tag{6.3}$$

Besides the exploration with the BIG data, we evaluated the detection and correction procedure in a second dataset by examining GMC images of 234 subjects (110 SZ patients and 124 healthy controls) from the MCIC study. First the data were corrected using SBM and general linear model (GLM), respectively. The SBM correction used the same procedure as described above, where components were extracted and subsequently assessed for associations with three scanning parameters while controlling for SZ diagnosis. The included parameters were scanner which was completely collinear with TR/TE (as shown in Table 6.3), field strength which was collinear with sequence, as well as slice thickness which was collinear with bandwidth and voxel dimensions. For GLM correction, the same three scanning parameters and SZ diagnosis were included as predictors. The estimated scanning effects were then regressed out at each voxel from the original data. The uncorrected, SBM- and GLM-corrected data were subsequently compared regarding the scanning and SZ effect sizes. Specifically, we investigated the scanning and disease effects using both component (ICA) and voxelwise (VBM) approaches. For ICA, the scanning and SZ effects were assessed based on the associations of extracted IC loadings with scanning parameters and SZ diagnosis. For VBM, a univariate analysis was performed to examine the associations between GMC and scanning parameters or SZ diagnosis at each voxel. Then voxels exhibiting significant scanning or SZ effects were identified using a false discovery rate (FDR) control for multiple comparisons.

Component Number Selection

PCA was applied before ICA for data whitening and reduction. It was noted that the PC variance curve turned between component 50 and 100, and the top 100 components explained a relatively larger amount of variance than the remainder. We thus performed ICA with the component numbers from 50 to 100, and found that the most significant scanning-related components (due to magnetic field strength and receiving coil) remained stable within the tested range. Meanwhile, with the increase of component numbers, edge effects appeared to be refined, manifested as increases in the level of significance. Given these observations, we chose to perform the SBM analysis with a component number of 100.

6. 3 Results

BIG Data

We applied ICA to extract 100 components from the GMC images. The 100 resulting ICs were then assessed for associations with 9 scanning parameters, as listed in Table 6.4.

Based on the resulting p-values, the threshold of significance (p_{th}) was estimated to be 1.40×10^{-23}, as shown in Figure 6.1.3. Nine ICs were significantly associated ($p <= p_{th}$, highlighted in bold in Table 6.4) with various parameters, including magnetic field strength and receiving coil. Figure 6.2 shows the spatial maps of the scanning-related ICs, thresholded at |Z| > 2.

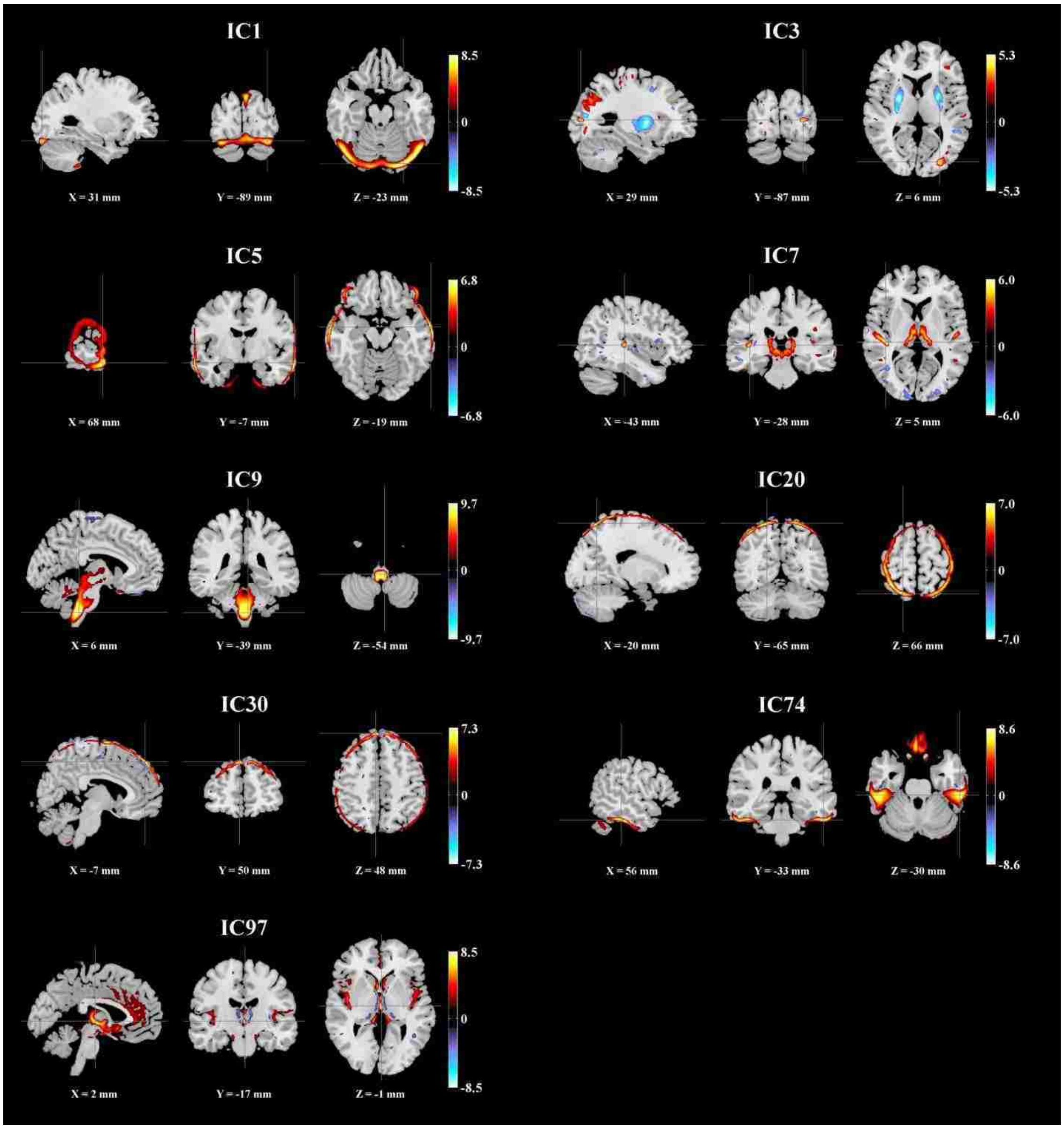

Figure 6.2: Spatial maps of the scanning-related components identified in the BIG data.

Table 6.4: Scanning effects in the BIG sMRI data.

Scanning/IC Index	1	3	5	7	9	20	30	74	97
StationName	**1.65E-40**	**1.44E-52**	**1.30E-28**	**2.98E-63**	**1.48E-241**	**9.80E-38**	1.63E-20	**7.80E-98**	**1.24E-25**
SequenceName	6.64E-01	1.91E-01	9.45E-01	9.41E-21	6.21E-17	6.01E-01	9.28E-01	3.10E-03	2.73E-03
TI-FieldStrength	2.08E-11	**2.73E-63**	**2.93E-32**	3.64E-06	**5.56E-239**	**6.10E-52**	**3.09E-29**	**1.32E-28**	1.09E-18
FieldStrength	2.40E-05	**3.98E-47**	2.11E-05	4.62E-07	**4.44E-216**	**1.65E-28**	3.50E-23	2.89E-23	1.45E-21
PixelBandwidth	8.86E-10	**6.17E-52**	2.25E-08	3.95E-11	**2.63E-247**	**1.50E-27**	1.13E-21	2.73E-23	7.89E-19
TransmittingCoil	1.12E-16	7.27E-08	8.09E-10	**1.22E-47**	**2.73E-26**	1.86E-03	5.91E-05	9.46E-14	1.26E-07
laccelFactPE	4.68E-13	2.77E-09	1.23E-12	7.78E-16	6.76E-15	**1.74E-25**	1.28E-04	3.91E-08	2.24E-03
tCoilID	3.83E-04	1.65E-15	1.06E-03	**3.86E-115**	1.50E-22	1.85E-10	2.11E-01	6.25E-11	3.07E-06
SNR	6.79E-01	8.34E-01	3.10E-06	**1.40E-23**	5.46E-05	1.75E-03	2.20E-01	5.02E-01	1.05E-06

Note: Significant scanning effects ($p < p_{th}$) are highlighted in bold.

Table 6.5: Talairach regions of the scanning-related components (BIG data).

Component	Brain region	Brodmann area	L/R volume (cm3)	L/R random effects, max Z (x,y,z)
IC3 pos	Superior Frontal Gyrus	8, 6, 9, 10, 11	9.1/10.2	3.73(-6,45,46)/4.07(4,45,46)
	Middle Frontal Gyrus	9, 6, 8, 10, 46	5.8/5.4	3.12(-30,46,35)/2.77(22,22,58)
	Postcentral Gyrus	5, 7, 3, 40, 2, 1	3.4/3.5	2.62(-12,-43,70)/2.60(16,-45,69)
	Inferior Parietal Lobule	40, 7, 39, 2	2.9/2.8	2.64(-50,-52,50)/2.88(48,-44,54)
	Superior Parietal Lobule	7	2.3/2.6	2.73(-34,-49,61)/3.01(24,-57,62)
	Precentral Gyrus	6, 4, 9, 44	2.1/2.4	2.55(-46,-1,53)/2.57(22,-20,67)
	Precuneus	7, 19, 39, 31	1.3/1.7	2.03(-22,-71,51)/2.51(4,-49,63)
IC3 neg	Lentiform Nucleus		3.7/4.8	7.75(-30,-14,1)/8.53(28,-17,3)
	Middle Temporal Gyrus	21, 20, 39, 37, 22, 19, 38	1.7/1.4	7.29(-42,1,-29)/6.68(53,-37,-3)
	Precuneus	7, 31, 19, 39	1.7/1.1	7.50(-22,-68,33)/5.72(16,-47,34)
	Middle Frontal Gyrus	10, 46, 8, 6, 11, 9, 47	1.5/0.9	6.36(-32,38,17)/6.35(40,13,20)
	Superior Temporal Gyrus	22, 13, 38, 39, 21, 41	1.4/0.8	6.18(-53,0,-3)/5.77(46,-42,24)
IC7 pos	Thalamus		4.2/4.4	6.00(-10,-23,1)/6.07(8,-23,1)
	Superior Temporal Gyrus	22, 41, 13, 39, 38, 21, 42	1.5/1.8	4.75(-50,-17,5)/6.79(44,-25,5)
	Middle Temporal Gyrus	21, 37, 22, 19, 39, 20	1.9/1.0	3.74(-57,7,-19)/3.65(50,-26,-7)
	Middle Frontal Gyrus	10, 6, 9, 11, 8, 46, 47	1.1/1.2	4.10(-26,64,6)/3.28(26,32,26)
IC9 pos	Thalamus		2.2/2.7	5.48(-6,-29,-4)/5.54(6,-29,-5)
	Rectal Gyrus	11	0.9/1.0	4.80(-10,16,-23)/7.18(4,14,-21)
IC97 pos	Superior Temporal Gyrus	38, 22, 13, 41, 42, 21, 39	4.2/3.8	6.58(-42,11,-16)/7.18(40,11,-16)
	Anterior Cingulate	25, 24, 33, 32	2.7/0.4	6.77(0,6,-5)/4.46(2,3,-10)
	Insula	13, 22, 40, 41, 47	2.4/3.2	5.91(-44,-11,10)/6.40(44,-6,0)
	Parahippocampal Gyrus	34, 28, 35, 27, 30, 19, 36	2.3/2.0	8.08(-10,-9,-16)/8.09(10,-7,-18)
	Inferior Frontal Gyrus	47, 13, 46, 44, 9, 45, 11	1.7/1.8	6.23(-40,15,-16)/6.90(38,13,-17)
	Medial Frontal Gyrus	25, 10, 9, 6, 11, 8, 32	1.7/0.8	4.66(0,20,-18)/4.72(10,7,-19)
	Thalamus		1.6/0.8	8.26(0,-16,1)/7.77(6,-31,3)
	Cingulate Gyrus	24, 32, 31, 23	1.4/0.1	4.80(0,15,27)/3.48(8,-31,35)
	Caudate		1.3/1.2	7.69(-6,10,3)/7.39(6,6,5)

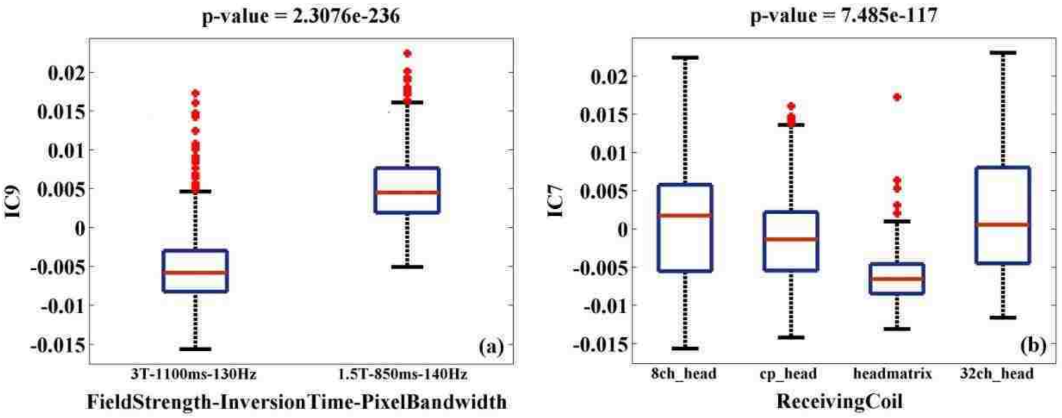

Figure 6.3: Boxplots of two components exhibiting the most significant scanning effects in the BIG data; (a) IC9 loadings versus magnetic field strength-inversion time-pixel bandwidth; (b) IC7 loadings versus receiving coil.

IC1, 5, 20, 30 and 74 reflected likely scanning effects at brain edges, while IC97 reflected scanning effects in the ventricle region. IC9 was predominantly located in the brainstem region and exhibited the most significant scanning effect, associated with station, TI-field strength, magnetic field strength and pixel bandwidth, among which slight collinearity ($r^2 > 0.088$) was observed. Specifically, 634 out of 1,460 subjects were scanned in 1.5T scanners with a 850ms inversion time and 140Hz pixel bandwidth. These subjects exhibited higher regional GMC in IC9 compared to 704 subjects scanned in 3T scanners with a 1,100ms inversion time and 130Hz pixel bandwidth, as shown in Figure 6.3a. Meanwhile, the type of receiving coil showed a unique effect on IC7. This component was primarily localized to the thalamus region and reflected higher regional GMC in subjects scanned with multichannel phased-array coils, as shown in Figure 6.3b. Table 6.5 provides a summary of the Talairach atlas labels (Lancaster et al., 1997; Lancaster et al., 2000) of for IC3, 7, 9 and 97 thresholded at |Z| > 2. It needs to be pointed out that in this work the components were mapped to the nearest gray matter, therefore brainstem areas were not included in the table.

MCIC Data

We applied the same procedure to the MCIC GMC images. 100 ICs were extracted and assessed for associations with scanner, field strength and slice thickness, while controlling for SZ diagnosis. The threshold of significance (p_{th}) was estimated to be 5.28×10^{-3}, based on which 8 ICs were identified as scanning-related, as summarized in Table 6.6. Figure 6.4 shows the spatial maps of these ICs, thresholded at |Z| > 2. It can be seen that scanning effects were again observed at brain edges, as represented by IC9, 63 and 92, as well as in the ventricle region, as reflected by IC 98. For the remaining scanning-related components, Table 6.7 provides a summary of the Talairach atlas labels of mapped gray matter regions with components thresholded at |Z| > 2. The most significant scanning effect was observed in IC 53 (inferior temporal region), associated with scanner and field strength. Again moderate collinearity was observed between these 2 parameters ($r^2 > 0.64$). Boxplots illustrated that scans acquired with lower field strength and shorter TR exhibited higher regional GMC in IC53, as shown in Figure 6.5. Finally the 8 scanning-related components were used for SBM-based data correction.

Table 6.6: Scanning effects in the MCIC sMRI data.

Scanning / IC Index	**7**	**9**	**33**	**38**	**53**	**63**	**92**	**98**
Scanner	2.06E-02	**1.94E-05**	5.51E-03	**5.18E-03**	**8.62E-53**	**2.68E-20**	**2.71E-09**	**2.96E-04**
FieldStrength	2.87E-01	**3.33E-06**	3.98E-01	8.17E-01	**3.57E-49**	1.54E-02	1.02E-01	**1.54E-03**
SliceThickness	**5.28E-03**	1.49E-01	**1.37E-03**	**3.84E-03**	5.88E-02	**4.14E-21**	**4.47E-10**	**4.69E-04**

Note: Significant scanning effects ($p < p_{th}$) are highlighted in bold.

Table 6.7: Talairach regions of the scanning-related components (MCIC data).

Component	Brain region	Brodmann area	L/R volume (cm^3)	L/R random effects, max Z (x,y,z)
IC7 pos	Middle Frontal Gyrus	11, 9, 8, 6, 46, 10, 47	2.7/2.1	4.76(-36,40,-10)/4.13(38,21,30)
	Inferior Frontal Gyrus	47, 9, 44, 45, 13, 10, 46	1.6/1.2	3.83(-55,15,29)/3.91(32,13,-17)
	Superior Frontal Gyrus	9, 6, 8, 10, 11	1.4/2.1	4.30(-16,58,28)/4.86(16,54,32)
	Superior Temporal Gyrus	21, 22, 39, 42, 38, 41, 13	1.4/1.9	4.48(-59,-8,-1)/4.34(40,-55,27)
IC7 neg	Middle Frontal Gyrus	6, 9, 11, 10, 8, 46, 47	2.6/2.4	4.32(-38,27,28)/4.79(28,8,44)
	Middle Temporal Gyrus	20, 39, 21, 19, 22, 37, 38	2.1/2.2	5.20(-53,-43,-10)/4.97(42,-59,20)
IC33 pos	Middle Frontal Gyrus	9, 6, 46, 10, 8, 11, 47	5.3/2.8	7.17(-36,13,31)/5.41(36,14,49)
	Superior Temporal Gyrus	41, 39, 42, 22, 38, 13, 21	3.8/3.0	8.24(-46,-39,6)/5.15(30,8,-29)
	Superior Frontal Gyrus	8, 6, 10, 11, 9	3.6/2.9	4.55(-12,30,48)/5.52(38,16,47)
	Precuneus	39, 7, 31, 19	3.3/3.0	7.03(-34,-62,36)/4.16(20,-76,28)
	Precentral Gyrus	9, 44, 6, 3, 43	3.2/1.6	7.65(-36,11,31)/7.28(48,12,10)
	Middle Temporal Gyrus	22, 21, 39, 37, 20, 38, 19	3.0/2.5	7.20(-48,-39,6)/5.96(55,-26,-10)
	Inferior Parietal Lobule	39, 40, 7, 2	2.6/1.6	7.48(-36,-62,38)/4.45(44,-48,45)
	Inferior Frontal Gyrus	44, 9, 46, 45, 13, 47, 10, 11	2.2/2.5	6.17(-34,9,29)/7.59(48,12,12)
IC33 neg	Middle Frontal Gyrus	6, 8, 10, 9, 46, 11, 47	3.3/2.3	5.10(-36,10,47)/4.57(38,14,53)
	Superior Frontal Gyrus	6, 10, 8, 9, 11	2.8/2.5	4.99(-20,58,25)/5.34(12,22,58)
	Middle Temporal Gyrus	39, 21, 20, 19, 37, 22, 38	2.2/1.9	5.40(-30,-61,31)/4.75(55,-9,-18)
	Inferior Parietal Lobule	40, 7, 39	2.2/1.3	6.11(-38,-49,39)/4.23(65,-42,24)
	Precuneus	7, 31, 19, 39	2.0/2.1	8.09(-24,-60,36)/4.63(14,-59,31)
IC38 pos	Middle Frontal Gyrus	10, 8, 6, 9, 11, 46, 47	3.4/2.6	6.06(-36,39,13)/5.14(32,37,11)
	Superior Frontal Gyrus	10, 6, 9, 8, 11	2.4/1.9	4.97(-24,44,18)/4.50(20,43,38)
	Cuneus	18, 7, 19, 17, 30, 23	2.2/2.3	6.07(-20,-81,21)/5.55(16,-74,26)
	Parahippocampal Gyrus	30, 19, 37, 36, 27, 28, 35	2.1/1.0	5.52(-12,-41,6)/4.35(24,-32,-9)
	Lingual Gyrus	19, 18:*, 17, 30	1.9/1.2	4.30(-20,-95,-2)/4.46(16,-47,-3)
	Inferior Frontal Gyrus	9, 47, 45, 46, 13, 10, 44, 6	1.5/1.8	4.74(-36,9,25)/4.33(30,28,-13)
IC38 neg	Middle Frontal Gyrus	46, 9, 10, 6, 8, 11, 47	2.9/1.9	4.65(-42,51,14)/3.83(32,23,38)
	Superior Frontal Gyrus	10, 6, 9, 11, 8	2.1/1.4	5.40(-20,46,23)/4.02(12,56,30)
	Middle Temporal Gyrus	22, 37, 39, 19, 21, 20, 38	1.8/1.9	4.80(-50,-41,4)/4.12(42,-55,-2)
	Parahippocampal Gyrus	30, 35, 27, 36, 28, 34, 19	1.8/0.9	6.64(-8,-39,4)/4.03(10,-41,-1)
	Superior Temporal Gyrus	38, 39, 21, 41, 22, 42, 13	1.6/1.5	3.56(-46,-39,6)/4.46(24,8,-26)
	Inferior Frontal Gyrus	45, 44, 47, 10, 9, 11, 13, 46	1.5/1.5	5.66(-38,45,0)/4.65(34,32,-18)
IC53 pos	Inferior Temporal Gyrus	20, 21, 37, 19	3.5/3.4	10.76(-46,-21,-28)/8.66(46,-21,-28)
	Superior Frontal Gyrus	11, 6, 10, 8, 9	3.1/2.7	6.24(-8,57,-23)/5.41(10,53,-23)
	Middle Frontal Gyrus	11, 9, 10, 8, 46, 6, 47	3.0/2.1	4.01(-32,39,11)/4.33(42,48,-14)
	Middle Temporal Gyrus	38, 39, 22, 21, 20, 37, 19	2.5/2.1	4.46(-36,10,-37)/4.36(55,-47,2)
	Fusiform Gyrus	20, 36, 18, 19, 37	2.3/2.6	8.67(-50,-23,-27)/9.08(50,-25,-26)
	Inferior Frontal Gyrus	47, 9, 11, 45, 46, 13, 44, 10	1.9/1.6	3.72(-50,38,-14)/3.56(53,3,27)
	Superior Temporal Gyrus	38, 22, 13, 39, 42, 41, 21	1.7/2.4	4.42(-22,10,-36)/5.44(30,4,-39)
IC53 neg	Precentral Gyrus	6, 4, 44, 9, 13, 43, 3	3.8/4.2	4.56(-30,-7,52)/4.84(18,-18,65)
	Superior Frontal Gyrus	6, 8, 11, 9, 10	3.4/2.5	4.64(-8,3,68)/5.54(14,-12,71)
	Middle Frontal Gyrus	6, 11, 8, 46, 10, 9, 47	2.2/1.2	5.57(-20,-1,61)/4.80(18,-7,59)
	Medial Frontal Gyrus	6, 9, 10, 8, 25, 32, 11	1.9/2.4	4.64(-6,-20,71)/5.01(6,-20,71)

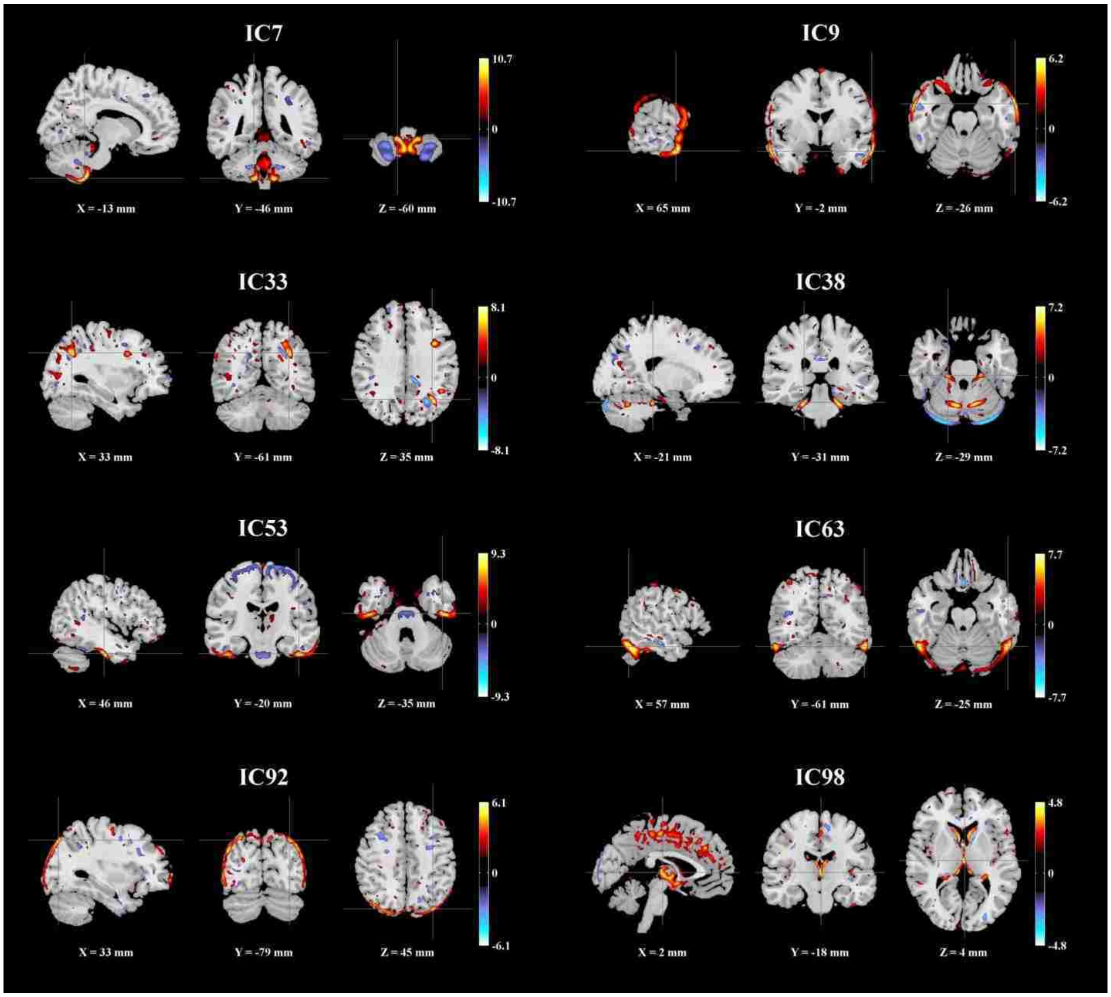

Figure 6.4: Spatial maps of the scanning-related components identified in the MCIC data.

Table 6.8: Comparisons of the scanning and SZ effects in the uncorrected, SBM- and GLM-corrected data evaluated with ICA and VBM (MCIC data).

a. ICA (p-value)	Uncorrected	SBM-corrected	GLM-corrected
Scanner	8.62E-53	0.01	0.32
MagneticFieldStrength	3.57E-49	0.05	0.48
SliceThickness	4.14E-21	0.01	0.35
SZ	5.50E-06 (IC94)	5.55E-07 (IC81)	1.40E-06 (IC12)
	1.72E-05 (IC98)		1.72E-04 (IC15)
	3.00E-04 (IC53)		7.24E-04 (IC34)
	2.36E-03 (IC7)		7.80E-04 (IC91)

b. VBM (number of voxels passing FDR)	Uncorrected	SBM-corrected	GLM-corrected
Scanner	20265	0	0
MagneticFieldStrength	22318	0	0
SliceThickness	10623	0	0
SZ	421	14	228

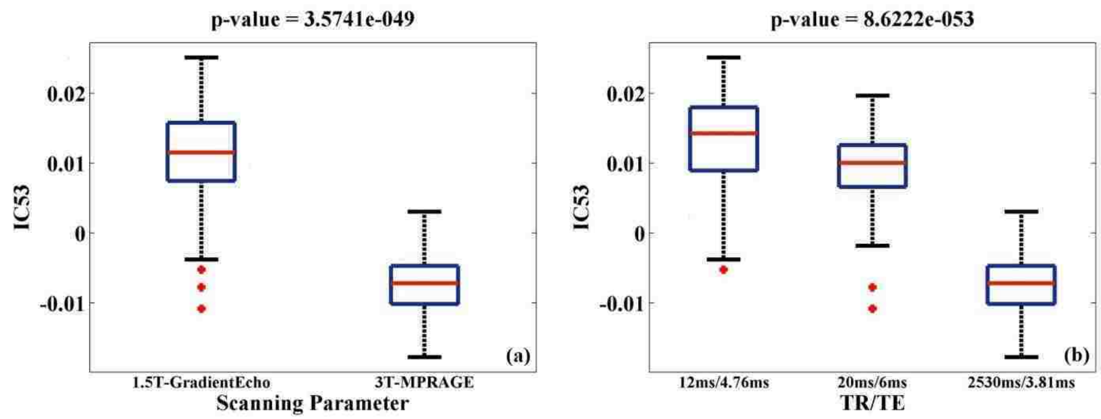

Figure 6.5: Boxplots of the component exhibiting the most significant scanning effect in the MCIC data; (a) IC53 loadings versus magnetic field strength-Sequence; (b) IC53 loadings versus TR/TE.

Both SBM and GLM correction appeared to effectively eliminate the scanning effects when evaluated with ICA, as summarized in Table 6.8a. Meanwhile, the SZ effect was refined with increasing significance levels. Figure 6.6a shows the spatial maps of the SZ-related components (thresholded at $|Z| > 2$) identified in the uncorrected, SBM- and GLM-corrected data. It can be seen that the mapped brain regions were highly consistent, highlighting a frontal-temporal network. IC94 identified in the uncorrected data exhibited a SZ effect with a p-value of 5.50×10^{-6} while controlling for scanning parameters. IC12 identified in the GLM-corrected data exhibited a more significant effect with p-value decreasing to 1.40×10^{-6}. The most significant SZ effect was observed from IC81 identified in the SBM-corrected data, presenting a p-value of 5.55×10^{-7}, as shown in Table 6.8a. The uncorrected data also presented marginal SZ effects captured by IC7, 53 and 98, whose spatial maps have been illustrated in Figure 6.4. Marginal SZ effects were also observed in the GLM-corrected data, as captured by IC15, 34 and 91, whose spatial maps are illustrated in Figure 6.6b.

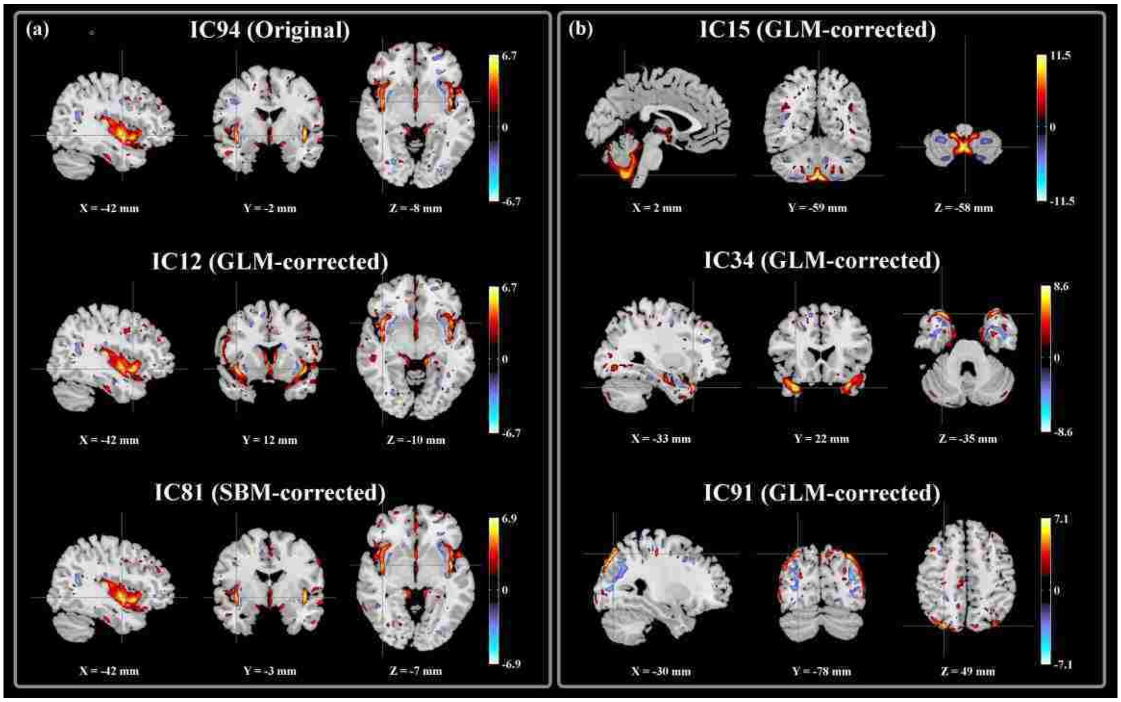

Figure 6.6: Spatial maps of the SZ-related components identified with ICA (|Z| > 2): (a) the frontal temporal network identified in the uncorrected (IC94), GLM-corrected (IC12) and SBM-corrected (IC81) data, respectively; (b) components exhibiting marginal SZ effects in the GLM-corrected data.

For VBM analyses, no voxels exhibited significant scanning effects passing FDR control after either SBM or GLM correction while number of voxels exhibiting significant SZ effect decreased after correction, as summarized in Table 6.8b. In the uncorrected data, 421 voxels showed SZ effect with scanning parameters included as covariates. In the SBM- and GLM-corrected data, 14 and 228 SZ-related voxels were identified, respectively. When we used a liberal threshold of 0.05, 29,853 voxels showed SZ effect in the uncorrected data, 23,836 voxels in the SBM-corrected data and 28,117 voxels in the GLM-corrected data. Figure 6.7 illustrates the spatial maps of the highlighted voxels for the three cases, where the uncorrected and the GLM-corrected data presented highly comparable maps, highlighting more voxels in the brainstem region compared to the map presented by the SBM-corrected data.

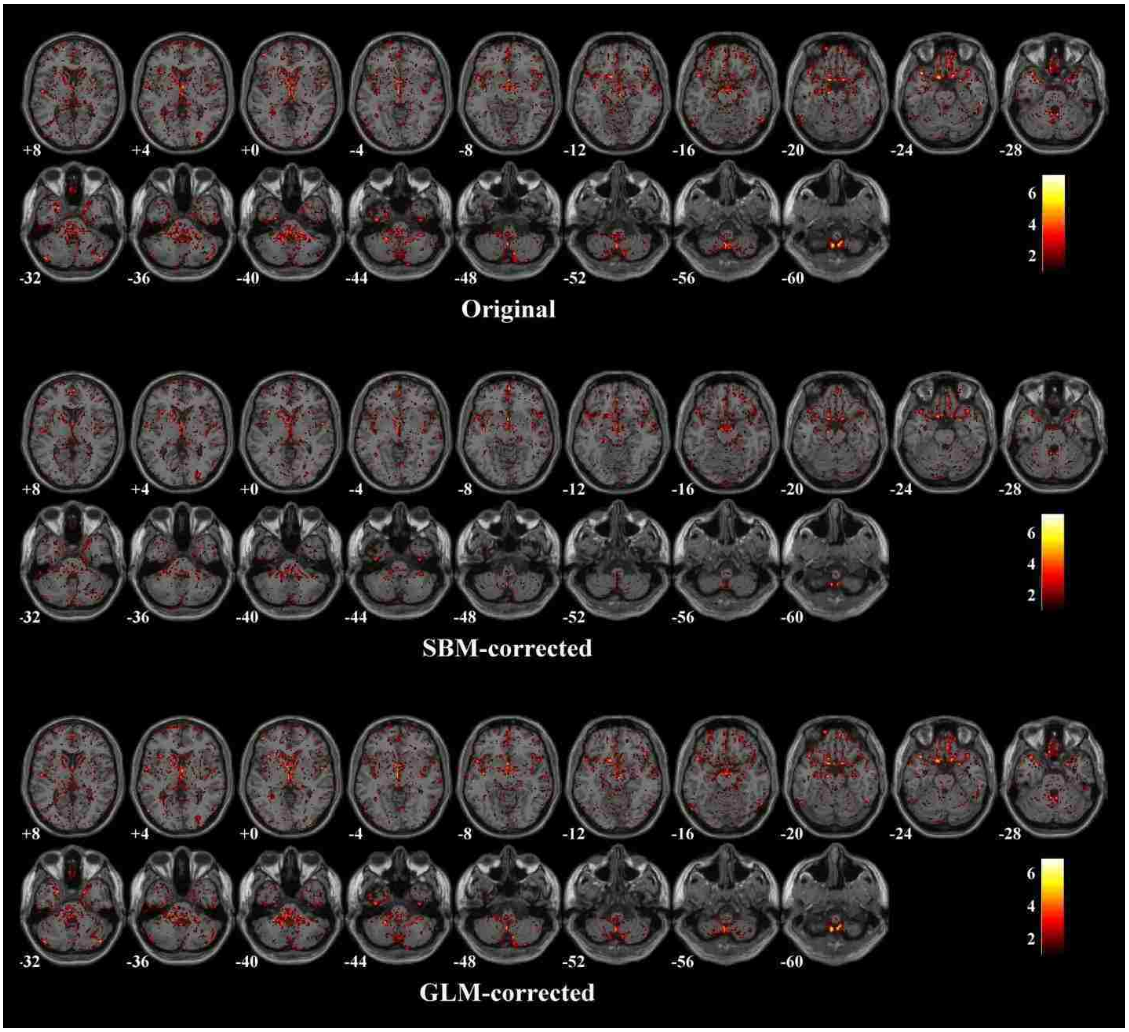

Figure 6.7: Spatial maps of the SZ-related voxels ($p < 0.05$) identified with VBM in the uncorrected, SBM-corrected and GLM-corrected data, respectively.

6. 4 Discussion

Pooling of multi-site structural MRI scans is desired in large scale brain morphometry analyses, which lead to increased statistical power and provide opportunities to identify reliable biomarkers. However, the true effect of interest may be confounded by the systematic differences introduced by inconsistent acquisition schemes and scanning platforms. This issue is especially inevitable in longitudinal studies. Thus it becomes important to investigate the image comparability when different scanning platforms are involved. In this study, we explore the effects of various scanning parameters and determine if a data correction is applicable in the SBM framework. The exploration was performed with GMC images collected from 1,460 healthy subjects. As expected, we observed significant scanning effects in distributed brain regions. The most pronounced effects were observed from magnetic field strength and receiving coil. In the second study with the MCIC data of 110 SZ patients and 124 healthy controls, the results confirmed significant scanning effects. In addition, it was also demonstrated that the SBM approach effectively separated scanning effects from the SZ-related GMC changes, thus enabling a correction which helped refine the true effect of interest.

BIG Data

In the exploration with the BIG data, the most significant scanning effect is observed in IC9. This component highlights the brainstem region, where the GMC exhibits differences among subjects scanned at various stations, which primarily involve differences in magnetic field strength, inversion time and pixel bandwidth. Magnetic field strength affects the T_1-relaxation and, hence, the imaging contrast between gray and white matter (Duewell et al., 1996), which is consistent with our observation. Higher field strength also results in increased magnetic susceptibility artifacts (Bernstein et al., 2006). The brainstem is particularly prone to these artifacts (Focke et al., 2011), which can cause geometric distortions and signal loss, and influence the effective excitation field and flip angle, thus affecting contrast in T_1-weighted images (Truong et al., 2006). Inversion time is typically chosen in line with T_1-relaxation (and hence field strength) as it determines the magnetization before excitation in each tissue, and thus the T_1-contrast. Pixel bandwidth, or receiver bandwidth, refers to the difference in magnetic resonance frequencies between adjacent pixels. This parameter is most commonly associated with chemical shift between fat and water and has a direct effect on image SNR (Schmitz et al., 2005). In the present study, it is difficult to disentangle the effects of individual parameters due to collinearity. However, it appears likely that the GMC variability observed in brainstem region is majorly attributable to the inversion time-field strength interaction.

A second notable scanning effect is observed in IC7. This component highlights the thalamus region and reflects GMC variability induced by RF coils, especially the receiving coil. As illustrated in Table 6.1, the majority of the cohorts (1259 out of 1460) were scanned using the same type of transmitting coil. Therefore, the present data may not appropriately reflect how transmitting coil influences the image pattern. Moreover, effects of transmitting coil only become more substantial at 7T or higher (Vaughan et al., 2001). Regarding the receiving coil, Figure 6.3b shows that subjects scanned with 32-channel head coil exhibit higher GMC in the highlighted thalamus region of IC7. Meanwhile it is noteworthy that IC7 is the only component associated with SNR ($r = 0.26$, $p = 1.40\times10^{-23}$). Not surprisingly, SNR exhibits a significant group difference among different types of receiving coils ($p = 3.28\times10^{-46}$), where 32-channel head coil yields the highest overall SNR and 8-channel head coil the second. This observation is consistent with previous studies that have found spatially dependent gains in SNR with the addition of element coils in multichannel phased-array head coils (de Zwart et al., 2004; Wintersperger et al., 2006). Overall, our finding reveals interrelationships between SNR and RF receiving coil, and indicates that coil design may lead to a significant variability in the image pattern. Given these observations, it is strongly recommended that in addition to calibrating magnetic field strength and inversion time, inconsistency in RF coil designs should also be avoided in aggregated sMRI analyses.

MCIC Data

The second study with the MCIC data confirms significant systematic differences in the image pattern induced by scanning parameters. The most affected component is IC53, highlighting inferior temporal region and showing a relationship with scanner and field strength. It should be noted that magnetic field strength is completely collinear with sequence, and scanner completely collinear with TR/TE in the MCIC data. Magnetic field strength, as discussed above, can significantly influence the T_1-contrast. The observation in the MCIC data is consistent with the BIG data in that scans obtained with lower field strength exhibit higher regional GMC, as shown in Figure 6.5a. Repetition time determines the recovery of magnetization and directly affects the T_1-contrast. The boxplot with TR/TE (Figure 6.5b) illustrates that MCIC scans acquired with shorter repetition time exhibit higher regional GMC, consistent with the general concept of shorter TR leading to higher contrast. In contrast, no linear relation is observed between the component loadings and TE, suggesting that the observed image variability is more attributable to repetition time instead of echo time. Although due to collinearity we cannot determine which parameter is the major contributor to the image variability observed in IC53, the results confirm that inconsistency in field strength and sequence design can introduce significant systematic differences in multi-site sMRI studies (Fennema-Notestine et al., 2007; Stonnington et al., 2008).

IC 53, IC98 and IC7 exhibit association with both scanning parameters and SZ diagnosis, as shown in Table 6.6 and Table 6.8a. Due to the collinearity, it is impossible to accurately estimate the effects of individual parameters. However, IC53 is much more significantly associated with scanning settings ($p = 8.62\times10^{-53}$, 3.57×10^{-49}, Table 6.6) than SZ diagnosis ($p = 3.00\times10^{-4}$, Table 6.8), suggesting that the observed variability is likely more attributable to scanning settings. IC98 highlights the ventricle region, making the observed SZ effect questionable. IC7 exhibits spatial overlapping with IC9 identified from the BIG data in brainstem and cerebellum regions. Since IC9 exhibits the most significant scanning effect in the BIG data, suggesting a high susceptibility to scanning settings in the region, we incline towards IC7 more likely representing scanning effects and decide to have it corrected in this study.

After SBM-correction, no significant scanning effects are observed and the SZ effect is refined when evaluated with ICA, as shown in Table 6.8a. The most significant SZ effect is observed from IC81, which is spatially consistent with those identified in the uncorrected (IC94) and GLM-corrected (IC12) data. The component suggests SZ-related gray matter reduction in a frontal-temporal network, one of the most consistently identified structural variations in SZ (Cannon et al., 2002; Glahn et al., 2008; Kuperberg et al., 2003; Turner et al., 2012; Xu et al., 2009). The SBM- and GLM-corrected data present more significant SZ-effects in the frontal-temporal network compared to the uncorrected data, suggesting that correcting for confounding effects helps refine the true effect of interest. Meanwhile, the GLM-corrected data also presents other marginal effects, as illustrated in Fig 6b. It can be seen that IC15 spatially overlaps with IC7-uncorrected which has been identified as scanning-related and eliminated in SBM-correction. IC34 and IC91 reflect sparse and edge effects. In contrast, the SBM-corrected data presents one single frontal-temporal network exhibiting the most significant SZ effect (p = 5.55×10^{-7}), suggesting that the data is more effectively corrected.

The evaluation with VBM echoes the results obtained with the ICA approach. Both SBM- and GLM-correction effectively eliminate the scanning effects. On the other hand, the uncorrected data present more SZ-related voxels than the GLM- and SBM-corrected data, as illustrated in Table 6.8b. For the GLM-corrected data, the identified 228 voxels are a subset of the 421 voxels identified in the uncorrected data. The difference is believed to result from the collinearity among regressors. It is expected that the scanning parameters capture more of the shared variance when they are modeled alone in the GLM-correction, leaving less variance for the SZ diagnosis in the subsequent VBM analysis. For the SBM-corrected data, a side-by-side slice view (Figure 6.7) illustrates that much fewer voxels are identified in brainstem and cerebellum regions. Clearly, these voxels are largely captured in the uncorrected data by IC7, which is subsequently eliminated in the SBM-correction. Therefore, no significant effects are expected in this region.

The comparison between GLM and SBM correction demonstrates that the former is more model-based while the latter more data-driven. In a GLM model, all the variances that can be explained by the predictors are regressed out. As shown in Table 6.8a, the GLM approach seems eliminating effects from scanner, field strength and slice thickness more completely than the SBM approach (p-values are higher than those obtained from the SBM approach although both are not significant). However, one concern is that the GLM model is not able to deal with embedded collinearity, such that variance shared between scanning parameters and traits of interest may also be eliminated. In contrast, in the SBM model, ICA is able to extract components attributed to different sources. This particularly manifests in the observation that the SZ effect is split into independent components (IC94, IC98, IC53 and IC7 in the uncorrected data, see Table 6.8a). While the GLM approach cannot separate the SZ-related voxels from each other, ICA acknowledges that they covary with different underlying patterns and extracts meaningful components. Subsequent analyses suggest that IC94 shows no scanning effects, while others are likely confounded. Thus, ICA-based SBM model enables the researchers to recognize the heterogeneity and allows flexibility on whether a correction is necessary for each component.

In summary, our study explores scanning effects in multi-site sMRI studies and demonstrates an effective approach for correction. The results suggest that magnetic field strength and sequence design (including repetition time, inversion time) are likely the most significant contributors to the image variability, although the individual effects could not be disentangled in the present data and await more investigations. Another significant confounder highlighted is RF receiving coil, which needs to be considered in the current atmosphere of data sharing and aggregation for large-scale analyses. The second study demonstrates that scanning effects can be isolated from the disease effect through SBM approach, and a correction could be further applied to refine the true effect of interest. Overall, consistent field strength, sequence design and RF coil are strongly recommended for multi-site sMRI studies, though SBM proves a flexible and effective approach to detect and clean scanning effects, which helps reduce the risk of false positives.

CHAPTER 7 GUIDED EXPLORATION OF GENOMIC RISK FOR GRAY MATTER ABNORMALITIES IN SCHZIOHRENIA

7. 1 Introduction

Schizophrenia (SZ) is a severe psychiatric disorder demonstrating a strong genetic component with heritability estimated up to 80% based on family and twin studies (Cardno and Gottesman, 2000). In the past decade, a number of susceptibility genes have been identified from linkage and association studies (Duan et al., 2010; Harrison and Owen, 2003). However, the associations between SZ diagnosis and individual polymorphisms were often weak (Duan et al., 2010). These results suggest a polygenic model for SZ (Gottesman and Shields, 1967), hypothesizing that multiple alleles with small individual effects may interact synergistically to increase the susceptibility to the disorder. The hypothesis is supported by a recent study, demonstrating the involvement of an aggregate of common (frequency > 0.05) single nucleotide polymorphisms (SNPs), collectively accounting for a substantial proportion of variation in risk to the disorder (Purcell et al., 2009b).

In response to the complex genetic architecture underlying SZ, a multivariate approach is well positioned to examine associations between SZ-related phenotypes and genetic components derived from various potential susceptibility alleles. Prata et al., examining epistasis between DAT and COMT genes, demonstrated that the nonadditive DAT-COMT interaction was associated with a SZ-altered modulation effect on cortical function during executive processing (Prata et al., 2009). Expanding the variables to 24

SNPs spanning 14 SZ putative risk genes, Meda et al. identified two genetic components (DAT and BDNF; SLC6A4, 5HTTLPR and 5HTTLPR_AG) correlating with three functional brain networks; the combined brain function-gene effects showed significant group differences in SZ (Meda et al., 2010). These positive findings have encouraged researchers to explore more genetic influences, including interactions among known risk genes and novel genes.

Simultaneously, an allied line of research has focused on defining endophenotypes given the heterogeneity of symptoms, course and outcome in SZ (Gottesman and Gould, 2003). Some identified endophenotypes are based on magnetic resonance imaging (MRI), which has demonstrated its specific value in identifying regional brain abnormalities (Rapoport et al., 2005), including structural endophenotypes (Lawrie et al., 2003; McDonald et al., 2006; Nelson et al., 1998; Sun et al., 2009) and functional networks recorded in fMRI that discriminate SZ patients from healthy controls.

In this work, we performed a guided exploration of genomic risk for SZ-related gray matter abnormalities using the proposed pICA-R approach. Whole-brain gray matter concentration images were analyzed in conjunction with genome-wide single nucleotide polymorphisms (SNPs) to investigate the genetic factors possibly underlying the regional variations in brain structure which relates to clinical manifestations. In particular, a genetic reference was derived from a previous SZ genome-wide association study with the largest sample size to guide the data decomposition such that reliable genetic components emphasizing specific biological mechanisms can be extracted.

7. 2 Materials and Methods

Participants

Structural MRI and SNP data were obtained from The Mind Clinical Imaging Consortium (MCIC), a collaborative effort of four research teams from University of New Mexico-Mind Research Network, Massachusetts General Hospital, University of Minnesota and University of Iowa) and from a local COBRE (Center of Biomedical Research Excellence) study. The institutional review board at each site approved the study and all participants provided written informed consents. All healthy participants were screened to ensure that they were free of any medical, neurological or psychiatric illnesses, including any history of substance abuse. The inclusion criteria for patients were based on a diagnosis of schizophrenia, schizophreniform or schizoaffective disorder confirmed by the Structured Clinical Interview for DSM-IV-TR disorders (SCID, (First et al., 1997)) or the Comprehensive Assessment of Symptoms and History (CASH, (Andreasen et al., 1992)). After preprocessing, we obtained a total of 300 participants (160 healthy controls and 140 SZ patients) for which both sMRI and SNP data were collected. Table 7.1 provides the demographic information.

Data Collection and Preprocessing

The brain images were T_1-weighted MRIs collected from 440 participants at multiple sites. 1.5T scanners were used at Massachusetts General Hospital (Siemens), University of New Mexico (Siemens) and University of Iowa (GE), and a 3T scanner was used at University of Minnesota (Siemens). Imaging parameters for the scans at MGH and New

Mexico were TR/TE = 12/4.76ms, slice thickness = 1.5mm, bandwidth = 110Hz, voxelsize

Table 7.1: Demographic information of participants.

Demographics		**SZ (140)**	**HC (160)**	**P-value**
Sex	Male	106	104	0.04
	Female	34	56	
Age	Mean ± SD	36 ± 12	33 ± 11	0.03
	Range	18 - 63	18 - 63	
Race/Ethnicity	Caucasian	109	140	0.19
	African American	20	8	
	Asian	5	5	
	Native Hawaiian	1	0	
	American Indian	1	2	
	Unreported	4	5	
Collection site	Harvard	28	24	0.85
	Iowa	32	59	
	Minnesota	30	19	
	New Mexico	50	58	

= 0.625×0.625×1.5mm. At Iowa the parameters were TR/TE = 20/6ms, slice thickness = 1.6mm, bandwidth = 122Hz, and voxel size = 0.664×0.664×1.6mm. At Minnesota the parameters were TR/TE=2530/3.81ms, slice thickness = 1.5mm, bandwidth = 110Hz, voxel size = 0.625×0.625×1.5mm (Segall et al., 2009). All scans were collected in a coronal orientation.

The T_1-weighted sMRI data were preprocessed in Statistical Parametric Mapping 5 (SPM5, http://www.fil.ion.ucl.ac.uk/spm) using voxel based morphometry (VBM) (Ashburner and Friston, 2005), a unified model where image registration, bias correction and tissue classification are integrated. Brains were segmented into gray matter, white matter and cerebrospinal fluid based on unmodulated normalized parameters. The resulting gray matter images consisted of voxelwise gray matter concentrations. Images were re-sliced to 2 × 2 × 2 mm, resulting in 91 × 109 × 91 voxels. The gray matter images were then smoothed with 10mm full width at half-maximum Gaussian kernel. In the subsequent quality check, we further excluded two participants whose images were four standard deviations away from the average gray matter image. A mask was then generated to include only the voxels inside the brain as well as exhibiting an average gray matter concentration greater than 0.1, resulting in a total of 253,632 voxels. Finally, a voxel-wise regression analysis was performed at each voxel to eliminate the effects from age, sex and collection site. The gray matter images corrected for the above variables were then analyzed in conjunction with the SNP data.

DNA was extracted from blood samples of 255 MCIC participants and saliva samples of 84 COBRE participants respectively (six participants appeared in both studies). Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium HumanOmni1-Quad assay spanning 1,140,419 SNP loci. BeadStudio was used to make the final genotype calls. No significant difference was observed in genotyping call rates between blood and saliva samples.

PLink (Purcell et al., 2007a) was used to perform a series of quality control procedures. Gender was imputed based on x-chromosome heterozygosity rates and checked against internal QC files; SNPs and participants were checked for a genotyping rate of less than 90%; SNPs were excluded if they showed deviation from Hardy-Weinberg Equilibrium in controls with a threshold of 10^{-6} or if they failed to be missing at random with a threshold of 10^{-10}; 4 participant was excluded due to relatedness with an identity-by-descent value > 0.1875; 2 participants were also excluded with heterozygosity ratios 3-SD away from the average; minor allele frequency cut-off was set to 0.01. Discrete numbers were then assigned to the categorical genotypes: 0 for no minor allele, 1 for one minor allele and 2 for two minor alleles. Subsequently, we replaced the missing genotypes using haplotype genotypes of high linkage disequilibrium loci if available (LD, correlation > 0.80). After the above procedures, 777,365 autosomal SNPs were retained for MCIC data and 823,733 autosomal SNPs were retained for COBRE data, resulting in the final dataset of 327 samples × 728,683 common SNPs. It was noted that the minor allele differed between MCIC and COBRE data for 23,716 SNPs. We then adjusted the minor allele codings of these 23,716 SNPs in COBRE data to be consistent with those in MCIC data. Finally, as we decided to admit participants from all ethnic groups, population stratification was investigated through PCA (Price et al., 2006). Specifically, the SNP data were decomposed into a linear combination of underlying components, four of which differed significantly among ethnicities ($p = 2.40\times10^{-99}$, 1.51×10^{-85}, 1.25×10^{-30}, 1.28×10^{-18}, respectively) while not differentiating between schizophrenia patients and healthy controls, suggesting weak associations with the disorder. Therefore, we eliminated these four components from the original data. Afterwards, a Q-Q plot (Chanock et al., 2007) for p-values of group differences between patients and controls tested against a uniform distribution showed no clear indication of population structure (Figure 7.1).

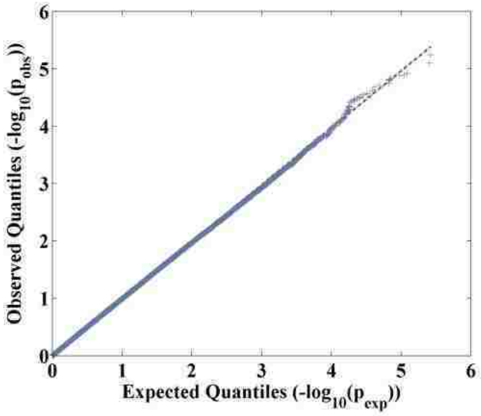

Figure 7.1: Q-Q plot of p-values (two sample t-test, group difference between patients and controls in terms of MAF) tested against a uniform distribution.

Association Analysis

Parallel ICA with reference (pICA-R) was used for the association analysis. Regarding the genetic reference, we leveraged the results from an independent genome-wide SZ study to obtain genetic references. First, we selected a potential susceptibility gene ANK3 with intragenic SNPs exhibiting top genome-wide associations in the Psychiatric Genomics Consortium (PGC) SZ study ((Ripke et al., 2011), Table S10), which is currently the SZ study with the largest sample size. This gene is involved in neuronal activities (Lambert et al., 1997; Zhou et al., 1998) and therefore poses a promising candidate to be a reference in this imaging genetics study. We then identified the corresponding SNPs in ANK3 and grouped neighboring SNPs with moderate LD ($|r| > 0.5$) into a cluster, which could serve as a reference set. The LD threshold was determined by a visual inspection of our data, while also considering that SNPs with $r^2 > 0.2$ are not considered independent (Ripke et al., 2011). For this proof-of-principle and method development study, our primary strategy for reference selection was that, in pICA-R, the reference loci are expected to contribute simultaneously to one single component, which is the case most likely to happen for SNPs in LD. Therefore, we chose to use LD clusters as references to elicit more SNPs contributing in a coordinated manner. Finally we tested three reference sets from ANK3, each spanning more than 40 SNPs, which were to yield at least 20 true loci with an accuracy of 0.5, a reasonable size as observed in simulations. It should be noted that we only examined a limited number of references in this work, as the major purpose was to demonstrate an application of the proposed approach instead of performing a complete SZ study. While there are also other genes that are of great importance, they will be left for future investigations.

For the purpose of validating our finding, the SNP component identified by pICA-R was examined for its SZ enrichment based on the independent results of the PGC SZ study (Ripke et al., 2011). We first selected out SNPs significantly contributing to the identified component. Next, we compared the ratios of SZ-related SNPs in the selected top contributing SNPs and in the whole genome. For each SNP, the SZ-relevance was determined based on the significance of association reported in the PGC SZ study, such that a SNP exhibiting SZ association with a p-value less than P_{th} was considered as SZ-related. To examine the enrichment across different significance levels, we tested a P_{th} range from the standard level of 0.05 to a more significant level of 0.001. Then based on this criterion of SZ-relevance, we performed Fisher's exact test to evaluate the significance of SZ enrichment in our finding compared to the whole genome.

In addition, we applied ICA, pICA and ICA-R to the sMRI-SNP dataset for a comparison. In case of ICA, we applied two separate regular ICAs to the sMRI and SNP data respectively. Then pairwise correlations were calculated based on the loadings. In case of pICA, the dataset were directly analyzed for inter-modality associations. In case of ICA-R, we applied regular ICA to the sMRI data while ICA-R was used to extract the SNP component given the same reference. As in pICA-R, the number of components was selected to be 10 for the sMRI data and 27 for the SNP data, if a component number estimation applied.

7. 3 Results

The number of components was estimated to be 10 on uncorrelated voxels of the sMRI data using minimum description length (MDL) (Rissanen, 1978). For the SNP data, 27 components were extracted based on the metric of consistency (Chen et al., 2012b). We tested the three reference sets generated from ANK3 (Ripke et al., 2011), and one reference set spanning 82 SNPs helped elicit significant inter-modality connection. These 82 SNPs exhibited moderate LD with an average correlation of 0.57 and were separated by an average of 1,276 base pairs. Guided by this reference, pICA-R identified one component pair exhibiting the highest correlation of -0.27 and a p-value of 1.64×10^{-6} (passing Bonferroni correction of 0.05/10/27). After regressing out variables (specifically age, sex, race/ethnicity, collection site and SZ diagnosis for the SNP component; race/ethnicity and SZ diagnosis for the sMRI component), the sMRI-SNP association remained significant, exhibiting a partial correlation of -0.24 ($p = 2.81\times10^{-5}$), as shown in Figure 7.2.

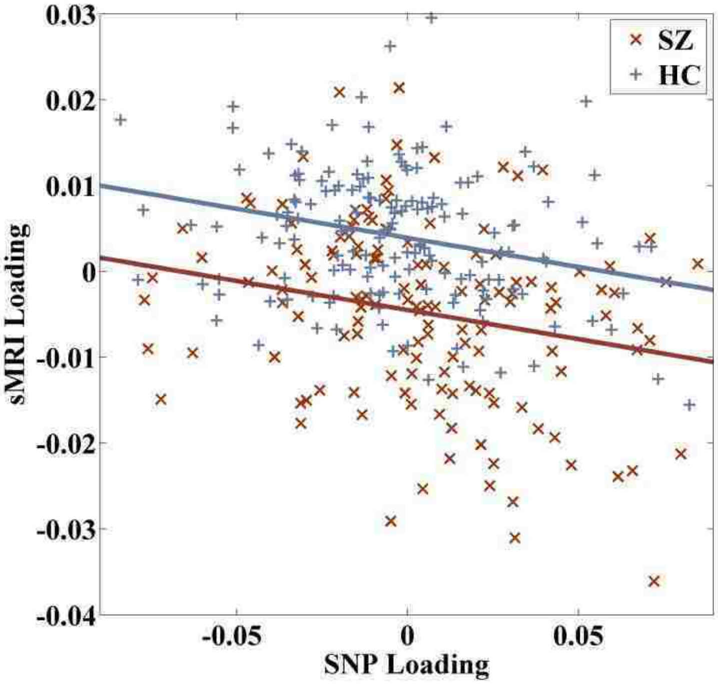

Figure 7.2: Scatter plots of loading coefficients associated with the identified sMRI and SNP components in patient and control group respectively. Controlling variables (age, sex, race/ethnicity, collection site) are corrected.

The loadings of the linked sMRI component significantly differed between SZ patients and healthy controls (two tailed t-test, $p = 1.33 \times 10^{-15}$). Note that effects from age, sex and collection site were already regressed out from the data and we did not observe any significant regression (two tailed t-test, p = 0.11) effect from the race/ethnicity on the sMRI component while controlling for diagnosis. We further examined whether medication affected the identified brain network in patients and found no significant regression effect (two-tailed t-test, p = 0.62) from the reported chlorpromazine equivalent dosage (Gardner et al., 2010) on the sMRI loadings while controlling for race/ethnicity. Figure 7.3 shows the spatial map of the sMRI component thresholded at |Z| > 3. The identified brain network included medial and inferior frontal gyri, superior temporal gyrus, insula and anterior cingulate, as listed in Table 7.2.

Table 7.2: Talairach labels of identified brain regions (|Z| > 3).

Brain region	Brodmann area	L/R volume (cm^3)	L/R random effects, max Z (x,y,z)
Medial Frontal Gyrus	9, 10, 6, 8	3.2/1.4	4.21(0,42,22)/3.98(2,49,10)
Inferior Frontal Gyrus	47, 13	2.6/2.8	5.09(-40,17,-14)/5.67(44,13,-9)
Superior Temporal Gyrus	38, 22, 13	2.3/3.8	4.94(-44,17,-13)/5.54(44,13,-11)
Insula	13, 22, 47	0.4/1.8	3.74(-44,9,-6)/5.28(44,9,-7)
Anterior Cingulate	32, 10	0.7/0.3	4.01(0,49,7)/3.86(2,47,9)

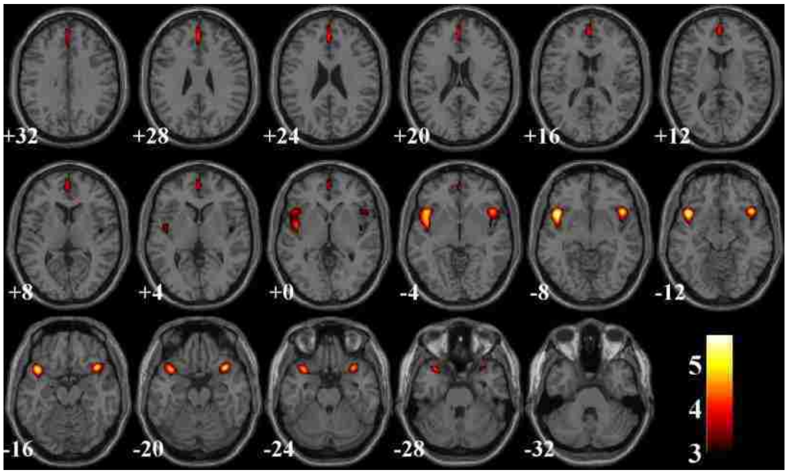

Figure 7.3: Spatial map of brain network for the identified sMRI component (|Z| > 3).

The loadings associated with the linked SNP component exhibited a significant group difference between patients and controls (two tailed t-test, p = 0.04). The SNP component followed a super-Gaussian distribution and Figure 7.4 shows a logistic fit to the histogram. Based on the normalized component weights, we selected out 1,030 top contributing SNPs (top 1,030 based on the absolute values of the normalized component weights, corresponding to |Z| > 3.60, p = 0.003 based on the logistic fit, see Figure 7.5) as our finding.

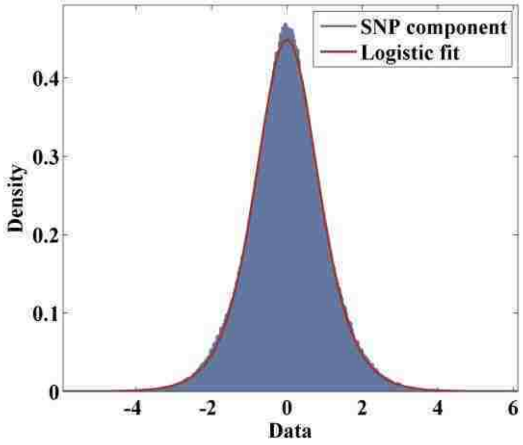

Figure 7.4: Logistic fit to the identified SNP component.

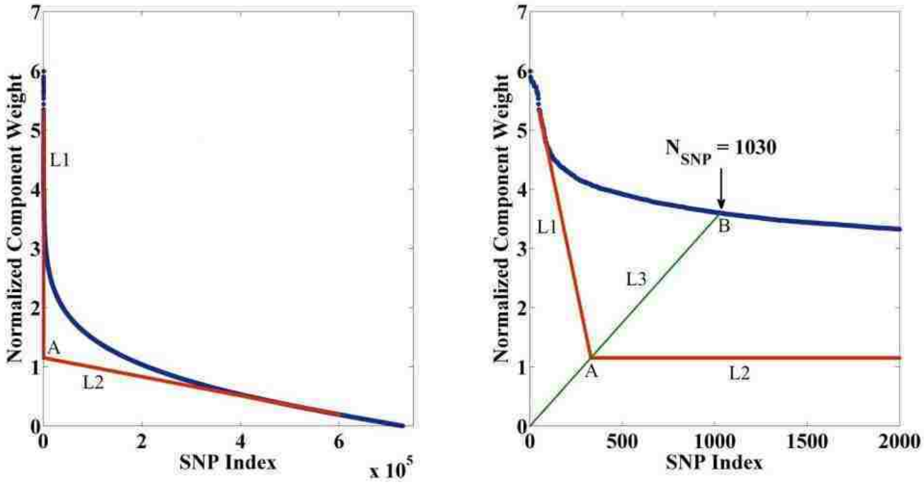

Figure 7.5: Illustration: how the number of top contributing SNPs is determined. The blue curve shows the plot of normalized component weights (descending absolute values); L1 and L2 represent the linear fits to the two segments of the component curve;

A denotes the intersection of L1 and L2; the green line (L3) represents the line connecting the origin and the intersection A; B denotes the intersection of L3 and the component curve.

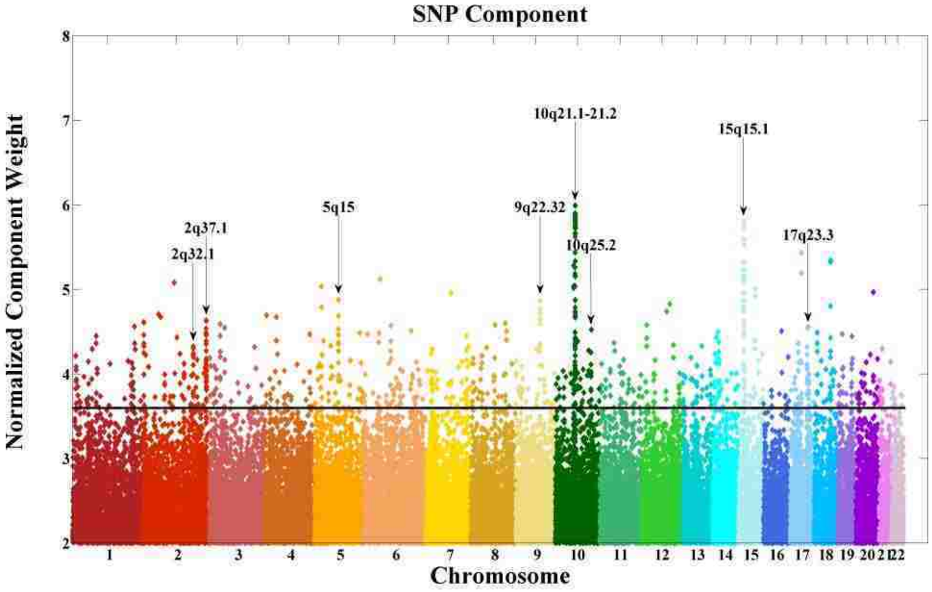

Figure 7.6: Manhattan plot for the identified SNP component (threshold at |Z| > 3.60 for top contributing SNPs).

Figure 7.6 shows a Manhattan plot of weights of loci for the identified SNP component, where clusters spanning more than 10 top contributing SNPs are marked. Table S1 provides a summary of the identified 1,030 SNPs, including SNP position, corresponding gene, normalized component weight, and MAFs in patient and control groups. Fifty-four out of the top 1,030 contributing SNPs were from the reference set and are marked in Table S1. A complete list of the 82 reference SNPs is also provided in Table S2. After these 54 reference SNPs were excluded, 656 out of the remaining 976 SNPs had been investigated in the PGC study for associations with SZ. We then conducted Fisher's exact test on SZ enrichment between these 656 matched SNPs and the whole genome of PGC data (spanning a total of 1,252,901 SNPs). As shown in Figure

7.7, significant SZ enrichment was consistently observed within the entire range of tested P_{th}'s.

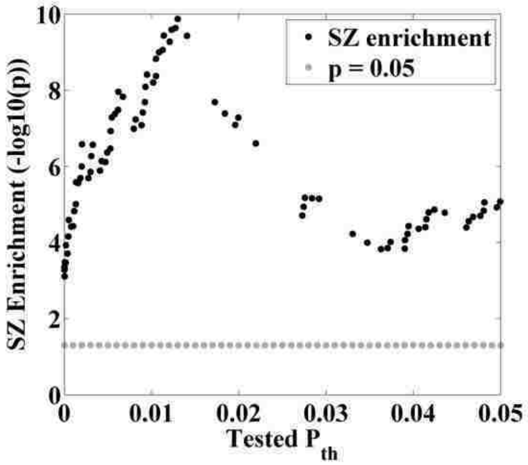

Figure 7.7: Fisher's exact test on SZ enrichment between the identified SNPs and the whole genome based on PGC results. P_{th} denotes the threshold p-value of SZ-relevance, ranging from 0.001 to 0.05.

We further investigated biological functions in which these top contributing SNPs are involved. While 522 out of 1,030 SNPs were mapped to 228 unique genes, Ingenuity Pathway Analysis (IPA: Ingenuity® Systems, http://www.ingenuity.com) indicated a significant enrichment of the domain of central nervous system development ($p = 2.88\times10^{-4}$) in our finding, where 7 genes were involved, as highlighted in Table 7.3a. The identified genes were also significantly overrepresented in glutamate receptor signaling ($p = 2.75\times10^{-2}$) and DARPP32 regulated pathway ($p = 4.07\times10^{-2}$), as well as synaptic long term depression (LTD, $p = 1.58\times10^{-2}$) and potentiation (LTP, $p = 3.24\times10^{-2}$), as highlighted in Table 7.3b. In addition, the DAVID (Database for Annotation, Visualization and Integrated Discovery) bioinformatics resource (Huang et al., 2009a, b) identified significant clusters functionally related to cell adhesion ($p = 1.14\times10^{-5}$), synaptic transmission ($p = 2.86\times10^{-4}$) and neuron projection morphogenesis (p =

1.75×10^{-3}) respectively, as highlighted in Table 7.3c. The identified canonical pathways and functional annotation clusters remained highly stable when the number of top contributing SNPs was adjusted from 1,000 to 5,000, as shown in Table 7.4 and 7.5.

Table 7.3: Biological Pathway analysis and functional annotation clustering.

1a. IPA biological function	Genes	P-value/P-value (B-H)
Coronary disease	ACE, ASIC2, CACNA1C, CERS6, CHRNA5, CSMD1, CSMD2, ITGB2, MECOM, MGAM, PPARA, PTPRM, SAMD12	2.24E-05/1.68E-02
Vascular disease	ACE, ASIC2, CACNA1A, CACNA1C, CERS6, CHRNA5, COL4A1, COL4A2 (includes EG:12827), CSMD1, CSMD2, ITGB2, MECOM, MGAM, PPARA, PTPRM, SAMD12, TEK	8.53E-05/2.25E-02
Aggregation of tumor cell lines	CMIP, DAPK3, IGF1R, ITGB2, PRKD1	9.70E-05/2.25E-02
Coronary artery disease	ASIC2, CACNA1C, CERS6, CSMD1, CSMD2, ITGB2, MECOM, MGAM, PPARA, PTPRM, SAMD12	1.20E-04/2.25E-02
Development of central nervous system	**ADAM22, ASIC2, CNTNAP2, DSCAML1, MYO16, PARK2, ZBTB16**	**2.88E-04/4.31E-02**
Atherosclerosis	ACE, ASIC2, CACNA1C, CERS6, CSMD1, CSMD2, ITGB2, MECOM, MGAM, PPARA, PTPRM, SAMD12	4.26E-04/5.33E-02

1b. IPA Canonical Pathway	Genes	P-value/P-value (B-H)
AMPK Signaling	PFKFB3, AK5, ACACB, PPP2R2C, PFKP, CHRNA5	4.17E-03/7.93E-01
Aldosterone Signaling in Epithelial Cells	DNAJC17, ASIC2, DNAJC18, PLCB1, DNAJC10, PRKD1	9.77E-03/8.08E-01
Synaptic Long Term Depression	**IGF1R, PLCB1, PPP2R2C, GRM4, PRKD1**	**1.58E-02/8.08E-01**
Maturity Onset Diabetes of Young (MODY) Signaling	CACNA1C, CACNA1A	2.04E-02/8.08E-01
Glutamate Receptor Signaling	**SLC1A1, GRM4, GNG2**	**2.75E-02/8.08E-01**
Synaptic Long Term Potentiation	**CACNA1C, PLCB1, GRM4, PRKD1**	**3.24E-02/8.08E-01**
Dopamine-DARPP32 Feedback in cAMP Signaling	**CACNA1C, PLCB1, PPP2R2C, PRKD1, CACNA1A**	**4.07E-02/8.08E-01**
Agrin Interactions at Neuromuscular Junction	ITGB2, NRG3, ARHGEF7	4.37E-02/8.08E-01
G Protein Signaling Mediated by Tubby	PLCB1, GNG2	5.01E-02/8.08E-01
RhoGDI Signaling	CDH12, ARHGEF7, CDH10, GNG2, ARHGAP8/PRR5-ARHGAP8	5.13E-02/8.08E-01

1c. DAVID functional annotation cluster	Genes	P-value/P-value (B-H)
Cell adhesion	PTPRM, CLSTN2, MAGI1, TNC, PCDH9, FBLIM1, DSCAML1, ITGB2, PTPRT, COL5A1, BTBD9, CDH12, SEMA5A, PKP2, TEK, PECAM1, CNTNAP2, RELN, CNTN4, IZUMO1, ADAM22, CDH10	1.14E-05/1.14E-02
Synaptic transmission	**GRM4, ACCN1, DLGAP1, GABRR1, CHRNA5, PARK2, VIPR1, CACNA1C, KCNIP1, RIMS1, SLC1A1, CACNA1A**	**2.86E-04/9.18E-02**
Neuron projection morphogenesis	**SEMA5A, IGF1R, PTPRM, ANK3, DSCAML1, CNTN4, RELN, GAS7, CACNA1A**	**1.75E-03/1.78E-01**

Note: P-value(B-H) represents the Benjamini-Hochberg corrected p-value of enrichment.

Table 7.4: IPA canonical pathways with varying numbers of top contributing SNPs.

IPA canonical pathway	N = 1000	N = 2000	N = 3000	N = 4000	N = 5000
	p/p (B-H)	p/p (B-H)	p/p (B-H)	p/p (B-H)	p/p (B-H)
Synaptic long term depression	1.58E-02/8.08E-01	4.68E-04/1.29E-01	2.57E-04/2.00E-02	9.33E-05/8.71E-03	2.63E-04/2.57E-02
	PRKCQ,GRM8,RYR2,GRM4,GRM7,GRM5,PLCD1,PRKG1,GRID1,PLCG2,PPP2R2B,IGF1R,PLCB1, PRKCE,PPP2R2C,PRKD3,PRKD1,PRKCB				
Glutamate receptor signaling	2.75E-02/8.08E-01	1.10E-02/6.00E-01	1.20E-02/3.51E-01	1.05E-02/1.33E-01	7.94E-03/1.78E-01
	GRM5,GRM7,GRIN2B,GRID1,GRM8,SLC1A1,GRM4,GNG2				
Synaptic long term potentiation	3.24E-02/8.08E-01	6.92E-03/6.00E-01	1.66E-04/1.74E-02	4.17E-05/5.13E-03	1.00E-04/1.32E-02
	GRIN2B,PRKCQ,GRM8,CACNA1C,GRM4,PRKAG1,GRM7,PLCD1,GRM5,PLCG2,PRKAG2,PRKCE, PLCB1,PRKD3,PRKD1,CAMK2B,PRKCB				
Dopamine-ARPP32 feedback in cAMP signaling	4.07E-02/8.08E-01	3.72E-02/7.21E-01	1.07E-03/5.62E-02	5.25E-04/2.45E-02	1.45E-03/6.31E-02
	KCNJ12,GRIN2B,PRKCQ,CACNA1C,CACNA1A,PRKAG1,PLCD1,PRKG1,KCNJ10,PLCG2,PPP2R2B, PRKAG2,PLCB1,PRKCE,PPP2R2C,PRKD3,PRKD1,PRKCB				

Note: *N* represents the number of top contributing SNPs; *p* and *p*(B-H) represent the uncorrected and Benjamini-Hochberg corrected p-values of enrichment.

Table 7.5: DAVID functional annotation clusters with varying numbers of top SNPs.

DAVID annotation cluster	N = 1000	N = 2000	N = 3000	N = 4000	N = 5000
	p/p (B-H)	p/p (B-H)	p/p (B-H)	p/p (B-H)	p/p (B-H)
Cell adhesion	1.14E-05/1.14E-02	4.40E-06/2.41E-03	2.64E-06/5.43E-03	1.73E-06/8.53E-04	1.44E-09/3.77E-06
	NRP2, CLSTN2, OPCML, PCDHGA1, CDH22, WISP1, ROBO1, CNTNAP2, ROBO2, IZUMO1, DSCAM, SYK, PTPRK, PTPRM, MAGI1, CNTN5, TRPM7, LEF1, PCDH9, FBLIM1, PTPRT, MFGE8, PTPRU, NRXN1, CERCAM, BTBD9, SLC26A6, PGM5, RELN, CNTN4, DST, DCHS2, GNE, TNC, DSCAML1, SPOCK1, IL32, ITGB2, SOX9, CDH4, ITGBL1, SEMA5A, FAT3, TNR, COL6A3, TEK, BAI1, CD2, SSX2IP, COL18A1, BMP1, GMDS, ADAM23, COL13A1, LPP, PPFIBP1, CELSR3, ITGA2, PCDH15, COL5A1, CDH12, CDH13, COL19A1, PKP2, PKP3, CDH18, FBLN5, PECAM1, ADAM22, CDH10, NTM, HABP2, MUC16				
Synaptic transmission	2.86E-04/9.18E-02	7.79E-08/1.28E-04	6.31E-06/4.32E-03	8.30E-07/1.02E-03	2.51E-06/5.50E-04
	SYT1, KCNC4, CACNB2, VIPR1, RIMS1, KCNIP1, AMPH, KCNQ5, GRIN2B, CHRNA5, SLC1A1, HTR1F, CHRNA3, KCNMA1, DLGAP1, GABRG3, DLGAP2, NRXN1, PARK2, PARK7, GRM5, GRM4, ACCN1, GABRR1, NPY, PNOC, GRM8, GRM7, KCNN3, SLC6A5, GHRL, UNC13C, CACNA1C, UNC13B, CACNA1A				
Neuron projection or development	1.75E-03/1.78E-01	7.99E-05/1.07E-02	2.03E-05/8.34E-03	1.64E-06/1.01E-03	3.69E-07/1.61E-04
	DCC, NRP2, NRTN, OPCML, DSCAML1, PRKG1, CDH4, SEMA5A, ATP2B2, IGF1R, ROBO1, ANK3, TNR, NUMB, BAI1, CNTNAP2, ROBO2, B3GNT2, DSCAM, KLF7, PTPRM, SPTBN4, NTNG1, CELSR3, NRXN1, LMX1A, FIG4, GAS7, TP73, SLIT3, SLC26A6, CLIC5, MAP2, GHRL, RELN, CNTN4, EFNA5, DST, NTM, CACNA1A, GFRA3				

Note: *N* represents the number of top contributing SNPs; *p* and *p*(B-H) represent the uncorrected and Benjamini-Hochberg corrected p-values of enrichment.

Compared to pICA-R, ICA did not identify any significant association between the two modalities (r = -0.16, p = 6.68×10^{-3}). pICA identified a significant sMRI-SNP association (r = -0.24, p = 3.13×10^{-5}). After regressing out possibly confounding variables (age, sex, collection site, ethnicity and diagnosis), the partial correlation was -0.18 (p = 1.40×10^{-3}). Using the same reference set as in pICA-R, ICA-R did not identify any significant sMRI-SNP association (r = -0.13, p = 0.03). Given that pICA identified a significant association, we then performed a cross validation for that SNP component based on the PGC SZ study, and did not observe any significant SZ enrichment compared to the whole genome (Figure 7.8). Thus pICA-R was the only approach that identified a significant sMRI-SNP association and also showed significant SZ enrichment in our validation study.

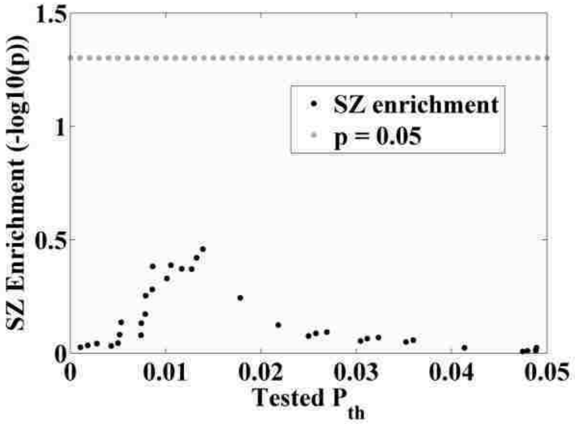

Figure 7.8: Fisher's exact test on SZ enrichment between the SNPs identified by pICA and the whole genome based on PGC results. p_{th} denotes the threshold p-value of SZ-relevance, ranging from 0.001 to 0.05.

7. 4 Discussion

Given a sample-to-SNP ratio around 4.12×10^{-4}, pICA-R identified one sMRI-SNP component pair exhibiting a significant association (r = 0.24, p = 2.81×10^{-5}) while controlling for age, sex, race/ethnicity, collection site and SZ diagnosis, indicating that the association was not mainly attributable to these factors. The loadings associated with the SNP component differentiated patients from healthy controls ($p = 0.04$), while the sMRI loadings showed a more significant group difference ($p = 1.33\times10^{-15}$). Overall, the results suggest that the identified genetic factor might underlie a proportion of variation in gray matter concentration that further contributes to SZ phenotypic symptoms (Harrison, 1999).

sMRI component: The loadings associated with the sMRI component were significantly lower in patients, indicating an overall SZ-related loss of gray matter, which has been indicated in a number of studies (Glahn et al., 2008; Gur et al., 2007; Narr et al., 2005; van Haren et al., 2007). The identified brain network consisted of dorsolateral (Brodmann Areas (BA) 9) and ventrolateral (BA6 and 47) prefrontal cortices (DLPFC and VLPFC), as well as anterior cingulate (BA32) and insular cortex (BA13). This network overlaps considerably with an SMRI component identified before in these data, and found to be heritable in a sibling-pair analysis (Turner et al., 2012). DLPFC is connected to a variety of brain areas and plays an important role in working memory (WM), executive function and other higher-order cognitive processes. Recent work also lends support for DLPFC contributing to the encoding of relational memory, which may further promote long-term memory (LTM) formation, through its role in WM organization (Blumenfeld et al., 2011; Murray and Ranganath, 2007). VLPFC, compared with DLPFC, is generally considered as involved in LTM formation, where the left frontal region is more associated with verbal memory while the non-verbal memory activates more of the right frontal region (Buckner et al., 1999). The anterior cingulate

(BA32) consists of affective and cognitive subdivisions, the former more associated with emotional processes and the latter more activated by tasks requiring cognitive and attentional control (Davidson et al., 2002; Pizzagalli, 2011). The above highlighted regions have been consistently reported to be altered in SZ patients, including reductions in gray matter and cortical thickness (Cannon et al., 2002; Glahn et al., 2008; Kuperberg et al., 2003; Shenton et al., 2001; Xu et al., 2008), as well as exhibiting abnormal task-related functional activation (Glahn et al., 2005; Manoach, 2002; Minzenberg et al., 2009). Overall, our findings are in line with a considerable evidence of gray matter abnormalities in prefrontal and temporal regions as one of the characteristic deficits in SZ.

SNP reference: In this work, we adopted the most straightforward strategy to generate a reference set based on LD clusters of one single gene (ANK3). Genome-wide association study (GWAS) is based on the premise that a causal variant is located on a haplotype, and thus a marker allele in LD with the causal variant should show (by proxy) an association with the trait of interest (Stranger et al., 2011). Therefore, SNPs in one LD cluster are more likely to contribute simultaneously to one single component and serve as good candidates for reference.

Although the SNP highlighted in the PGC study (rs10994359 from ANK3) is not covered in our data, the nearest SNP (rs10761503, 307bp upstream, in LD with rs10994359 with a D-prime of 1 according to the HapMap CEU LD data) is in moderate LD with the reference set (exhibiting a mean correlation of 0.43). In addition, we mapped the selected reference SNPs to the PGC data. 18 out of the 82 reference SNPs were investigated in the PGC study, and 12 were implicated for SZ relevance ($p < 0.05$), leading to a true causal loci ratio of 0.67 (12/18). Given that the 18 PGC-mapped SNPs were uniformly distributed along the 82 reference SNPs, this ratio of 0.67 provided a reference for estimating the number of true casual loci in our reference set, which should be about 55 (0.67*82). In fact, our results did echo this true causal loci ratio, where 54 out of the 82 reference SNPs were identified as top-contributing. The 54 identified SNPs included 9 PGC-implicated causal loci, and the remaining 45 SNPs demonstrated very high LD with the PGC findings. According to the HapMap CEU LD data, 16 SNPs are in complete LD with the 12 PGC-implicated SNPs (D-prime = 1), and another 4 demonstrate a D-prime of 0.871, 0.875, 0.939 and 0.883, respectively. For other 25 SNPs not covered in the HapMap CEU LD data, we evaluated in our data their relations with the 12 PGC-implicated SNPs and found high correlations ($r > 0.96$) except for one locus. These observations suggest that LD can provide good guidance in reference selection. When limited causal loci are known, searching clusters of SNPs exhibiting LD with them may be the most effective approach to generate a testable reference in this pICA-R model.

SNP component: The SNP component was significantly associated with the sMRI component. On average, SZ patients carried higher loadings on the SNP component while exhibiting lower gray matter concentration in the identified regions of the sMRI component. The SNP component was predominantly contributed to by 1,030 SNPs exhibiting top component weights. Cross-evaluation based on PGC results confirmed that the top contributing SNPs were significantly overrepresented in terms of SZ-relevance, which validated our finding. It is noted that when the threshold of SZ-relevance (P_{th}) increased, the enrichment diminished, which is reasonable. The top contributing SNPs comprised a number of clusters distributed across the whole genome, which is not surprising given our model, where SNPs in LDs would exhibit comparable effects. Clusters spanning more than 10 top contributing SNPs are highlighted in Figure 7.6 and marked by the corresponding cytogenetic bands, some of which have been implicated in previous studies, such as 5q15 for bipolar disorder (Scott et al., 2009), 15q15.1 for attention deficit/hyperactivity disorder (Bakker et al., 2003), as well as 17q23.3 for autism (Girirajan et al., 2011) and schizophrenia (Wahlbeck et al., 2000).

Among the 1,030 top contributing SNPs, 522 reside in a total of 228 unique genes. The remaining 508 intergenic SNPs lie within sequences not presently annotated but they could have a regulatory function on large non-coding RNAs and other regulatory non-coding RNAs.

Pathway analyses of the 228 known genes revealed that they participate in a number of neurotransmitter and nervous signaling pathways, including glutamate receptor signaling and DARPP32 regulated pathway, as well as synaptic LTP and LTD. It was noted that some pathways and clusters were no longer significant after the Benjamini–Hochberg correction; however this does not necessarily indicate a false positive finding. First, the correction was performed for all candidate pathways, which may not be independent from each other, indicating a possibility of over-correction. Second, the identified canonical pathways and functional annotation clusters remained highly stable when we adjusted the number of top contributing SNPs from 1,000 to 5,000. In particular, the enrichment became significant even after the correction at some point (Table 7.4 and 7.5). Finally, as emphasized by IPA, the enrichment score simply provides guidance for interpretation, and it is more important to further explore the functions of involved genes to interpret the finding. In this study, the pathway analyses results are provided to help unravel the genetic architecture. The involved genes are discussed in more details to understand the biological connections between the identified component and the disorder.

Glutamate receptor signaling (SLC1A1, GRM4, GNG2): Glutamate receptor signaling plays a crucial role in neurocognitive processes and aberrant glutamate neurotransmission may be associated with positive and negative symptoms as well as cognitive deficits in SZ (Coyle, 2006; Egan et al., 2004; Krystal et al., 2010). Recent work also provides evidence for an association between perturbed glutamate function and gray matter volume variation in prodromal SZ (Stone et al., 2009). In particular, one SNP in GNG2 (encoding guanine nucleotide-binding protein, gamma-2**)** has been identified, with its minor allele relating to an increased gray matter volume in medial prefrontal cortex (Chavarria-Siles et al., 2012). Also, some glutamate transporters including SLC1A1 (encoding excitatory amino-acid transporter 3) are believed to have pivotal functions in mediating neurotoxicity, which raises the possibility of underlying structural changes in SZ (Deutsch et al., 2001; Olney and Farber, 1995). In our finding, three SNPs contributed to the glutamate pathway, including rs2150195_A (SLC1A1, ‘A’ denotes the minor allele), rs1873249_G (GRM4) and rs10150721_G (GNG2). The first SNP contributed with a positive weight, indicating an increased MAF being associated with lower gray matter concentration; and the latter two SNPs presented negative weights, implying gray matter loss being associated with decreased MAFs.

Dopamine-DARPP32 signaling (CACNA1A, CACNA1C, PLCB1, PPP2R2C, PRKD1): These proteins modulate dopamine and DARPP32 regulated gene expression and function, which likely influences synaptic plasticity such as LTP and LTD (Jay, 2003; Svenningsson et al., 2004) as well as being associated with SZ risk (Albert et al., 2002; Howes and Kapur, 2009). In our finding, five genes are involved in this pathway, including CACNA1A (rs4926278_C and rs4926279_C), CACNA1C (rs2238070_T), PLCB1 (rs2745764_T), PPP2R2C (rs7688267_G) and PRKD1 (rs12883327_T). CACNA1C is likely a major risk gene for bipolar disorder (Ferreira et al., 2008). Meanwhile, it is of particulate interest that CACNA1A and CACNA1C (calcium channels, voltage-dependent) also participate in calcium signaling, which plays an important role in neuronal processes (Lidow, 2003; Mattson, 1992) and may also contribute to the reduction in neuronal number given its suggested role in cell death (Sastry and Rao, 2000; Toescu, 1998).

Synaptic LTP and LTD (IGF1R, PLCB1, PPP2R2C, GRM4, PRKD1, CACNA1C): synaptic LTP and LTD are two forms of synaptic plasticity resulting in altered synaptic strength, which underlie learning and memory (Collingridge et al., 2010; Cooke and Bliss, 2006; Linden and Connor, 1995). While learning and memory impairments are well documented in SZ (Aleman et al., 1999; Paulsen et al., 1995), direct evidence has also been provided for disrupted LTP/LTD in SZ (Frantseva et al., 2008; Weng et al., 2011). In our finding, three genes are involved in both LTD and LTP processes, including PLCB1, GRM4 and PRKD1. GRM4 (encoding metabotropic glutamate receptor 4) is also implicated in glutamate signaling, while PLCB1 (1-phosphatidylinositol 4, 5-bisphosphate phosphodiesterase beta-1), PRKD1 (Serine/threonine-protein kinase D1**)** and PPP2R2C (Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B gamma isoform) are also implicated in DARPP32 regulated pathway, indicating a possible convergence in pathology. On the other hand, IGF1R (Insulin-like growth factor 1 receptor**,** rs8038015_C and rs6598542_G) is involved only in LTD, where both of two SNPs contributed with positive weights.

Besides those genes implicated in the aforementioned neurotransmitter and nervous signaling pathways, it is noteworthy that a number of the remaining detected genes have been implicated in other neuronal processes. For instance, CNTNAP2 (encoding contactin-associated protein-like 2) and RELN (encoding reelin), as reported by DAVID, are among the functional cluster of cell adhesion, which plays an important role in brain development (Edelman, 1983; Rutishauser and Jessell, 1988). CNTNAP2 is shown to mediate intercellular interactions during latter phases of neuroblast migration and laminar organization (Strauss et al., 2006). This gene exhibits a high expression in anterior temporal and prefrontal regions in humans, yet low or absent expression in rodents (Abrahams et al., 2007), suggesting a possible role in higher cognitive functions such as language (Vernes et al., 2008). RELN is suggested to regulate neurogenesis and migration, as well as enhance synaptic LTP (Hoe et al., 2009; Pujadas et al., 2010; Spalice et al., 2009). In addition, RELN mutations have been associated with SZ (Guidotti and Di-Giorgi-Gerevini, 2002; Wedenoja et al., 2008).

It's noted that IPA indicates an enrichment of coronary artery and vascular disease in the identified component, as shown in Table 7.3a. While comorbidity between these diseases and SZ has been documented, most of the previous works highlighted environmental factors, such as cigarette smoking and metabolic syndrome (Hennekens et al., 2005; Jeste et al., 1996). This issue may deserve further investigation.

Combining the sMRI and SNP findings, pICA-R revealed an association between one genetic component and SZ-related reduction in gray matter concentration in distributed brain regions. The identified brain regions are among those shown to exhibit gray matter deficits partly attributable to genetic factors (Cannon et al., 2002; Thompson et al., 2001). The genetic component reflects enrichment in neuronal processes. It is noteworthy that both genetic and imaging findings show a particular relevance to cognition, especially memory function. While the underlying mechanism remains to be elucidated, our finding strongly suggests that the identified genetic component may affect neurobiological conditions that play a role in the cognitive deficits of SZ.

CHAPTER 8 CREB-BDNF GUIDED EXPLORATION OF GENOMIC RISK FOR CUE-ELICITED HYPERACTIVATION IN ALCOHOL DEPENDENCE

8. 1 Introduction

Alcohol addiction is a prevalent devastating disease. It is estimated that ~14% of alcohol users experience dependence, presenting a substantial healthy and economic issue (Grant et al., 2001). Genetic factors have been shown to affect liability to alcohol dependence, with the heritability estimated to be 40-80 % while the remainder variances might be majorly attributable to environmental factors (Heath et al., 1997; Knopik et al., 2004; Uhl, 2004). Great efforts have been made towards unraveling the genetic etiology of alcoholism. Targeted gene and large-scale genome-wide studies have provided evidences for a number of susceptibility variants, highlighting genes involved in a variety of neural signaling pathways, including dopaminergic (Conner et al., 2005; Filbey et al., 2008c; Noble, 2000), glutamatergic (Krystal et al., 2003; Mayfield et al., 2008; Schumann et al., 2008) and GABAergic (Bierut et al., 2010; Enoch et al., 2009; Radel et al., 2005) systems. Genes encoding alcohol dehydrogenase (ADH) enzymes playing a key role in alcohol metabolism are also implicated in the vulnerability (Edenberg and Foroud, 2006; Luo et al., 2007).

Despite the growing knowledge on susceptibility loci contributing to the individual differences in drinking behavior, the genetic findings in general suffer small effect sizes.

For instance, in a large genome-wide association study of alcohol dependence where thousands of subjects were included for investigations, no SNP could pass the genome-wide significance threshold of 5×10^{-8} (Stranger et al., 2011). Instead the highlighted 15 SNPs yielded suggestive associations with $p < 10^{-5}$ (Bierut et al., 2010), yet none of them could be replicated in two independence studies with nominal threshold of 0.05, and nor did they replicate findings of a previous GWAS (Treutlein et al., 2010). This is essentially a common challenge in complex genetic trait mapping. Like many other complex disorders, addiction is also suggested as polygenic (Enoch and Goldman, 1999; Goldman, 1993; Johnson et al., 2006), such that the underlying genetic factor involves many loci with small individual effect sizes. The interpretation of genetic effect becomes even more complicated due to heterogeneity with different genetic variants exert influences on phenotypes through different biological mechanisms (Pickens et al., 1991; Wong and Schumann, 2008).

In this work, we employ parallel ICA with multiple references for the investigation of genetic effect on alcohol dependence. The multivariate approach assesses many variables for aggregate effects, posing a promising model for polygenicity. In addition, prior knowledge is incorporated to guide the data decomposition, such that genetic factors of specific attributes can be elicited from high-dimensional complex data. Furthermore, instead of directly linking genetic factors to behavioral assessments, the associations with neurobiological traits are emphasized. Specifically, brain activations were measured from subjects exposed to the taste of alcohol which has been shown to appropriately draw forth altered functions related to alcohol abuse (Myrick et al., 2008; Tapert et al., 2004). Thus, the approach provides a three-way translational framework for exploring the genetic underpinnings of neuronal functions, which might ultimately lead to clinical manifestations of the disorder.

8. 2 Materials and Methods

Participants

A total of 326 subjects participated in the study to investigate genetic and neurobiological traits in heavy drinking (Claus et al., 2011). The institutional review board approved the study. All the participants were recruited from the greater Albuquerque metropolitan region and provided written informed consents. The inclusion criterion was based on alcohol consumption, requiring participants to have at least 5 (for men) or 4 (for women) drinks per drinking occasion at least five times in the past month. The exclusion criteria included a history of severe alcohol withdrawal, brain-related medical problems, or symptoms of psychosis. In addition, participants were required to be sober during the data collection, with a breath alcohol concentration of 0.00. After preprocessing, 315 participants were admitted into the analysis, for which good quality fMRI and SNP data were collected. Table 8.1 provides the demographic information.

Table 8.1: Demographic information of participants.

Number of participants		**Male (220)**	**Female (95)**
Race/Ethnicity	Caucasian	99	43
	African American	4	2
	Asian	2	0
	Latino	54	28
	Native American	13	3
	Mixed	47	19
	Unreported	1	0
Age	Mean ± SD	31.74 ± 9.43	32.52 ± 10.58
	Range	21 – 56	21 – 55

Table 8.2: Alcohol dependence assessment.

Assessment	Sub-category	Description
ADS	ADS-con	Loss of behavior control
	ADS-obs	Obsessive drinking style
	ADS-per	Psychoperceptual withdrawal
	ADS-phy	Psychophysical withdrawal
	ADS-tot	Total ADS
AUDIT	AUDIT-1	How often do you have a drink containing alcohol?
	AUDIT-2	How many drinks do you have on a typical day when you are drinking?
	AUDIT-3	How often do you have 6 or more drinks on one occasion?
	AUDIT-4	How often during the last year have you found that you were unable to stop drinking once you had started?
	AUDIT-5	How often during the last year have you failed to do what was normally expected from you because of drinking?
	AUDIT-6	How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?
	AUDIT-7	How often during the last year have you had a feeling of guilt or remorse after drinking?
	AUDIT-8	How often during the last year have you been unable to remember what happened the night before because you had been drinking?
	AUDIT-9	Have you or someone else been injured as the result of your drinking?
	AUDIT-10	Has a relative, friend, or a doctor or other health worker been concerned about your drinking or suggested you cut down?
	AUDIT-tot	Total AUDIT score
	AUDIT-consump	Alcohol consumption total (sum of AUDIT-1, -2, and -3)
	AUDIT-dep	Alcohol dependence total (sum of AUDIT-4, -5, and -6)
	AUDIT-probs	Alcohol problems total (sum of AUDIT-7, -8, -9 and -10)
ICS	ICS-total	Total ICS
	ICS-ac	Attempted control
	ICS-fc	Failed control
	ICS-pc	Percieved control
Alcohol symptom count	PA-count	Past alcohol abuse symptom count
	CA-count	Current alcohol abuse symptom count
	PD-count	Past alcohol dependence symptom count
	CD-count	Current alcohol dependence symptom count
Drinking history	NewAgeDrink	Probable age that regular drinking first occurred
	NewYearsDrink	Probable number of years of regular drinking
	AgeFirstDrink	Probable age of first drink
	LastDrink	Number of days since last drink from Time-Line Follow-Back
Stress	EStress-tot	Total early stress for ages before 19 (0-18 years old)
	Stress-tot	Early Stress total (all ages reported)
	BDI-tot	Total Beck Depression Inventory

Data Collection and Preprocessing

Behavioral assessment: The assessment was administered through a variety of questionnaires, including the Alcohol Dependence Scale (ADS) (Skinner and Horn, 1984), the Alcohol Use Disorder Identification Test (AUDIT) (Babor et al., 2001) and the

Failed Control Subscale of Impaired Control Scale (ICS) (Heather et al., 1998; Heather et al., 1993). We excluded those relatively incomplete measurements where data were missing for more than 25 subjects. Finally a total of 34 related behavioral measurements were used in the subsequent analysis, as listed in Table 8.2. The missing ratio was no greater than 4/315. It should be noted that most of these behavioral assessments showed significant associations with age, except for AUDIT-9, AgeFirstDrink, LastDrink, EStress-tot and Stress-tot.

Functional MRI: Brain activation data were collected during an alcohol craving task (Claus et al., 2011; Filbey et al., 2008a). Participants were exposed to small amounts of alcoholic (subjects preferred) or juice (litchi) beverages pseudorandomly presented to them during the MRI scans. Figure 8.1 shows the schematic of a single taste cue trial. Each trial sequentially consisted of a 2s “Ready” prompt, a 24s taste cue presentation and a 16s washout period. During the cue presentation, participants tasted the presented beverage (second 1-10 and 12-22) and then swallowed (second 10-12 and 22-24). No stimuli were presented during the washout and participants viewed the word “Rest”. Two 9min runs were conducted for each participant, with a single run spanning 12 trials, 6 for each tastant. A 3T Siemens Trio was used for the data collection. The echo-planar gradient-echo pulse sequence was configured as follows: TR = 2s, TE = 29ms, flip angle = 75°, voxel size = 3.75mm × 3.75mm × 4.55mm. The collected fMRI data were preprocessed with FSL ((Smith et al., 2004), http://www.fmrib.ox.ac.uk/fsl/). Standard motion correction was performed and images were normalized to the Montreal Neurological Institute (MNI) template (Jenkinson et al., 2002). An 8mm full-width half-maximum Gaussian kernel was used for spatial smoothing. Finally alcohol versus juice contrast images spanning a total of 54,937 voxels were extracted for subsequent association analyses.

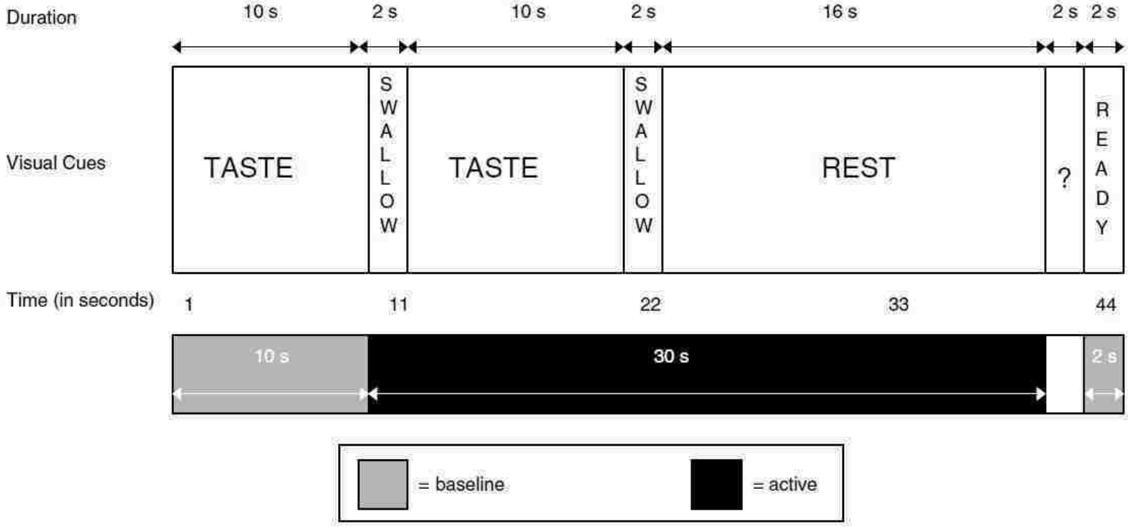

Figure 8.1: Schematic of a single taste cue trial (Filbey et al., 2008a).

SNP data: Saliva samples were collected from participants for DNA extraction. Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium Human 1M-Duo assay spanning 1,199,187 SNP loci. BeadStudio was used to make the final genotype calls. PLink (Purcell et al., 2007a) was used to perform a series of quality control procedures as described in the previous chapter for the MCIC and COBRE data. Specifically, SNPs and participants were first examined for a genotyping rate threshold of 95%; SNPs were excluded if they showed deviation from Hardy-Weinberg Equilibrium with a threshold of 10^{-6} or if they failed to be missing at random with a threshold of 10^{-10}; 2 participants were excluded due to high heterozygosity (3-SD greater than the mean across all subjects); Another 2 participants were excluded due to relatedness with an identity-by-descent value > 0.1875; Minor allele frequency cut-off was set to 0.01. After the standard quality control, discrete numbers were then assigned to the categorical genotypes: 0 for no minor allele, 1 for one minor allele and 2 for two minor alleles. Subsequently, we replaced the missing genotypes using haplotype genotypes of high LD loci if available (correlation > 0.80). After the above procedures, missing genotypes were still observed in 123,159 out of a total of 735,938 autosomal SNPs. We then excluded 18,809 SNPs with a missing ratio greater than 1% and saved the rest through replacing the missing genotypes using the major alleles of individual loci. The resulting data spanning 717,129 autosomal SNPs were then investigated for population stratification using PCA (Price et al., 2006), as we decided to admit participants from all ethnic groups. Specifically, the SNP data were decomposed into a linear combination of underlying PCs, three of which (PC1, 2, and 4) differed significantly among ethnicities (p = 9.85×10^{-79}, 3.23×10^{-86}, 3.21×10^{-55}, respectively) while exhibiting no association with alcohol dependence. These three components were then eliminated from the original data. Afterwards, a Q-Q plot (Chanock et al., 2007) for p-values of AUDIT associations tested against a uniform distribution showed no clear indication of population structure (Figure 8.2).

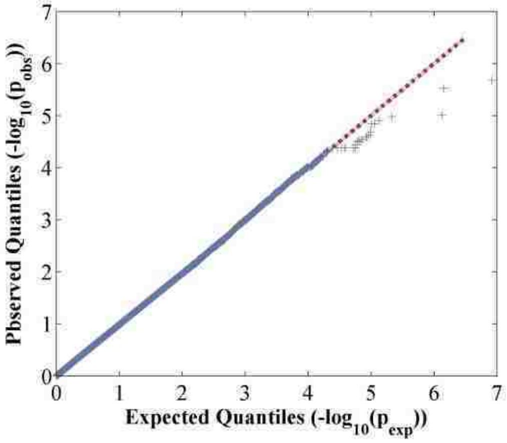

Figure 8.2: Q-Q plot of p-values (correlation test between SNP and AUDIT-tot) tested against a uniform distribution.

Association Analysis

The fMRI contrast images were analyzed in conjunction with the SNP data using parallel ICA with multiple references. The number of fMRI components was estimated by minimum description length (MDL) (Rissanen, 1978) on uncorrelated voxels. The number of SNP components was estimated based on the metric of consistency (Chen et al., 2012b). With respect to genetic references, hotspots or pathways harboring susceptibility genes implicated in previous studies were recruited, as listed in Table 8.3. The gene cluster CHRNA3-CHRNA5-CHRNB4 is the most replicated finding for nicotine dependence (Bierut, 2010; Caporaso et al., 2009) while also implicated in risk of alcohol dependence (Wang et al., 2009). SNPs in 4p12 hosting GABA receptors were shown to present moderate odds ratios in a GWAS on alcohol dependence (Bierut et al., 2010) and independent contributions from individual receptors have been suggested. Another cluster of GABA receptors resides in 5q34. Both mouse and human studies provided evidences for their modulatory roles in alcohol dependence (Radel et al., 2005). Alcohol dehydrogenase (ADH) is a primary enzyme involved in alcohol dependence and variants in the investigated gene cluster, ADH1A-ADH1B-ADH1C have been found associated with risk of alcoholism (Edenberg and Foroud, 2006; Luo et al., 2007). The opioidergic system is considered as mediating drug-induced feelings and playing an important role in substance rewarding properties (Gianoulakis, 2001). Related polymorphisms have been identified as associated with alcohol dependence (Filbey et al., 2008b; Zhang et al., 2008). CREB is a key transcriptional factor for neuronal growth and regulates the expression of BDNF. These two genes have been found to interact in a variety of brain regions and play a critical role in addiction (Carlezon et al., 2005; Crews et al., 2007; Pandey, 2003).

Table 8.3: Tested genetic references.

Reference	Genes	Evidence
15q24-25	CHRNA3, CHRNA5, CHRNB4	Bierut, 2010; Caporaso et al., 2009; Wang et al., 2009
4p12	GABRA4, GABRA2, GABRG1, GABRB1	Bierut et al., 2010; Enoch et al., 2009
5q34	GABRB2, GABRA6, GABRA1, GABRG2	Radel et al., 2005
4q23	ADH1A, ADH1B, ADH1C	Edenberg et al., 2006; Luo et al., 2007
Opioid system	OPRM1, OPRK1, OPRD1	Filbey et al., 2008; Zhang et al., 2008
CREB-BDNF	CREB1, CREB5, BDNF	Carlezon et al., 2005; Crews et al., 2007; Pandey, 2003

Following the design of parallel ICA with multiple references, we tested each of these hotspots or pathways separately, where each produced multiple referential SNP sets. Specifically, for each hotpot or pathway, a reference matrix was generated with each row representing a reference vector highlighting a group of SNP loci selected from a single gene. In this study, most of the referential genes spanned tens of SNPs forming a single LD block, which were directly used to generate a reference vector. One exception was CREB5 hosting 228 SNPs, for which multiple LD blocks were identified (neighboring SNPs with moderate LD |r| > 0.5). The LD threshold was determined by a visual inspection of our data, and also considering that SNPs with $r^2 > 0.2$ are not considered independent (Ripke et al., 2011). The CREB5 referential set was then derived from each LD block and combined with the referential sets derived from BDNF and CREB1 to from the reference matrix. It should be noted that we only examined a limited number of hotspots and pathways in this work. While there are also other pathways such as dopaminergic and glutamatergic systems that are of great importance, they will be left for future investigations.

To assess the fidelity of the identified association, we applied a subset evaluation test. Ten runs, each with 90% of the subjects, were conducted for the evaluation of association, and the fMRI-SNP association identified in the full dataset was tested for its replication in the subset results. More informatively, we performed a permutation test to assess the validity of the identified fMRI-SNP association, that is, to investigate the possibility of the identified association occurring in randomly rearranged subjects. Given the estimated numbers of components, parallel ICA with multiple references extracted 165 fMRI-SNP component pairs in each of the 1,000 permutations. We then calculated the tail probability to evaluate the significance level of the identified fMRI-SNP association based on the top linked component pair of each run.

The identified linked fMRI-SNP components were further investigated for connections with behavioral assessments (Table 8.2) to confirm the functional influences. Partial correlation or regression analysis was conducted to evaluate the association while controlling for sex and race/ethnicity. False discovery rate (FDR) was applied to correct for multiple comparisons given the associations among most of the behavioral measurements. Due to the collinearity, age was not included as a covariate as it is impossible to isolate its contribution from those of phenotypes.

8. 3 Results

The numbers of components was estimated to be 15 and 11 for the fMRI and SNP data, respectively. Among the tested hotspots or pathways listed in Table 8.3, three referential SNP sets derived from the CREB-BDNF pathway were identified as contributing to the same genetic component, which showed a significant correlation with an fMRI component (r = -0.38, p = 3.98×10^{-12}, passing Bonferroni correction of 0.05/15/11). Table 8.4 summarizes the recruited referential loci, which consisted of all the genotyped loci in BDNF (15 SNPs) and CREB1 (20 SNPs), and an LD block spanning 20 SNPs in CREB5. After regressing out controlling variables (age, sex, race/ethnicity), the fMRI-SNP association remained significant, exhibiting a partial correlation of -0.36 (p = 2.98×10^{-11}), as shown in Figure 8.3a.

Table 8.4: Recruited reference SNPs for the CREB-BDNF pathway.

Gene	SNP	Chr	Position
BDNF	rs1519480	11	27632288
	rs7124442	11	27633617
	rs6265	11	27636492
	rs11030104	11	27641093
	rs12291063	11	27650677
	rs11030108	11	27652040
	rs7103411	11	27656701
	rs10835211	11	27657941
	rs1013402	11	27668957
	rs7127507	11	27671460
	rs11030119	11	27684678
	rs2030323	11	27685115
	rs7934165	11	27688559
	rs962369	11	27690996
	rs12273363	11	27701435
CREB1	rs889895	2	208107174
	rs2551640	2	208116138
	rs2551641	2	208118512
	rs2709356	2	208120337
	rs2709357	2	208120777
	rs2551642	2	208121742
	rs11904814	2	208135043
	rs6740584	2	208137596
	rs2551921	2	208143800
	rs2709387	2	208150340
	rs2254137	2	208152273
	rs2551645	2	208159033
	rs2464978	2	208166163
	rs13029936	2	208173789
	rs2551928	2	208174023
	rs1045780	2	208175395
	rs6785	2	208176242
	rs2551929	2	208176717
	rs2256941	2	208177429
	rs2551931	2	208179725
CREB5	rs160337	7	28594314
	rs160338	7	28594727
	rs1008262	7	28602266
	rs310353	7	28603649
	rs310361	7	28607509

rs13233942	7	28607958
rs310338	7	28609570
rs41273	7	28611382
rs1637457	7	28615963
rs17156919	7	28616653
rs41276	7	28616788
rs160375	7	28621226
rs917275	7	28625047
rs160342	7	28630355
rs160343	7	28632292
rs41295	7	28633161
rs160357	7	28635027
rs41298	7	28635305
rs160359	7	28635520
rs41305	7	28640066

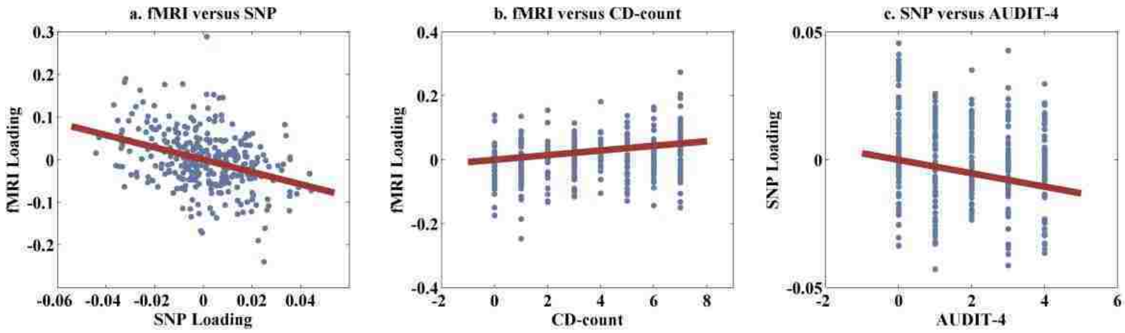

Figure 8.3: Scatter plots of: (a) the fMRI and SNP loadings; (b) the fMRI loading and CD-count; (c) the SNP loading and AUDIT-4.

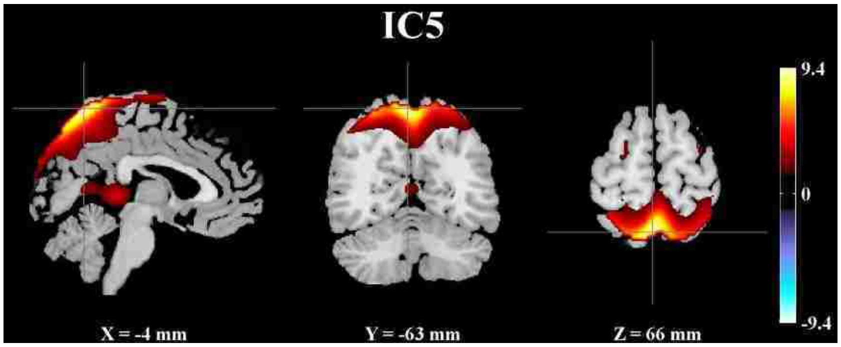

Figure 8.4: Spatial map of brain network for the identified fMRI component (|Z| > 2).

Table 8.5: Talairach labels of identified brain regions (|Z| > 2).

Brain region	**Brodmann area**	**L/R volume (cm^3)**	**L/R random effects, max Z (x,y,z)**
Precuneus	7, 19, 39, 31	16.8/14.1	9.18(0,-58,61)/9.41(3,-58,64)
Superior Parietal Lobule	7, 5	8.9/7.6	8.55(-3,-67,56)/8.72(6,-64,58)
Postcentral Gyrus	7, 5, 3, 2, 40, 1	5.3/4.5	8.06(0,-46,66)/9.03(3,-52,66)
Inferior Parietal Lobule	40, 7, 39	3.5/3.5	4.28(-39,-49,61)/5.11(39,-52,58)
Cuneus	19, 18, 7, 30	2.7/3.7	4.53(0,-82,40)/4.59(27,-83,37)
Paracentral Lobule	5, 4, 6, 7	2.9/1.8	7.28(0,-46,63)/5.71(3,-37,68)
Posterior Cingulate	29, 30, 23	1.6/1.0	3.25(-6,-41,5)/2.99(6,-41,5)

The identified fMRI-SNP pair was replicated in all subset evaluations, where we observed stable fMRI-SNP correlations ranging from 0.23 to 0.33 with a median of 0.27 in the 10 subset evaluations. More importantly, in the 1000-run permutation, the top linked component pair of each run exhibited an fMRI-SNP correlation ranging from 0.11 to 0.38 with a median of 0.16 (absolute values). Only one permuted sample yielded an fMRI-SNP association higher than that observed in the original data, resulting in a significant p-value of 0.001 (1/1000) for the identified association (r = -0.38).

The linked fMRI component, after FDR control, was found significantly associated with a number of behavioral assessments, including CD-count, ICS-fc, and AUDIT-4. The most significant association was observed from CD-count, exhibiting a partial correlation of 0.25 ($p = 7.04\times10^{-06}$) after regressing out sex and race/ethnicity, as shown in Figure 8.3b. In addition, the partial correlation remained significant ($r = 0.19$, $p = 6.45\times10^{-04}$) after age was further regressed out, suggesting that the observed altered activation might be more related to current dependence symptoms than alcohol use history. The fMRI loadings also showed significant associations with ICS-fc and AUDIT-4 (correlations of 0.24 and 0.21, respectively), where both measures were highly correlated with CD-count (correlations of 0.68 and 0.76, respectively). Due to this collinearity, the individual effect could not be disentangled and we chose to focus on the most significantly linked symptom CD-count in the following discussion. Figure 8.4 shows the spatial map of the identified fMRI component thresholded at $|Z| > 2$. The brain network included precuneus, superior and inferior parietal lobules, as well as postcentral gyrus, as listed in Table 8.5.

The linked SNP component, with FDR control, exhibited a significant partial correlation with one single behavioral assessment AUDIT-4 ($r = -0.21$, $p = 1.57\times10^{-04}$) after regressing out sex and race/ethnicity, as shown in Figure 8.3c. Again this association might be majorly attributable to current drinking behavior, given that a significant partial correlation ($r = -0.18$, $p = 1.05\times10^{-03}$) was still observed after further regressing out age. Figure 8.5 demonstrates that 2,020 top contributing SNPs were selected out based on the absolute values of the normalized component weights using the same approach as introduced in section 7.3. Figure 8.6 shows a Manhattan plot of weights of loci for the identified SNP component, where clusters spanning more than 15 top contributing SNPs are marked. Table S3 provides a summary of all the identified SNPs, including SNP position, corresponding gene and normalized component weight.

While 1,019 out of 2,020 SNPs were mapped to 457 unique genes, Ingenuity Pathway Analysis (IPA: Ingenuity® Systems, http://www.ingenuity.com) indicated a significant enrichment of neurological diseases in our finding, including bipolar disorder (7.56×10^{-4}), schizophrenia (5.50×10^{-3}) and major depression (4.37×10^{-2}), as highlighted in Table 8.6a. The identified genes were also significantly overrepresented in neuritogenesis (2.81×10^{-4}) and other developmental functions, as listed in Table 8.6b. IPA also revealed a number of enriched canonical pathways, including synaptic long term depression (LTD, 1.70×10^{-5}) and potentiation (LTP, 5.89×10^{-3}), CREB Signaling in Neurons (6.31×10^{-4}), protein kinase A (PKA) signaling (1.26×10^{-2}), as well as GABA receptor signaling (2.24×10^{-2}). Table 8.6c provides a complete summary. In addition, the DAVID (Database for Annotation, Visualization and Integrated Discovery) bioinformatics resource (Huang et al., 2009a, b) identified ion channel activity (1.07×10^{-5}), cell adhesion (9.76×10^{-5}) and transmission of nerve impulse (5.37×10^{-4}) to be the top enriched functional clusters, as highlighted in Table 8.6d. The identified pathways remained highly stable when the number of top contributing SNPs was adjusted from 1,000 to 5,000, as shown in Table 8.7.

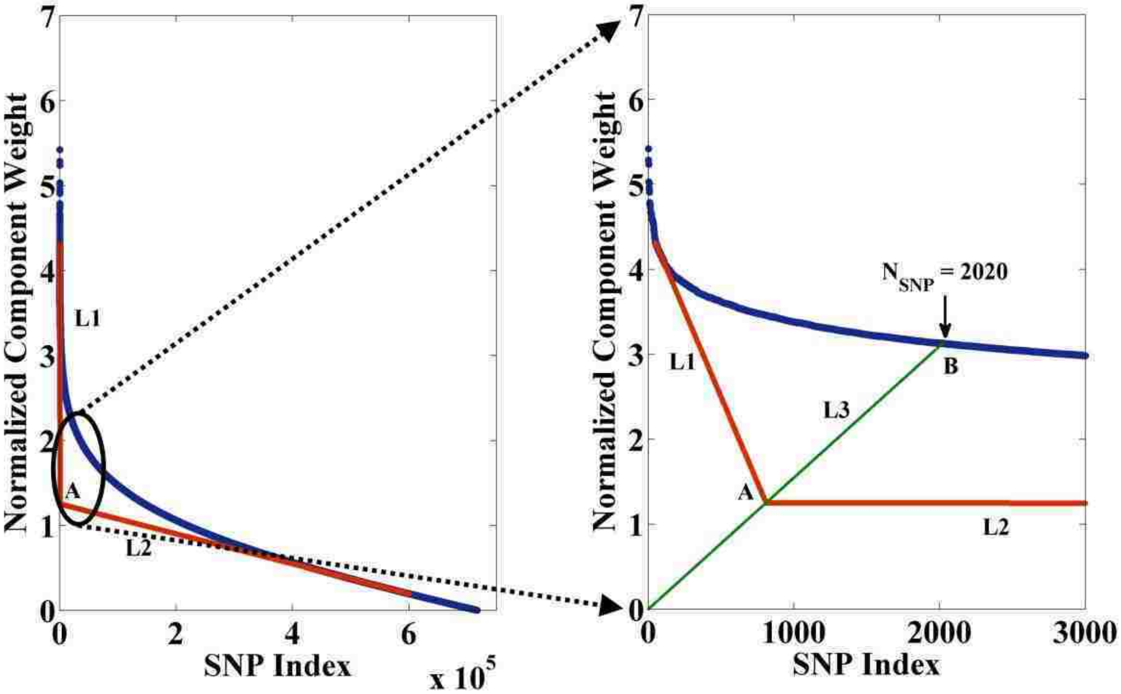

Figure 8.5: Number of top contributing SNPs. The blue curve shows the plot of descending normalized component weights; L1 and L2 represent the linear fits to the two segments of the component curve; A denotes the intersection of L1 and L2; L3 represents the line connecting the origin and the intersection A; B denotes the intersection of L3 and the component curve.

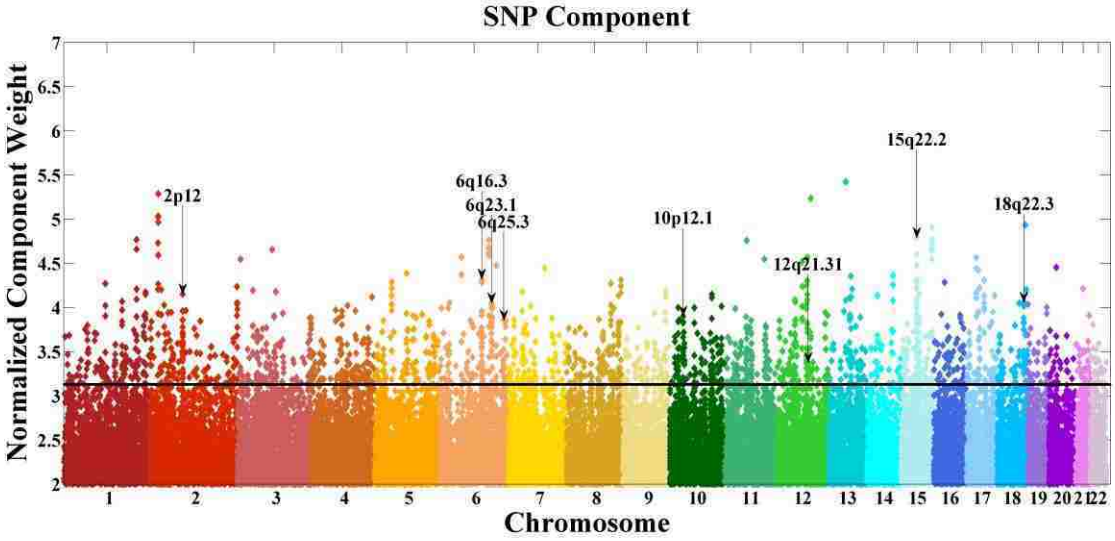

Figure 8.6: Manhattan plot for the identified SNP component.

Table 8.6: Pathway analyses.

a. IPA neurological disease	**Genes**	**P-value**	**P-value (B-H)**
Bipolar disorder	**ALG9,CA10,CACNA1G,CMTM8,DLAT,FLJ35024, GABRG3,GRID1,GRIN2A,GRM5,HTR4,ME1,MRPL48, NCAM1,PRKCA,PRKCB,SYN3,THRB,TRPM2**	**7.56E-04**	**1.75E-01**
Pervasive developmental disorder	AUTS2,CNTNAP2,GNA14,GRIN2A,MBD5,NLGN1	3.49E-03	1.75E-01
Schizophrenia	**ACSBG1,CACNA1G,CACNG2,CNTNAP2,DPYD, GABRG3,GRID1,GRIN2A,GRM5,GRM7,NCAM1, NELL1,NELL2,PRKCB,RBFOX1,ROBO1,RTN4,SHANK3, SND1,ST8SIA2,SYN3,TEKT5,TMTC1,UHMK1,ZFHX3**	**5.50E-03**	**1.75E-01**
Partial seizure	CACNA1G,CACNA2D3,GABRG3,GABRR3,GRIN2A,KCNQ5	9.86E-03	1.75E-01
Autosomal dominant mental retardation	CACNG2,KIF1A,MBD5	1.08E-02	1.75E-01
Complex partial seizure	CACNA1G,CACNA2D3,GABRG3,GABRR3	1.51E-02	1.75E-01
Localization-related epilepsy	CACNA1G,CACNA2D3,CNTNAP2,GABRG3,GABRR3	1.69E-02	1.75E-01
Epilepsy	CACNA1G,CACNA2D3,CH25H,CNTNAP2,GABRG3,GABRR3, GRIN2A,KCNQ5,KCTD7,ME2,STXBP1,TMTC1,TRIB1	2.79E-02	2.03E-01
Absence seizure	CACNA1G,GABRG3,GABRR3	4.14E-02	2.34E-01
Major depression	**CEP68,GABRG3,GRIN2A,GRM5,ITIH4,KCNK2,NCAM1, RCAN1,RSU1**	**4.37E-02**	**2.34E-01**
b. IPA nervous system development and function	**Genes**	**P-value**	**P-value (B-H)**
Neuritogenesis	**CAMK1D,CNTN4,DCC,LRRC4C,MARCO,NCAM1,RAC1**	**2.81E-04**	**9.88E-02**
Formation of neurites	CAMK1D,MARCO,RAC1,ROBO1	1.55E-03	1.75E-01
Formation of dendrites	CAMK1D,MARCO,RAC1	2.39E-03	1.75E-01
Cell-cell adhesion of neurons	CNTN4,NLGN1	5.21E-03	1.75E-01
Axonogenesis	CNTN4,DCC,LRRC4C	5.68E-03	1.75E-01
c. IPA canonical pathway	**Genes**	**P-value**	**P-value (B-H)**
Synaptic Long Term Depression	**GNA14,ITPR1,GRM5,GRM7,GNAI3,GRID1,PLB1, RYR3,LYN,GNAT2,PPP2R1B,PRKCB,PRKCA**	**1.70E-05**	**5.50E-03**
CREB Signaling in Neurons	**GRM5,GRM7,GNAI3,GRIN2A,GRID1,GNAT2, PIK3CD,GNA14,ITPR1,CREB5,PRKCA,PRKCB**	**6.31E-04**	**1.02E-01**
Neuropathic Pain Signaling In Dorsal Horn Neurons	**GRM5,GRM7,GRIN2A,CAMK1D,PIK3CD,ITPR1, PRKCA,PRKCB**	**2.24E-03**	**2.08E-01**

CXCR4 Signaling	GNAI3,RAC1,LYN,GNAT2,PIK3CD,GNA14,ITPR1, ELMO1,PRKCA,PRKCB	2.95E-03	2.08E-01
Virus Entry via Endocytic Pathways	AP2M1,ITSN1,RAC1,PIK3CD,AP1B1,PRKCA,PRKCB	4.90E-03	2.08E-01
Axonal Guidance Signaling	**LRRC4C,ITSN1,KALRN,RAC1,GNA14,ROBO1, ADAMTS2,GNAI3,SRGAP3,NTRK3,DCC,ADAM19,RTN4, ADAM23,GNAT2,PIK3CD,WNT5B,PRKCA,PRKCB**	**5.13E-03**	**2.08E-01**
Heparan Sulfate Biosynthesis (Late Stages)	AADAC,EXT1,EXTL2,HS3ST4,CHST15	5.13E-03	2.08E-01
Breast Cancer Regulation by Stathmin1	GNAI3,CAMK1D,RAC1,UHMK1,PIK3CD,ARHGEF3, ITPR1,E2F3,PPP2R1B,PRKCA,PRKCB	5.25E-03	2.08E-01
Synaptic Long Term Potentiation	**GRM5,GRM7,GRIN2A,GNA14,ITPR1,CREB5,PRKCA, PRKCB**	**5.89E-03**	**2.11E-01**
Ephrin B Signaling	GNAI3,KALRN,ITSN1,RAC1,GNAT2,GNA14	7.24E-03	2.30E-01
Heparan Sulfate Biosynthesis	AADAC,EXT1,EXTL2,HS3ST4,CHST15	9.12E-03	2.65E-01
GNRH Signaling	**GNAI3,RAC1,GNA14,ITPR1,CREB5,NFKB1,PRKCA, PRKCB**	**1.00E-02**	**2.68E-01**
Protein Kinase A Signaling	**PTPN7,PTPRD,PTPN3,ITPR1,NFKB1,CREB5,PDE1C, GNAI3,HHAT,ADD3,RYR3,DCC,PTPRS,PDE8B, PRKCB,PRKCA**	**1.26E-02**	**2.74E-01**
Thrombin Signaling	GNAI3,CAMK1D,GNAT2,PIK3CD,ARHGEF3, GNA14,ITPR1,NFKB1,PRKCA,PRKCB	1.29E-02	2.74E-01
fMLP Signaling in Neutrophils	GNAI3,RAC1,PIK3CD,ITPR1,NFKB1,PRKCA,PRKCB	1.29E-02	2.74E-01
Role of IL-17F in Allergic Inflammatory Airway Diseases	TRAF6,RPS6KA2,CREB5,NFKB1	1.51E-02	2.85E-01
G-Protein Coupled Receptor Signaling	**GRM5,GRM7,GNAI3,HTR4,PDE8B,PIK3CD, GNA14,CREB5,NFKB1,PDE1C,PRKCA,PRKCB**	**1.66E-02**	**2.85E-01**
Role of NFAT in Regulation of the Immune Response	RCAN1,GNAI3,LYN,GNAT2,PIK3CD,GNA14, ITPR1,NFKB1,LCP2	1.70E-02	2.85E-01
Endothelin-1 Signaling	GNAI3,PLB1,GNAT2,EDNRA,PIK3CD,GNA14, ITPR1,PRKCA,PRKCB	1.70E-02	2.85E-01
GABA Receptor Signaling	**GABRG3,GABRR3,AP2M1,AP1B1**	**2.24E-02**	**3.56E-01**
Nitric Oxide Signaling in the Cardiovascular System	CACNA1E,PIK3CD,ITPR1,PDE1C,PRKCA,PRKCB	2.40E-02	3.66E-01
LPS-stimulated MAPK Signaling	RAC1,PIK3CD,NFKB1,PRKCA,PRKCB	2.82E-02	3.94E-01
Tec Kinase Signaling	GNAI3,LYN,GNAT2,PIK3CD,GNA14,NFKB1, PRKCA,PRKCB	2.88E-02	3.94E-01
Dopamine-DARPP32 Feedback in cAMP Signaling	**GNAI3,GRIN2A,CACNA1E,ITPR1,CREB5, PPP2R1B,PRKCA,PRKCB**	**3.16E-02**	**3.94E-01**
Role of p14/p19ARF in Tumor Suppression	RAC1,PIK3CD,NPM2	3.24E-02	3.94E-01
Relaxin Signaling	GNAI3,GNAT2,PDE8B,PIK3CD,GNA14,NFKB1,PDE1C	3.55E-02	3.94E-01
NGF Signaling	TRAF6,RAC1,PIK3CD,RPS6KA2,CREB5,NFKB1	3.80E-02	3.94E-01
Role of Tissue Factor in Cancer	RAC1,LYN,PIK3CD,GNA14,RPS6KA2,PRKCA	3.98E-02	3.94E-01
Renin-Angiotensin Signaling	RAC1,PIK3CD,ITPR1,NFKB1,PRKCA,PRKCB	4.17E-02	3.94E-01
Glutamate Receptor Signaling	**GRM5,GRM7,GRIN2A,GRID1**	**4.17E-02**	**3.94E-01**
Fc Epsilon RI Signaling	RAC1,LYN,PIK3CD,LCP2,PRKCA,PRKCB	4.27E-02	3.94E-01
Androgen Signaling	GNAI3,GNAT2,GNA14,NFKB1,PRKCA,PRKCB	4.47E-02	3.94E-01
FGF Signaling	RAC1,PIK3CD,ITPR1,CREB5,PRKCA	4.90E-02	3.94E-01

d. DAVID functional annotation	**Genes**	**P-value**	**P-value (B-H)**
Ion channel activity	KCNK17,GABRG3,KCND3,NOX5,GRIN2A,CACNG2, CACNA2D3,KCNIP1, KCNK2, ITPR1, KCNIP4, TRPM2,KCTD7,GABRR3,ACCN1,KCNQ5,KCNK9,RYR3, GRM7,CACNA1G,ABCC4,CACNA1E,NALCN,GRID1	1.07E-05	2.70E-03
Cell adhesion	PPFIA2,MPZL2,MYBPC2,PKHD1,GNE,NELL1,NELL2, CTNND2,EDIL3,ITGBL1,ROBO1,RAC1,CNTNAP2, COL12A1,CD6,FLRT2,SELP,HAPLN1,GMDS,CNTNAP4, ADAM23,NLGN1,STXBP1,PTPRS,AJAP1,CTNNA3, CDH12,NCAM1,SIGLEC1,SNED1,CNTN4,NTM	9.76E-05	1.57E-01
Transmission of nerve impulse	PRKCA,GABRG3,GRIN2A,NLGN1,CACNG2,KCNIP1, ACSBG1,GRM5,HCRTR2,ACCN1,KCNQ5,GABRR3, CBLN1.GRM7,SYN3,SLC22A3,CNTNAP2,CACNA1E,UNC13C	5.37E-04	2.70E-01

Table 8.7: IPA canonical pathways with varying numbers of top contributing SNPs.

IPA canonical pathway	N = 1000	N = 2000	N = 3000	N = 4000	N = 5000
	p/p (B-H)	p/p (B-H)	p/p (B-H)	p/p (B-H)	p/p (B-H)
Synaptic Long Term Depression	5.62E-03/7.29E-01	7.24E-05/2.34E-02	2.04E-05/7.08E-03	3.55E-04/1.37E-01	1.86E-3/1.87E-01
	ITPR1,GNA14,GRM5,GRM7,GNAI3,GRID1,PLB1,PLCG2,RYR3,LYN,IGF1R, GNAT2,PPP2R1B,PRKCB,PRKCA				
CREB Signaling in Neurons	1.62E-02/7.29E-01	2.00E-03/2.28E-01	2.51E-03/2.93E-01	1.35E-03/1.95E-01	2.82E-03/1.89E-01
	GRIN2A,ITPR1,GNA14,CREB5,GRM5,GRM7,GNAI3,GRID1, PLCG2,CREB1,PRKAG2,GNAT2,PIK3CD,GRIK1,ATM,PRKCA,PRKCB				
Protein Kinase A Signaling	1.29E-02/7.29E-01	1.17E-02/3.59E-01	5.50E-03/2.93E-01	2.40E-02/4.65E-01	6.46E-04/1.31E-01
	FLNB,SMAD3,NFKB1,TGFBR2,HHAT,RYR3,CREB1,DCC,TGFB2,PTPRT,EYA2,PRKCA,PTPN7, PTPRG,PTPRD,PTPN3,TCF7L1,PDE4B,ITPR1,PDE4D,CREB5,PDE1C,GNAI3,ADD3,PLCG2,PDE1B, PTPRS,PRKAG2,PDE8B,SIRPA,PRKCB,AKAP1				
GABA Receptor Signaling	1.74E-02/7.29E-01	2.19E-02/4.42E-01	5.50E-02/4.73E-01	2.29E-03/1.95E-01	4.68E-05/1.91E-02
	GABRG3,GABBR2,GABRR3,AP2M1,UBQLN1,GAD1,GABRB1,AP1B1,GPHN,GABRB2				
Synaptic Long Term Potentiation	1.62E-01/7.97E-01	1.86E-02/4.42E-01	1.05E-02/3.61E-01	1.86E-02/4.65E-01	2.19E-02/4.88E-01
	GRM5,GRM7,GRIN2A,PLCG2,CREB1,PRKAG2,GNA14,ITPR1,CREB5,PRKCA,PRKCB				

Note: *N* represents the number of top contributing SNPs; *p* and *p*(B-H) represent the uncorrected and Benjamini-Hochberg corrected p-values of enrichment.

We further performed a regression analysis between the fMRI and SNP loadings with additional controlling variables of associated behavioral assessments, as shown in (8.1). The SNP component still showed a significant regression effect ($p = 8.37\times10^{-11}$) on the fMRI component when controlling for CD-count, AUDIT-4 and ICS-fc.

$$loading_{fmri} = b_0 + b_1 \cdot loading_{snp} + b_2 \cdot CD\text{-}count + b_3 \cdot AUDIT4 + b_4 \cdot \text{ICS-fc} + $$
$$+ b_5 \cdot age + b_6 \cdot sex + b_7 \cdot race/ethnicity \qquad (8.1)$$

8. 4 Discussion

One fMRI-SNP component pair was identified presenting a significant association ($r = -0.38$, $p = 3.98\times10^{-12}$). Both the genetic and fMRI components were found to be associated with clinical measures of alcohol dependence (fMRI versus CD-count: $p = 7.04\times10^{-06}$ and SNP versus AUDIT-4: $p = 1.57\times10^{-04}$). The fMRI-SNP association was not due to age, sex or race/ethnicity, and it was beyond the shared variance of behavioral assessments, as confirmed with the partial correction and regression analysis. The genetic component was elicited by a reference comprising three SNP sets derived from the CREB-BDNF pathway. Parallel ICA with multiple references detected that the three referential sets contributed to the same genetic component, suggesting convergent functional influence on neurobiological conditions. The linked fMRI component reflected regional hyperactivation for more severe alcohol dependence. Overall, the results suggest that the CREB-BDNF pathway plays a key role in the genetic factor underlying a proportion of variation in cue-elicited brain activations, which might play a role in phenotypic symptoms of alcohol dependence.

fMRI component: The loadings associated with the fMRI component exhibits a positive correlation with CD-count while the highlighted voxels present positive component weights, indicating that subjects experiencing more severe alcohol dependence symptoms have higher regional activations when exposed to the taste of alcohol. The hyperactivated region majorly comprises precuneus, superior parietal lobule (SPL) and posterior cingulate cortex (PCC), as listed in Table 8.5. Precuneus belongs to associative cortices and is known to be directly involved in a wide spectrum of highly integrated tasks, including episodic memory retrieval (Addis et al., 2004; Lundstrom et al., 2005), self-referential processes (den Ouden et al., 2005; Ochsner et al., 2004) and consciousness (Laureys et al., 2004). Also, precuneus may subserve a variety of functional processes given its wide-spread connections with both adjacent areas such as PCC and SPL, and frontal lobes including prefrontal cortex and anterior cingulate cortex (Cavanna and Trimble, 2006). Although not generally targeted for addiction, precuneus and adjacent parietal regions have been robustly implicated in craving studies, where hyperactivation elicited by drug-related cues are found relating to severity of dependence (Claus et al., 2011; Liu et al., 2013; Park et al., 2007; Tapert et al., 2004; Yalachkov et al., 2010). PCC is suggested as functioning to rapidly associate particularly familiar sensory stimuli (Buccino et al., 2001) and frequently implicated in the processing of drug-related stimuli (Garavan et al., 2000; Kosten et al., 2006; Tapert et al., 2004; Wrase et al., 2007). In particular, there is evidence that increased activation to cocaine cues in PCC is associated with relapse to cocaine abuse (Kosten et al., 2006). Of particular interest, as implicated in a meta-analysis on fMRI studies of alcohol cue reactivity, brain activation in precuneus and PCC, instead of the mesolimbic system, most effectively differentiates cases from controls in terms of alcohol use severity (Schacht et al., 2013). Overall, the identified brain network echoes considerable similar findings and deserves more attention to elucidate the neuropathology of addiction.

SNP component: The linked SNP component exhibits a negative correlation with the fMRI component and behavioral assessment, indicating that subjects carrying lower loadings on the SNP component present higher brain activation in the identified precuneus region. Top 2,020 SNPs are selected out as predominantly contributing to the identified component, among which 1,019 reside in 457 unique genes. Pathway analyses delineate a complex genetic architecture emphasizing synaptic plasticity and other neural signaling pathways. Considering that a meta-analysis identified 3,800 genes differentially expressed between models of high and low amounts alcohol consumption (Mulligan et al., 2006), our finding highlights one side of the story where CREB serves as a convergence point of various neurocircuitries related to alcohol dependence.

CREB signaling: Given that the reference is derived from CREB and BDNF genes, it's not surprising that the related pathway is enriched in the finding. cAMP-response element-binding protein (CREB) functions as a transcription factor and is well known for its role in neuronal plasticity and long-term memory (Carlezon et al., 2005; Lonze and Ginty, 2002; Silva et al., 1998). Brain derived neurotrophic factor (BDNF) is a CREB regulated gene that also plays an important role in synaptic plasticity (Bramham and Messaoudi, 2005; Mattson et al., 2004). While drug addiction might be partly considered as a result of adaptations in specific brain neurons due to repeated exposure to a substance of abuse, there is considerable evidence that learning/memory and addiction converge in a variety of aspects, one of which being that both are modulated by transcription factor CREB and neurotrophic factors (Bolanos and Nestler, 2004; Nestler, 2002, 2005; Pandey et al., 2005; Pierce and Bari, 2001). In particular, one SNP in BDNF (rs6265_A or Val66Met, 'A' represents the minor allele) has been identified as predicting relapse in alcohol dependence patients, where minor allele carriers show decreased vulnerability to relapse (Wojnar et al., 2009). This is consistent with our finding where the same SNP exhibits a positive weight, indicating that minor allele carriers exhibit lower brain activation which is associated with less severe alcohol dependence.

Synaptic LTD and LTP: Synaptic LTP and LTD are two forms of synaptic plasticity which enhances or weakens, respectively, the synchronized stimulations between neurons, thus allowing the refinement of neuronal circuits underlying learning and memory (Malenka and Bear, 2004). It's commonly recognized that synaptic plasticity plays an important role in the development of addiction, through which use of drug progresses from impulsive to compulsive behavior (Kauer and Malenka, 2007; Mameli and Luscher, 2011; Nestler, 2001). Genetic variations may also account for a proportion of variation in synaptic plasticity. As revealed by a meta-analysis, LTD and LTP are among the top enriched pathways for the 396 addiction-related genes implicated in two or more independent studies (Li et al., 2008). In our finding, the synaptic LTD and LTP pathways strongly overlap with the CREB signaling pathway, highlighting metabotropic glutamate receptors (GRM5 and GRM7) and ionotropic glutamate receptors (GRID1 and GRIN2A), as well as protein kinase C (PRKCA and PRKCB). The SNP in GRIN2A (rs4628972_G) and three SNPs in PRKCA (rs17688881_C, rs721429_A and rs7217618_C) present negative weights, indicating that the minor allele carriers show high brain activation and more severe alcohol dependence. The opposite is observed for the rest (rs1000061_G in GRM5, rs1353832_C in GRM7, rs1863824_C in GRID1, rs8077110_T in PRKCA and rs880824_A in PRKCB) which all present positive weights.

Protein kinase A signaling (PKA): The cAMP-PKA pathway is a primary signaling cascade that modulates numerous cellular events in neurons, including synaptic plasticity (Abel and Kandel, 1998; Skalhegg and Tasken, 2000; Waltereit and Weller, 2003). It's documented that all drugs of abuse alter cAMP-PKA signaling and activation of the cAMP-PKA pathway leads to increased activity of the transcription factor CREB (Ron and Jurd, 2005). Direct evidence is also provided that genetic mutation reducing cAMP-PKA signaling results in increased sensitivity to the sedative effects of ethanol (Wand et al., 2001). In our finding, a number of genes are involved in this pathway, 4 of which encode protein tyrosine phosphatase (PTPN3, PTPN7, PTPRD and PTPRS), known to be signaling molecules regulating a variety of cellular processes including cell growth and differentiation (Denhertog et al., 1993). Specifically, both two SNPs in

PTPN3 (rs10979861_C and rs7046039_T) exhibit negative weights while the rest (rs10920336_C in PTPN7, rs10815927_C and rs10756029_C in PTPRD, and rs1015675_A in PTPRS) contribute with positive weights.

GABA receptor signaling: As a major inhibitory neurotransmitter in the central nervous system, GABAergic signaling has been implicated in addiction in numerous studies (Kauer and Malenka, 2007; Pandey, 1998). It's reported that chronic cocaine uses decrease GABAergic synapse function, such that LTP induction is not effectively suppressed at excitatory synapses, which is believed to increase the likelihood of firing to a stimulus (Liu et al., 2005). Polymorphisms in GABA receptors are consistently identified as susceptibility loci to addiction in genome-wide association studies (Bierut et al., 2010; Enoch et al., 2009; Radel et al., 2005). In particular, a number of SNPs in GABRG3 are shown to be associated with alcohol dependence (Dick et al., 2004). In our finding, all the three identified SNPs in GABRG3 (rs12439549_G, rs4438262_G and rs3922613_G) contribute with positive weights. Regarding GABRR3, two SNPs (rs1874864_G and rs7638369_T) present positive weights while the rest (rs1844934_T, rs1688378_A and rs1492054_C) exhibit negative weights. Although not targeted as references, these two genes deserve more attention in future investigations.

The identified genetic component is not specific to alcohol use disorders. As suggested by IPA, the genes harboring top contributing SNPs are overrepresented for other neurobiological diseases, including bipolar disorder, schizophrenia and major depression. This suggests a genetic basis for the comorbidity among these disorders, for which accumulated evidence has been provided (Johnson et al., 2009; Kendler et al., 2003; Purcell et al., 2009a). The genetic component also captures neurodevelopmental functions such as neuritogenesis and cell adhesion, which conforms with the notion that adolescent cortical development is a critical period of vulnerability for both addiction (Crews et al., 2007) and schizophrenia (Rapoport et al., 2012). On the other hand, it needs to be emphasized that the genetic component is more tightly associated with the imaging component, explaining a relatively large amount of variance in the observed alterations in brain activation. As the fMRI-SNP association is not majorly due to the shared relationship with alcohol dependence, the observed variation of brain function is partly regulated by the genetic variations. The remainder variances might be due to other genetic or environmental factors or their interactions which awaits further investigations.

In summary, our finding confirms the genetic variation induced vulnerability to disruptions in brain function which might play a role in alcohol dependence. The identified genetic component reflects polygenicity and heterogeneity, involving a variety of synaptic plasticity and neural signaling pathways implicated in a wide spectrum of disorders including addiction. Association analyses reveal that this complex genetic factor strongly affects neurobiological conditions, contributing to altered cue-elicited brain activations in precuneus. While the identified brain network is known to participate in many cognitive processes and robustly implicated in craving-related studies, our work emphasizes the genetic underpinnings and highlights a key role of the CREB-BDNF pathway.

CHAPTER 9 CONCLUSIONS AND FUTURE WORK

In summary, our work makes several important contributions to advance the application of ICA to imaging genetics studies. First, as presented in Chapter 3, we designed a consistency-based order estimation approach to locate the order range which allows ICA to extract relatively accurate, consistent components and loadings for genotype data. Subsequently, to assist the decomposition of high-dimensional data, we developed a semi-blind multivariate model, pICA-R, as presented in Chapter 4. The new approach assesses many variables for aggregate effects while incorporating prior knowledge. It helps pinpoint a particular component of interest embedded in a large complex dataset, thus improving the robustness of the results. Guided by a single referential SNP set derived from ANK3, pICA-R identified a significant sMRI-SNP association, revealing a complex genetic component underlying the SZ-related gray matter concentration reduction in frontal and temporal regions. In Chapter 5, we further extended pICA-R to accommodate multiple referential SNP sets whose interrelationship is unknown. The extended model enables robust investigations on potentially related genetic variants implicated in molecular, cellular or system biology. When the extended parallel ICA with multiple references was employed to study the genetic influence on alcohol dependence, three referential SNP sets derived from the CREB-BDNF pathway were identified as contributing to the same SNP component significantly linked to altered regional brain activation. The results strongly suggest that the CREB-BDNF pathway functions as a convergence point of various synaptic and neural signaling processes underlying the variation in cue-elicited brain activations in precuneus, which might be involved in phenotypic symptoms of alcohol dependence.

There are still many aspects to be explored in the future regarding the application of ICA in the imaging genetics field. For instance, from the method's point of view, a regularization of sparsity can be introduced to further improve the robustness of the model in high-dimensional space. The assumption is that when millions of genetic variants are included in a model as potential causal loci, it's more likely that most of them do not contribute significantly to the trait of interest. Consequently, the underlying component (or source) is expected to be sparse. Sparsity-regularized linear models have been used for model selection in the context of genome wide association (Cantor et al., 2010; Vounou et al., 2010; Wu et al., 2009), however yet investigated in commonly used ICA frameworks. On the other hand, it would also be interesting to investigate the genetic factors underlying neurobiological traits captured with other imaging modalities, such as functional network connectivity in resting-state fMRI and tractography in diffusion tensor imaging. For instance, it has been widely acknowledged that SZ patients tend to have less integrated, more diverse profiles of brain functional connectivity (Lynall et al., 2010). This is at least partly due to deficits in white matter tracts which result in a disconnected configuration of gray matter regions (Bassett et al., 2008; Zhou et al., 2008). Investigation on the genetic underpinnings is expected to further improve our understanding of the biological mechanisms underlying the disorder.

Some of the works in this dissertation have been published. The consistency-based order estimation approach appears in (Chen et al., 2012a); the preliminary investigation of guided exploration of genomic risk for SZ is in (Chen et al., 2012c) and the pICA-R approach appears in (Chen et al., 2013). A manuscript is ready for submission for the exploration of scanning effects and another manuscript is under preparation for the extended parallel ICA with multiple references and its application to alcohol dependence.

REFERENCES

Abel, T., Kandel, E., 1998. Positive and negative regulatory mechanisms that mediate long-term memory storage. Brain Research Reviews 26, 360-378.

Abrahams, B.S., Tentler, D., Perederiy, J.V., Oldham, M.C., Coppola, G., Geschwind, D.H., 2007. Genome-wide analyses of human perisylvian cerebral cortical patterning. Proceedings of the National Academy of Sciences of the United States of America 104, 17849-17854.

Addis, D.R., McIntosh, A.R., Moscovitch, M., Crawley, A.P., McAndrews, M.P., 2004. Characterizing spatial and temporal features of autobiographical memory retrieval networks: a partial least squares approach. Neuroimage 23, 1460-1471.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N Petrov, F.C. (Ed.), Proc. of 2nd International Symposium on Information Theory. Academiai Kiado, Budapest, pp. 267-281.

Albert, K.A., Hemmings, H.C., Adamo, A.I.B., Potkin, S.G., Akbarian, S., Sandman, C.A., et al., 2002. Evidence for decreased DARPP-32 in the prefrontal cortex of patients with schizophrenia. Archives of General Psychiatry 59, 705-712.

Aleman, A., Hijman, R., de Haan, E.H., Kahn, R.S., 1999. Memory impairment in schizophrenia: a meta-analysis. The American journal of psychiatry 156, 1358-1366.

Amari, S., 1998. Natural Gradient Works Efficiently in Learning. Neural Computation 10, 251-276.

Amari, S., Cichocki, A., Yang, H.H., 1996. A new learning algorithm for blind signal separation. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (Eds.), Advances in

Neural Information Processing Systems. MIT press, Cambridge, MA, pp. 752--763.

Andreasen, N.C., Flaum, M., Arndt, S., 1992. The Comprehensive Assessment of Symptoms and History (Cash) - an Instrument for Assessing Diagnosis and Psychopathology. Archives of General Psychiatry 49, 615-623.

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry--the methods. Neuroimage 11, 805-821.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839-851.

Babor, T.F., Higgins-Biddle, J.C., Sauders, J.B., Monteiro, M.G., 2001. The Alcohol Use Disorders Identification Test: Guide for Use in Primary Care. World Health Organization, Geneva, Switzerland.

Bakker, S.C., van der Meulen, E.M., Buitelaar, J.K., Sandkuijl, L.A., Pauls, D.L., Monsuur, A.J., et al., 2003. A whole-genome scan in 164 Dutch sib pairs with attention-deficit/hyperactivity disorder: suggestive evidence for linkage on chromosomes 7p and 15q. American Journal of Human Genetics 72, 1251-1260.

Barnett, J.H., Smoller, J.W., 2009. The Genetics of Bipolar Disorder. Neuroscience 164, 331-343.

Bassett, D.S., Bullmore, E., Verchinski, B.A., Mattay, V.S., Weinberger, D.R., Meyer-Lindenberg, A., 2008. Hierarchical organization of human cortical networks in health and schizophrenia. Journal of Neuroscience 28, 9239-9248.

Bell, A.J., Sejnowski, T.J., 1995a. An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7, 1129-1159.

Bell, A.J., Sejnowski, T.J., 1995b. An Information Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation 7, 1129-1159.

Bernstein, M.A., Huston, J., 3rd, Ward, H.A., 2006. Imaging artifacts at 3.0T. Journal of magnetic resonance imaging : JMRI 24, 735-746.

Bierut, L.J., 2010. Convergence of genetic findings for nicotine dependence and smoking related diseases with chromosome 15q24-25. Trends in Pharmacological Sciences 31, 46-51.

Bierut, L.J., Agrawal, A., Bucholz, K.K., Doheny, K.F., Laurie, C., Pugh, E., et al., 2010. A genome-wide association study of alcohol dependence. Proc Natl Acad Sci U S A 107, 5082-5087.

Blumenfeld, R.S., Parks, C.M., Yonelinas, A.P., Ranganath, C., 2011. Putting the pieces together: the role of dorsolateral prefrontal cortex in relational memory encoding. Journal of Cognitive Neuroscience 23, 257-265.

Bolanos, C.A., Nestler, E.J., 2004. Neurotrophic mechanisms in drug addiction. Neuromolecular Medicine 5, 69-83.

Bramham, C.R., Messaoudi, E., 2005. BDNF function in adult synaptic plasticity: The synaptic consolidation hypothesis. Progress in Neurobiology 76, 99-125.

Buccino, G., Binkofski, F., Fink, G.R., Fadiga, L., Fogassi, L., V, G., et al., 2001. Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. European Journal of Neuroscience 13, 400-404.

Buckner, R.L., Kelley, W.M., Petersen, S.E., 1999. Frontal cortex contributes to human memory formation. Nature neuroscience 2, 311-314.

Calhoun, V.D., Adali, T., Giuliani, N.R., Pekar, J.J., Kiehl, K.A., Pearlson, G.D., 2006. Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. Human Brain

Mapping 27, 47-62.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Human Brain Mapping 14, 140-151.

Cannon, T.D., Thompson, P.M., van Erp, T.G.M., Toga, A.W., Poutanen, V.P., Huttunen, M., et al., 2002. Cortex mapping reveals regionally specific patterns of genetic and disease-specific gray-matter deficits in twins discordant for schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 99, 3228-3233.

Cantor, R.M., Lange, K., Sinsheimer, J.S., 2010. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. American Journal of Human Genetics 86, 6-22.

Caporaso, N., Gu, F.Y., Chatterjee, N., Jin, S.C., Yu, K., Yeager, M., et al., 2009. Genome-Wide and Candidate Gene Association Study of Cigarette Smoking Behaviors. PloS one 4.

Cardno, A.G., Gottesman, I.I., 2000. Twin studies of schizophrenia: From bow-and-arrow concordances to star wars mx and functional genomics. American Journal of Medical Genetics 97, 12-17.

Cardoso, J.F., Soloumiac, A., 1993. Blind Beamforming for Non Gaussian Signals. Radar and Signal Processing, IEE Proceedings F 140, 362-370.

Carlezon, W.A., Duman, R.S., Nestler, E.J., 2005. The many faces of CREB. Trends in Neurosciences 28, 436-445.

Cavanna, A.E., Trimble, M.R., 2006. The precuneus: a review of its functional anatomy and behavioural correlates. Brain : a journal of neurology 129, 564-583.

Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., et al., 2007. Replicating genotype-phenotype associations. Nature 447, 655-660.

Chavarria-Siles, I., Rijpkema, M., Lips, E., Arias-Vasquez, A., Verhage, M., Franke, B., et al., 2012. Genes Encoding Heterotrimeric G-proteins Are Associated with Gray Matter Volume Variations in the Medial Frontal Cortex. Cerebral Cortex.

Chen, J., Calhoun, V.D., Liu, J., 2012a. ICA order selection based on consistency: application to genotype data. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 2012, 360-363.

Chen, J., Calhoun, V.D., Liu, J., 2012b. ICA Order Selection Based on Consistency: Application to Genotype Data. 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, p. in press.

Chen, J., Calhoun, V.D., Pearlson, G.D., Ehrlich, S., Turner, J.A., Ho, B.C., et al., 2012c. Multifaceted genomic risk for brain function in schizophrenia. Neuroimage 61, 866-875.

Chen, J., Calhoun, V.D., Pearlson, G.D., Perrone-Bizzozero, N., Sui, J., Turner, J.A., et al., 2013. Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. Neuroimage 83C, 384-396.

Claus, E.D., Ewing, S.W., Filbey, F.M., Sabbineni, A., Hutchison, K.E., 2011. Identifying neurobiological phenotypes associated with alcohol use disorder severity. Neuropsychopharmacology 36, 2086-2096.

Collingridge, G.L., Peineau, S., Howland, J.G., Wang, Y.T., 2010. Long-term depression in the CNS. Nature reviews. Neuroscience 11, 459-473.

Comon, P., 1994. Independent Component Analysis, a New Concept. Signal Processing 36, 287-314.

Conner, B.T., Noble, E.P., Berman, S.M., Ozkaragoz, T., Ritchie, T., Antolin, T., et al., 2005. DRD2 genotypes and substance use in adolescent children of alcoholics. Drug and Alcohol Dependence 79, 379-387.

Cooke, S.F., Bliss, T.V., 2006. Plasticity in the human central nervous system. Brain : a journal of neurology 129, 1659-1673.

Correa, N., Adali, T., Calhoun, V.D., 2007. Performance of blind source separation algorithms for fMRI analysis using a group ICA method. Magnetic Resonance Imaging 25, 684-694.

Coyle, J.T., 2006. Glutamate and schizophrenia: beyond the dopamine hypothesis. Cellular and molecular neurobiology 26, 365-384.

Crabbe, J.C., 2002. Genetic contributions to addiction. Annual review of psychology 53, 435-462.

Crews, F., He, J., Hodge, C., 2007. Adolescent cortical development: A critical period of vulnerability for addiction. Pharmacology Biochemistry and Behavior 86, 189-199.

Davidson, R.J., Pizzagalli, D., Nitschke, J.B., Putnam, K., 2002. Depression: perspectives from affective neuroscience. Annual review of psychology 53, 545-574.

Dawy, Z., Sarkis, M., Hagenauer, J., Mueller, J., 2005. A Novel Gene Mapping Algorithm Based on Independent Component Analysis. Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processings. IEEE International

Conference on Acoustics, Speech, and Signal Processings, Philadelphia, pp. 381-384.

de Zwart, J.A., Ledden, P.J., van Gelderen, P., Bodurka, J., Chu, R., Duyn, J.H., 2004. Signal-to-noise ratio and parallel imaging performance of a 16-channel receive-only brain coil array at 3.0 Tesla. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 51, 22-26.

den Ouden, H.E.M., Frith, U., Frith, C., Blakemore, S.J., 2005. Thinking about intentions. Neuroimage 28, 787-796.

Denhertog, J., Pals, C.E.G.M., Peppelenbosch, M.P., Tertoolen, L.G.J., Delaat, S.W., Kruijer, W., 1993. Receptor Protein-Tyrosine Phosphatase-Alpha Activates Pp60(C-Src) and Is Involved in Neuronal Differentiation. Embo Journal 12, 3789-3798.

Deutsch, S.I., Rosse, R.B., Schwartz, B.L., Mastropaolo, J., 2001. A revised excitotoxic hypothesis of schizophrenia: therapeutic implications. Clinical neuropharmacology 24, 43-49.

Dick, D.M., Edenberg, H.J., Xuei, X.L., Goate, A., Kuperman, S., Schuckit, M., et al., 2004. Association of GABRG3 with alcohol dependence. Alcoholism-Clinical and Experimental Research 28, 4-9.

Duan, J.B., Sanders, A.R., Gejman, P.V., 2010. Genome-wide approaches to schizophrenia. Brain Research Bulletin 83, 93-102.

Duewell, S., Wolff, S.D., Wen, H., Balaban, R.S., Jezzard, P., 1996. MR imaging contrast in human brain tissue: assessment and optimization at 4 T. Radiology 199, 780-786.

Edelman, G.M., 1983. Cell adhesion molecules. Science 219, 450-457.

Edenberg, H.J., Foroud, T., 2006. The genetics of alcoholism: identifying specific genes through family studies. Addict Biol 11, 386-396.

Egan, M.F., Straub, R.E., Goldberg, T.E., Yakub, I., Callicott, J.H., Hariri, A.R., et al., 2004. Variation in GRM3 affects cognition, prefrontal glutamate, and risk for schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 101, 12604-12609.

Enoch, M.A., Goldman, D., 1999. Genetics of alcoholism and substance abuse. Psychiatr Clin North Am 22, 289-299, viii.

Enoch, M.A., Hodgkinson, C.A., Yuan, Q., Albaugh, B., Virkkunen, M., Goldman, D., 2009. GABRG1 and GABRA2 as independent predictors for alcoholism in two populations. Neuropsychopharmacology 34, 1245-1254.

Erhardt, E.B., Allen, E.A., Wei, Y., Eichele, T., Calhoun, V.D., 2011. SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability. Neuroimage.

Farrer, L.A., Cupples, L.A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R., et al., 1997. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease - A meta-analysis. Jama-Journal of the American Medical Association 278, 1349-1356.

Fennema-Notestine, C., Gamst, A.C., Quinn, B.T., Pacheco, J., Jernigan, T.L., Thal, L., et al., 2007. Feasibility of multi-site clinical structural neuroimaging studies of aging using legacy data. Neuroinformatics 5, 235-245.

Ferreira, L.K., Diniz, B.S., Forlenza, O.V., Busatto, G.F., Zanetti, M.V., 2011.

Neurostructural predictors of Alzheimer's disease: a meta-analysis of VBM studies. Neurobiology of aging 32, 1733-1741.

Ferreira, M.A., O'Donovan, M.C., Meng, Y.A., Jones, I.R., Ruderfer, D.M., Jones, L., et al., 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nature Genetics 40, 1056-1058.

Filbey, F.M., Claus, E., Audette, A.R., Niculescu, M., Banich, M.T., Tanabe, J., et al., 2008a. Exposure to the taste of alcohol elicits activation of the mesocorticolimbic neurocircuitry. Neuropsychopharmacology 33, 1391-1401.

Filbey, F.M., Ray, L., Smolen, A., Claus, E.D., Audette, A., Hutchison, K.E., 2008b. Differential neural response to alcohol priming and alcohol taste cues is associated with DRD4 VNTR and OPRM1 genotypes. Alcoholism-Clinical and Experimental Research 32, 1113-1123.

Filbey, F.M., Ray, L., Smolen, A., Claus, E.D., Audette, A., Hutchison, K.E., 2008c. Differential neural response to alcohol priming and alcohol taste cues is associated with DRD4 VNTR and OPRM1 genotypes. Alcoholism-Clinical and Experimental Research 32, 1113-1123.

First, M.B., Gibbon, M., Spitzer, R.L., Williams, J.B.W., Benjamin, L.S., 1997. Structured clinical interview for DSM-IV axis II personality disorders, (SCID-II), 4th ed. American Psychiatric Press, Washington, DC.

Fischl, B., Dale, A.M., 2000. Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences of the United States of America 97, 11050-11055.

Focke, N.K., Helms, G., Kaspar, S., Diederich, C., Toth, V., Dechent, P., et al., 2011.

Multi-site voxel-based morphometry--not quite there yet. Neuroimage 56, 1164-1170.

Fornito, A., Yucel, M., Dean, B., Wood, S.J., Pantelis, C., 2009. Anatomical abnormalities of the anterior cingulate cortex in schizophrenia: bridging the gap between neuroimaging and neuropathology. Schizophrenia Bulletin Epub 35, 973-993.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nature Reviews Neuroscience 8, 700-711.

Frantseva, M.V., Fitzgerald, P.B., Chen, R., Moller, B., Daigle, M., Daskalakis, Z.J., 2008. Evidence for impaired long-term potentiation in schizophrenia and its relationship to motor skill learning. Cerebral Cortex 18, 990-996.

Garavan, H., Pankiewicz, J., Bloom, A., Cho, J.K., Sperry, L., Ross, T.J., et al., 2000. Cue-induced cocaine craving: Neuroanatomical specificity for drug users and drug stimuli. American Journal of Psychiatry 157, 1789-1798.

Gardner, D.M., Murphy, A.L., O'Donnell, H., Centorrino, F., Baldessarini, R.J., 2010. International consensus study of antipsychotic dosing. The American journal of psychiatry 167, 686-693.

Georgiev, P., Cichocki, A., 2001. Blind source separation via symmetric eigenvalue decomposition. Isspa 2001: Sixth International Symposium on Signal Processing and Its Applications, Vols 1 and 2, Proceedings, 17-20.

Gianoulakis, C., 2001. Influence of the endogenous opioid system on high alcohol consumption and genetic predisposition to alcoholism. Journal of Psychiatry & Neuroscience 26, 304-318.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H.M., et al.,

2003. The International HapMap Project. Nature 426, 789-796.

Giedd, J.N., 2004. Structural magnetic resonance imaging of the adolescent brain. Adolescent Brain Development: Vulnerabilities and Opportunities 1021, 77-85.

Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., et al., 2011. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. Plos Genetics 7, e1002334.

Giuliani, N.R., Calhoun, V.D., Pearlson, G.D., Francis, A., Buchanan, R.W., 2005. Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia. Schizophrenia Research 74, 135-147.

Glahn, D.C., Laird, A.R., Ellison-Wright, I., Thelen, S.M., Robinson, J.L., Lancaster, J.L., et al., 2008. Meta-analysis of gray matter anomalies in schizophrenia: application of anatomic likelihood estimation and network analysis. Biological Psychiatry 64, 774-781.

Glahn, D.C., Ragland, J.D., Abramoff, A., Barrett, J., Laird, A.R., Bearden, C.E., et al., 2005. Beyond hypofrontality: A quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. Human Brain Mapping 25, 60-69.

Goldman, D., 1993. Recent developments in alcoholism:genetic transmission. Recent Dev Alcohol 11, 231-248.

Gollub, R., Shoemaker, J.M., King, M., White, T., Ehrlich, S., Sponheim, S., et al., in press. The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. Journal of

NeuroInformatics.

Gottesman, I.I., Gould, T.D., 2003. The endophenotype concept in psychiatry: Etymology and strategic intentions. American Journal of Psychiatry 160, 636-645.

Gottesman, I.I., Shields, J., 1967. A Polygenic Theory of Schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 58, 199-205.

Grant, B.F., Stinson, F.S., Harford, T.C., 2001. Age at onset of alcohol use and DSM-IV alcohol abuse and dependence: A 12-year follow-up. Journal of Substance Abuse 13, 493-504.

Guidotti, A., Di-Giorgi-Gerevini, V., 2002. Decrease in reelin and glutamic acid decarboxylase(67) (GAD(67)) expression in schizophrenia and bipolar disorder (vol 57, pg 1061, 2000). Archives of General Psychiatry 59, 12-12.

Gunderson, K.L., Steemers, F.J., Ren, H.G., Ng, P., Zhou, L.X., Tsan, C., et al., 2006. Whole-genome genotyping. DNA Microarrays Part A: Array Platforms and Wet-Bench Protocols 410, 359-+.

Gur, R.E., Nimgaonkar, V.L., Almasy, L., Calkins, M.E., Ragland, J.D., Pogue-Geile, M.F., et al., 2007. Neurocognitive Endophenotypes in a Multiplex Multigenerational Family Study of Schizophrenia. The American journal of psychiatry 164, 813-819.

Harrison, P.J., 1999. The neuropathology of schizophrenia - A critical review of the data and their interpretation. Brain : a journal of neurology 122, 593-624.

Harrison, P.J., Owen, M.J., 2003. Genes for schizophrenia? Recent findings and their pathophysiological implications. Lancet 361, 417-419.

Hashemi, R.H., G., B.W., Lisanti, C.J., 2003. MRI: The Basics. Lippincott Williams & Wilkins.

Heath, A.C., Bucholz, K.K., Madden, P.A., Dinwiddie, S.H., Slutske, W.S., Bierut, L.J., et al., 1997. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. Psychological Medicine 27, 1381-1396.

Heather, N., Booth, P., Luce, A., 1998. Impaired control scale: cross-validation and relationships with treatment outcome. Addiction 93, 761-771.

Heather, N., Tebbutt, J.S., Mattick, R.P., Zamir, R., 1993. Development of a scale for measuring impaired control over alcohol consumption: a preliminary report. J Stud Alcohol 54, 700-709.

Henkelman, R.M., 1985. Measurement of signal intensities in the presence of noise in MR images. Medical physics 12, 232-233.

Hennekens, C.H., Hennekens, A.R., Hollar, D., Casey, D.E., 2005. Schizophrenia and increased risks of cardiovascular disease. American heart journal 150, 1115-1121.

Hoe, H.S., Lee, K.J., Carney, R.S., Lee, J., Markova, A., Lee, J.Y., et al., 2009. Interaction of reelin with amyloid precursor protein promotes neurite outgrowth. The Journal of neuroscience : the official journal of the Society for Neuroscience 29, 7459-7473.

Honea, R., Crow, T.J., Passingham, D., Mackay, C.E., 2005. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. The American journal of psychiatry 162, 2233-2245.

Howes, O.D., Kapur, S., 2009. The dopamine hypothesis of schizophrenia: version III--the final common pathway. Schizophrenia Bulletin 35, 549-562.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009a. Bioinformatics enrichment tools:

paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37, 1-13.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4, 44-57.

Hyvarinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis 1ed. Wiley, New York.

Hyvarinen, A., Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. Neural Computation 9, 1483-1492.

Jay, T.M., 2003. Dopamine: a potential substrate for synaptic plasticity and memory mechanisms. Progress in Neurobiology 69, 375-390.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825-841.

Jeste, D.V., Gladsjo, J.A., Lindamer, L.A., Lacro, J.P., 1996. Medical comorbidity in schizophrenia. Schizophrenia Bulletin 22, 413-430.

Johnson, C., Drgon, T., Liu, Q.R., Walther, D., Edenberg, H., Rice, J., et al., 2006. Pooled association genome scanning for alcohol dependence using 104,268 SNPs: Validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 141B, 844-853.

Johnson, C., Drgon, T., McMahon, F.J., Uhl, G.R., 2009. Convergent Genome Wide Association Results for Bipolar Disorder and Substance Dependence. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 150B, 182-190.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., et al., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30, 436-443.

Kauer, J.A., Malenka, R.C., 2007. Synaptic plasticity and addiction. Nature Reviews Neuroscience 8, 844-858.

Kendler, K.S., Prescott, C.A., Myers, J., Neale, M.C., 2003. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. Archives of General Psychiatry 60, 929-937.

Klamroth, K., Tind, J., 2007. Constrained optimization using multiple objective programming. Journal of Global Optimization 37, 325-355.

Knopik, V.S., Heath, A.C., Madden, P.A.F., Bucholz, K.K., Slutske, W.S., Nelson, E.C., et al., 2004. Genetic effects on alcohol dependence risk: re-evaluating the importance of psychiatric and other heritable risk factors. Psychological Medicine 34, 1519-1530.

Kosten, T.R., Scanley, B.E., Tucker, K.A., Oliveto, A., Prince, C., Sinha, R., et al., 2006. Cue-induced brain activity changes and relapse in cocaine-dependent patients. Neuropsychopharmacology 31, 644-650.

Krystal, J.H., Mathew, S.J., D'Souza, D.C., Garakani, A., Gunduz-Bruce, H., Charney, D.S., 2010. Potential psychiatric applications of metabotropic glutamate receptor agonists and antagonists. CNS drugs 24, 669-693.

Krystal, J.H., Petrakis, I.L., Mason, G., Trevisan, L., D'Souza, D.C., 2003. N-methyl-D-aspartate glutamate receptors and alcoholism: reward, dependence, treatment, and vulnerability. Pharmacology & Therapeutics 99, 79-94.

Kuperberg, G.R., Broome, M.R., McGuire, P.K., David, A.S., Eddy, M., Ozawa, F., et al., 2003. Regionally localized thinning of the cerebral cortex in schizophrenia. Archives of General Psychiatry 60, 878-888.

Lambert, S., Davis, J.Q., Bennett, V., 1997. Morphogenesis of the node of Ranvier: co-clusters of ankyrin and ankyrin-binding integral proteins define early developmental intermediates. The Journal of neuroscience : the official journal of the Society for Neuroscience 17, 7025-7036.

Lancaster, J.L., Rainey, L.H., Summerlin, J.L., Freitas, C.S., Fox, P.T., Evans, A.C., et al., 1997. Automated labeling of the human brain: a preliminary report on the development and evaluation of a forward-transform method. Human Brain Mapping 5, 238-242.

Lancaster, J.L., Woldorff, M.G., Parsons, L.M., Liotti, M., Freitas, C.S., Rainey, L., et al., 2000. Automated Talairach atlas labels for functional brain mapping. Human Brain Mapping 10, 120-131.

Laureys, S., Owen, A.M., Schiff, N.D., 2004. Brain function in coma, vegetative state, and related disorders. Lancet Neurology 3, 537-546.

Lawrie, S.M., Abukmeil, S.S., 1998. Brain abnormality in schizophrenia. A systematic and quantitative review of volumetric magnetic resonance imaging studies. The British journal of psychiatry : the journal of mental science 172, 110-120.

Lawrie, S.M., Whalley, H.C., Job, D.E., Johnstone, E.C., 2003. Structural and functional abnormalities of the amygdala in schizophrenia. Amygdala in Brain Function: Bacic and Clinical Approaches 985, 445-460.

Li, C.Y., Mao, X.Z., Wei, L.P., 2008. Genes and (Common) pathways underlying drug addiction. Plos Computational Biology 4.

Li, Y.O., Adali, T., Calhoun, V.D., 2007. Estimating the number of independent components for functional magnetic resonance imaging data. Human Brain Mapping 28, 1251-1266.

Lidow, M.S., 2003. Calcium signaling dysfunction in schizophrenia: a unifying approach. Brain Research Reviews 43, 70-84.

Lightbody, A.A., Reiss, A.L., 2009. Gene, brain, and behavior relationships in fragile X syndrome: evidence from neuroimaging studies. Developmental disabilities research reviews 15, 343-352.

Lin, Q.H., Liu, J.Y., Zheng, Y.R., Liang, H.L., Calhoun, V.D., 2010. Semiblind Spatial ICA of fMRI Using Spatial Constraints. Human Brain Mapping 31, 1076-1088.

Linden, D.J., Connor, J.A., 1995. Long-term synaptic depression. Annual review of neuroscience 18, 319-357.

Littmann, A., Guehring, J., Buechel, C., Stiehl, H.S., 2006. Acquisition-related morphological variability in structural MRI. Academic radiology 13, 1055-1061.

Liu, J., Calhoun, V.D., Chen, J., Claus, E.D., Hutchison, K.E., 2013. Effect of homozygous deletions at 22q13.1 on alcohol dependence severity and cue-elicited BOLD response in the precuneus. Addict Biol 18, 548-558.

Liu, J., Ghassemi, M.M., Michael, A.M., Boutte, D., Wells, W., Perrone-Bizzozero, N., et al., 2012. An ICA with reference approach in identification of genetic variation and associated brain networks. Frontiers in human neuroscience 6, 21.

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N.I., Calhoun, V., 2009. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Human Brain Mapping 30, 241-255.

Liu, Q.S., Pu, L., Poo, M.M., 2005. Repeated cocaine exposure in vivo facilitates LTP induction in midbrain dopamine neurons. Nature 437, 1027-1031.

Logothetis, N.K., 2003. The underpinnings of the BOLD functional magnetic resonance imaging signal. Journal of Neuroscience 23, 3963-3971.

Lonze, B.E., Ginty, D.D., 2002. Function and regulation of CREB family transcription factors in the nervous system. Neuron 35, 605-623.

Lundstrom, B.N., Ingvar, M., Petersson, K.M., 2005. The role of precuneus and left inferior frontal cortex during source memory episodic retrieval. Neuroimage 27, 824-834.

Luo, X.G., Kranzler, H.R., Zuo, L.J., Wang, S., Schork, N.J., Gelernter, J., 2007. Multiple ADH genes modulate risk for drug dependence in both African- and European-Americans. Human Molecular Genetics 16, 380-390.

Lynall, M.E., Bassett, D.S., Kerwin, R., McKenna, P.J., Kitzbichler, M., Muller, U., et al., 2010. Functional connectivity and brain networks in schizophrenia. The Journal of neuroscience : the official journal of the Society for Neuroscience 30, 9477-9487.

Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S., et al., 2000. Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences of the United States of America 97, 4398-4403.

Malenka, R.C., Bear, M.F., 2004. LTP and LTD: An embarrassment of riches. Neuron 44, 5-21.

Mameli, M., Luscher, C., 2011. Synaptic plasticity and addiction: Learning mechanisms gone awry. Neuropharmacology 61, 1052-1059.

Manoach, D.S., 2002. Prefrontal cortex dysfunction during working memory performance in schizophrenia: Reconciling discrepant findings. Biological Psychiatry 51, 104s-104s.

Mattson, M.P., 1992. Calcium as Sculptor and Destroyer of Neural Circuitry. Experimental Gerontology 27, 29-49.

Mattson, M.P., Maudsley, S., Martin, B., 2004. BDNF and 5-HT: a dynamic duo in age-related neuronal plasticity and neurodegenerative disorders. Trends in Neurosciences 27, 589-594.

Mayfield, R.D., Harris, R.A., Schuckit, M.A., 2008. Genetic factors influencing alcohol dependence. British journal of pharmacology 154, 275-287.

McDonald, C., Marshall, N., Sham, P.C., Bullmore, E.T., Schulze, K., Chapple, B., et al., 2006. Regional brain morphometry in patients with schizophrenia or bipolar disorder and their unaffected relatives. American Journal of Psychiatry 163, 478-487.

Meda, S.A., Jagannathan, K., Gelernter, J., Calhoun, V.D., Liu, J.Y., Stevens, M.C., et al., 2010. A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia. Neuroimage 53, 1007-1015.

Minzenberg, M.J., Laird, A.R., Thelen, S., Carter, C.S., Glahn, D.C., 2009. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. Archives of General Psychiatry 66, 811-822.

Muhle, R., Trentacoste, S.V., Rapin, I., 2004. The genetics of autism. Pediatrics 113, E472-E486.

Mulligan, M.K., Ponomarev, I., Hitzemann, R.J., Belknap, J.K., Tabakoff, B., Harris, R.A., et al., 2006. Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. Proceedings of the National Academy of Sciences of the United States of America 103, 6368-6373.

Murray, L.J., Ranganath, C., 2007. The dorsolateral prefrontal cortex contributes to successful relational memory encoding. Journal of Neuroscience 27, 5515-5522.

Myrick, H., Anton, R.F., Li, X., Henderson, S., Randall, P.K., Voronin, K., 2008. Effect of naltrexone and ondansetron on alcohol cue-induced activation of the ventral striatum in alcohol-dependent people. Arch Gen Psychiatry 65, 466-475.

Narr, K.L., Bilder, R.M., Toga, A.W., Woods, R.P., Rex, D.E., Szeszko, P.R., et al., 2005. Mapping cortical thickness and gray matter concentration in first episode schizophrenia. Cerebral Cortex 15, 708-719.

Nelson, M.D., Saykin, A.J., Flashman, L.A., Riordan, H.J., 1998. Hippocampal volume reduction in schizophrenia as assessed by magnetic resonance imaging - A meta-analytic study. Archives of General Psychiatry 55, 433-440.

Nestler, E.J., 2001. Molecular basis of long-term plasticity underlying addiction. Nature Reviews Neuroscience 2, 119-128.

Nestler, E.J., 2002. Common molecular and cellular substrates of addiction and memory. Neurobiology of Learning and Memory 78, 637-647.

Nestler, E.J., 2005. Is there a common molecular pathway for addiction? Nature neuroscience 8, 1445-1449.

Noble, E.P., 2000. Addiction and its reward process through polymorphisms of the D2 dopamine receptor gene: a review. European Psychiatry 15, 79-89.

Ochsner, K.N., Knierim, K., Ludlow, D.H., Hanelin, J., Ramachandran, T., Glover, G., et al., 2004. Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. Journal of Cognitive Neuroscience 16, 1746-1772.

Oliphant, A., Barker, D.L., Stuelpnagel, J.R., Chee, M.S., 2002. BeadArray (TM) technology: Enabling an accurate, cost-effective approach to high throughput genotyping. Biotechniques, 56-+.

Olney, J.W., Farber, N.B., 1995. Glutamate Receptor Dysfunction and Schizophrenia. Archives of General Psychiatry 52, 998-1007.

Pandey, S.C., 1998. Neuronal signaling systems and ethanol dependence. Molecular Neurobiology 17, 1-15.

Pandey, S.C., 2003. Anxiety and alcohol abuse disorders: a common role for CREB and its target, the neuropeptide Y gene. Trends in Pharmacological Sciences 24, 456-460.

Pandey, S.C., Chartoff, E.H., Carlezon, W.A., Zou, J., Zhang, H.B., Kreibich, A.S., et al., 2005. CREB gene transcription factors: Rrole in molecular mechanisms of alcohol and drug addiction. Alcoholism-Clinical and Experimental Research 29, 176-184.

Park, M.S., Sohn, J.H., Suk, J.A., Kim, S.H., Sohn, S., Sparacio, R., 2007. Brain substrates of craving to alcohol cues in subjects with alcohol use disorder. Alcohol Alcohol 42, 417-422.

Paulsen, J.S., Heaton, R.K., Sadek, J.R., Perry, W., Delis, D.C., Braff, D., et al., 1995. The nature of learning and memory impairments in schizophrenia. J Int Neuropsychol Soc 1, 88-89.

Paus, T., Collins, D.L., Evans, A.C., Leonard, G., Pike, B., Zijdenbos, A., 2001. Maturation of white matter in the human brain: a review of magnetic resonance studies. Brain Research Bulletin 54, 255-266.

Pickens, R.W., Svikis, D.S., Mcgue, M., Lykken, D.T., Heston, L.L., Clayton, P.J., 1991. Heterogeneity in the Inheritance of Alcoholism - a Study of Male and Female Twins. Archives of General Psychiatry 48, 19-28.

Pierce, R.C., Bari, A.A., 2001. The role of neurotrophic factors in psychostimulant-induced behavioral and neuronal plasticity. Reviews in the Neurosciences 12, 95-110.

Pizzagalli, D.A., 2011. Frontocingulate dysfunction in depression: toward biomarkers of treatment response. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 36, 183-206.

Prata, D.P., Mechelli, A., Fu, C.H.Y., Picchioni, M., Toulopoulou, T., Bramon, E., et al., 2009. Epistasis between the DAT 3 ' UTR VNTR and the COMT Val158Met SNP on cortical function in healthy subjects and patients with schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 106, 13600-13605.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904-909.

Pujadas, L., Gruart, A., Bosch, C., Delgado, L., Teixeira, C.M., Rossi, D., et al., 2010. Reelin regulates postnatal neurogenesis and enhances spine hypertrophy and long-term potentiation. The Journal of neuroscience : the official journal of the Society for Neuroscience 30, 4636-4649.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al., 2007a. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559-575.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., et al., 2007b. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559-575.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., et al., 2009a. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., et al., 2009b. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.

Radel, M., Vallejo, R.L., Iwata, N., Aragon, R., Long, J.C., Virkkunen, M., et al., 2005. Haplotype-based localization of an alcohol dependence gene to the 5q34 gamma-aminobutyric acid type a gene cluster. Archives of General Psychiatry 62, 47-55.

Rapoport, J.L., Addington, A.M., Frangou, S., Psych, M., 2005. The neurodevelopmental model of schizophrenia: update 2005. Molecular Psychiatry 10, 434-449.

Rapoport, J.L., Giedd, J.N., Gogtay, N., 2012. Neurodevelopmental model of schizophrenia: update 2012. Molecular Psychiatry 17, 1228-1238.

Rasch, B., Papassotiropoulos, A., de Quervain, D.F., 2010. Imaging genetics of cognitive functions: Focus on episodic memory. Neuroimage 53, 870-877.

Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., et al., 2011. Genome-wide association study identifies five new schizophrenia loci. Nature

Genetics 43, 969-976.

Rissanen, J., 1978. Modeling by Shortest Data Description. Automatica 14, 465-471.

Ron, D., Jurd, R., 2005. The "ups and downs" of signaling cascades in addiction. Sci STKE 2005, re14.

Rose, E.J., Donohoe, G., 2013. Brain vs behavior: an effect size comparison of neuroimaging and cognitive studies of genetic risk for schizophrenia. Schizophrenia Bulletin Epub 39, 518-526.

Rutishauser, U., Jessell, T.M., 1988. Cell adhesion molecules in vertebrate neural development. Physiological reviews 68, 819-857.

Sastry, P.S., Rao, K.S., 2000. Apoptosis and the nervous system. Journal of neurochemistry 74, 1-20.

Schacht, J.P., Anton, R.F., Myrick, H., 2013. Functional neuroimaging studies of alcohol cue reactivity: a quantitative meta-analysis and systematic review. Addict Biol 18, 121-133.

Schmitz, B.L., Aschoff, A.J., Hoffmann, M.H., Gron, G., 2005. Advantages and pitfalls in 3T MR brain imaging: a pictorial review. AJNR. American journal of neuroradiology 26, 2229-2237.

Schumann, G., Johann, M., Frank, J., Preuss, U., Dahmen, N., Laucht, M., et al., 2008. Systematic analysis of Glutamatergic neurotransmission genes in alcohol dependence and adolescent risky drinking behavior. Archives of General Psychiatry 65, 826-838.

Scott, L.J., Muglia, P., Kong, X.Q., Guan, W., Flickinger, M., Upmanyu, R., et al., 2009. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proceedings of the National Academy of Sciences of the United

States of America 106, 7501-7506.

Segall, J.M., Allen, E.A., Jung, R.E., Erhardt, E.B., Arja, S.K., Kiehl, K., et al., 2012. Correspondence between structure and function in the human brain at rest. Frontiers in neuroinformatics 6, 10.

Segall, J.M., Turner, J.A., van Erp, T.G., White, T., Bockholt, H.J., Gollub, R.L., et al., 2009. Voxel-based morphometric multisite collaborative study on schizophrenia. Schizophrenia Bulletin 35, 82-95.

Shen, R., Fan, J.B., Campbell, D., Chang, W.H., Chen, J., Doucet, D., et al., 2005. High-throughput SNP genotyping on universal bead arrays. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis 573, 70-82.

Shenton, M.E., Dickey, C.C., Frumin, M., McCarley, R.W., 2001. A review of MRI findings in schizophrenia. Schizophrenia Research 49, 1-52.

Silva, A.J., Kogan, J.H., Frankland, P.W., Kida, S., 1998. CREB and memory. Annual Review of Neuroscience 21, 127-148.

Skalhegg, B.S., Tasken, K., 2000. Specificity in the cAMP/PKA signaling pathway. Differential expression, regulation, and subcellular localization of subunits of PKA. Frontiers in Bioscience 5, D678-D693.

Skinner, H.A., Horn, J.L., 1984. Alcohol Dependence Scale: Users Guide. Alcohol Research Foundation, Toronto, Canada.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208-219.

Spalice, A., Parisi, P., Nicita, F., Pizzardi, G., Del Balzo, F., Iannetti, P., 2009. Neuronal migration disorders: clinical, neuroradiologic and genetics aspects. Acta paediatrica 98, 421-433.

Stone, J.M., Day, F., Tsagaraki, H., Valli, I., McLean, M.A., Lythgoe, D.J., et al., 2009. Glutamate dysfunction in people with prodromal symptoms of psychosis: relationship to gray matter volume. Biological Psychiatry 66, 533-539.

Stonnington, C.M., Tan, G., Kloppel, S., Chu, C., Draganski, B., Jack, C.R., Jr., et al., 2008. Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. Neuroimage 39, 1180-1185.

Stranger, B.E., Stahl, E.A., Raj, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187, 367-383.

Strauss, K.A., Puffenberger, E.G., Huentelman, M.J., Gottlieb, S., Dobrin, S.E., Parod, J.M., et al., 2006. Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. New England Journal of Medicine 354, 1370-1377.

Sun, J., Maller, J.J., Guo, L., Fitzgerald, P.B., 2009. Superior temporal gyrus volume change in schizophrenia: a review on region of interest volumetric studies. Brain Research Reviews 61, 14-32.

Svenningsson, P., Nishi, A., Fisone, G., Girault, J.A., Nairn, A.C., Greengard, P., 2004. DARPP-32: an integrator of neurotransmission. Annual review of pharmacology and toxicology 44, 269-296.

Tapert, S.F., Brown, G.G., Baratta, M.V., Brown, S.A., 2004. fMRI BOLD response to alcohol stimuli in alcohol dependent young women. Addictive Behaviors 29, 33-50.

Thompson, P.M., Cannon, T.D., Narr, K.L., van Erp, T., Poutanen, V.P., Huttunen, M., et al., 2001. Genetic influences on brain structure. Nature neuroscience 4, 1253-1258.

Toescu, E.C., 1998. Apoptosis and cell death in neuronal cells: where does Ca2+ fit in? Cell Calcium 24, 387-403.

Tong, L., Soon, V.C., Huang, Y.F., Liu, R., 1990. Amuse - a New Blind Identification Algorithm. 1990 Ieee International Symp on Circuits and Systems, Vols 1-4, 1784-1787.

Treutlein, J., Cichon, S., Ridinger, M., Wodarz, N., Soyka, M., Maier, W., et al., 2010. Genome-Wide Association Study of Alcohol Dependence. Alcoholism-Clinical and Experimental Research 34, 43a-43a.

Truong, T.K., Chakeres, D.W., Beversdorf, D.Q., Scharre, D.W., Schmalbrock, P., 2006. Effects of static and radiofrequency magnetic field inhomogeneity in ultra-high field magnetic resonance imaging. Magnetic Resonance Imaging 24, 103-112.

Turner, J.A., Calhoun, V.D., Michael, A., van Erp, T.G., Ehrlich, S., Segall, J.M., et al., 2012. Heritability of multivariate gray matter measures in schizophrenia. Twin research and human genetics : the official journal of the International Society for Twin Studies 15, 324-335.

Turner, J.A., Smyth, P., Macciardi, F., Fallon, J.H., Kennedy, J.L., Potkin, S.G., 2006. Imaging phenotypes and genotypes in schizophrenia. Neuroinformatics 4, 21-49.

Uhl, G.R., 2004. Molecular genetic underpinnings of human substance abuse vulnerability: likely contributions to understanding addiction as a mnemonic process. Neuropharmacology 47, 140-147.

van Haren, N.E., Hulshoff Pol, H.E., Schnack, H.G., Cahn, W., Mandl, R.C., Collins, D.L., et al., 2007. Focal gray matter changes in schizophrenia across the course of the illness: a 5-year follow-up study. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 32, 2057-2066.

Vaughan, J.T., Garwood, M., Collins, C.M., Liu, W., DelaBarre, L., Adriany, G., et al., 2001. 7T vs. 4T: RF power, homogeneity, and signal-to-noise comparison in head images. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine 46, 24-30.

Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., et al., 2008. A functional genetic link between distinct developmental language disorders. The New England journal of medicine 359, 2337-2345.

Vounou, M., Nichols, T.E., Montana, G., Initia, A.D.N., 2010. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. Neuroimage 53, 1147-1159.

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. Ieee Transactions on Medical Imaging 26, 405-421.

Wahlbeck, K., Ahokas, A., Nikkila, H., Miettinen, K., Rimon, R., 2000. Cerebrospinal fluid angiotensin-converting enzyme (ACE) correlates with length of illness in schizophrenia. Schizophrenia Research 41, 335-340.

Waltereit, R., Weller, M., 2003. Signaling from cAMP/PKA to MAPK and synaptic plasticity. Molecular Neurobiology 27, 99-106.

Wand, G., Levine, M., Zweifel, L., Schwindinger, W., Abel, T., 2001. The cAMP-protein kinase A signal transduction pathway modulates ethanol consumption and sedative effects of ethanol. The Journal of neuroscience : the official journal of the Society for Neuroscience 21, 5297-5303.

Wang, J.C., Grucza, R., Cruchaga, C., Hinrichs, A.L., Bertelsen, S., Budde, J.P., et al., 2009. Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. Molecular Psychiatry 14, 501-510.

Wang, K., Abbott, D., 2008. A principal components regression approach to multilocus genetic association studies. Genetic Epidemiology 32, 108-118.

Wax, M., Kailath, T., 1985. Detection of Signals by Information Theoretic Criteria. Ieee Transactions on Acoustics Speech and Signal Processing 33, 387-392.

Wedenoja, J., Loukola, A., Tuulio-Henriksson, A., Paunio, T., Ekelund, J., Silander, K., et al., 2008. Replication of linkage on chromosome 7q22 and association of the regional Reelin gene with working memory in schizophrenia families. Molecular Psychiatry 13, 673-684.

Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S.G., Yu, Z., et al., 2011. SNP-based pathway enrichment analysis for genome-wide association studies. BMC bioinformatics 12, 99.

Wintersperger, B.J., Reeder, S.B., Nikolaou, K., Dietrich, O., Huber, A., Greiser, A., et al., 2006. Cardiac CINE MR imaging with a 32-channel cardiac coil and parallel imaging: impact of acceleration factors on image quality and volumetric accuracy. Journal of magnetic resonance imaging : JMRI 23, 222-227.

Wojnar, M., Brower, K.J., Strobbe, S., Ilgen, M., Matsumoto, H., Nowosad, I., et al., 2009. Association Between Val66Met Brain-Derived Neurotrophic Factor (BDNF) Gene Polymorphism and Post-Treatment Relapse in Alcohol Dependence. Alcoholism-Clinical and Experimental Research 33, 693-702.

Wong, C.C.Y., Schumann, G., 2008. Genetics of addictions: strategies for addressing heterogeneity and polygenicity of substance use disorders. Philosophical Transactions of the Royal Society B-Biological Sciences 363, 3213-3222.

Wrase, J., Schlagenhauf, F., Kienast, T., Wustenberg, T., Bermpohl, F., Kahnt, T., et al., 2007. Dysfunction of reward processing correlates with alcohol craving in detoxified alcoholics. Neuroimage 35, 787-794.

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25, 714-721.

Xu, L., Groth, K.M., Pearlson, G., Schretlen, D.J., Calhoun, V.D., 2009. Source-Based Morphometry: The Use of Independent Component Analysis to Identify Gray Matter Differences With Application to Schizophrenia. Human Brain Mapping 30, 711-724.

Xu, L., Liu, j., Adali, T., Calhoun, V.D., 2008. Source based morphometry using structural mri phase images to identify sources of gray matter and white matter relative differences in schizophrenia versus controls. International Conference on Acoustics, Speech, and Signal Processing (ICASSP),, Las Vegas, NV.

Yalachkov, Y., Kaiser, J., Naumer, M.J., 2010. Sensory and motor aspects of addiction. Behavioural brain research 207, 215-222.

Zhang, H., Kranzler, H.R., Yang, B.Z., Luo, X., Gelernter, J., 2008. The OPRD1 and OPRK1 loci in alcohol or drug dependence: OPRD1 variation modulates substance dependence risk. Molecular Psychiatry 13, 531-543.

Zhou, D., Lambert, S., Malen, P.L., Carpenter, S., Boland, L.M., Bennett, V., 1998. AnkyrinG is required for clustering of voltage-gated Na channels at axon initial segments and for normal action potential firing. The Journal of cell biology 143, 1295-1304.

Zhou, Y.A., Shu, N., Liu, Y., Song, M., Hao, Y.H., Liu, H.H., et al., 2008. Altered resting-state functional connectivity and anatomical connectivity of hippocampus in schizophrenia. Schizophrenia Research 100, 120-132.