KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING
FACULTY OF ENGINEERING SCIENCE
FACULTY OF MEDICINE
FACULTY OF SCIENCES

Multivariate Genotype-Phenotype Associations on Global-to-Local Cortical Brain Shape

Promoters:
Prof. Peter Claes
Department of Human Genetics
Department of Electrical Engineering (ESAT)
Division for Processing Speech and Images (PSI)
Laboratory for Imaging Genetics

Prof. Isabelle Cleynen
Department of Human Genetics
Laboratory for Complex Genetics

Dissertation presented in
fulfillment of the requirements
for the degree of Master of Science:
Bioinformatics

Seppe GOOVAERTS

June 2021

This dissertation is part of the examination and has not been corrected after defense for eventual errors. Use as a reference is permitted subject to written approval of the promotor stated on the front page.

Abstract

The genetic variation between individuals, caused for a large part by common single nucleotide polymorphisms (SNPs) has been found to attribute to complex diseases or traits in general. Genome-wide association studies (GWAS) aim to find genotype-phenotype associations through means of a statistical framework. Most commonly, SNPs are tested individually for association, however, such approach fails to detect SNPs with very small effect sizes. SNP-set GWAS aims to combine these small effects by simultaneously testing for association with multiple SNPs close together in the genome. Grouping SNPs leads to fewer tests thus alleviating the multiple testing burden and relaxing the significance threshold. In addition, highly complex phenotypes such as the brain or face are hard to describe univariately and benefit from a multivariate description. Canonical correlation analysis (CCA) offers a bi-multivariate statistical framework where multiple SNPs can be tested for association with such a multivariate phenotype. This work explored the potential and features of bi-multivariate, genome-wide, brain-wide association testing on cortical surface morphology in 19,643 individuals of European ancestry.

SNP-set GWAS were conducted based on gene-based, haplotype-based, and window-based grouping of SNPs, and found 120 genes, 124 loci, and 124 loci respectively influencing cortical surface morphology. By comparing the results to a recent GWAS by Naqvi et al. (2021) it was demonstrated that SNP-set GWAS is able to detect both known and novel associations. Window-based GWAS were conducted using a range of window sizes (5 – 200 kb) and showed that larger SNP-sets result in fewer associations. In addition, it was shown that there exists an optimal range of window sizes and thus group sizes for finding the loci that per-SNP GWAS fails to detect.

This thesis proposes an adapted implementation of the genomic control factor specifically suited for SNP-set association testing with the CCA framework. In addition, it was demonstrated that the genomic control factor can be used on pooled test statistics to investigate whether certain group sizes yield disproportionately inflated test statistics and as such result in bias. It was then demonstrated that a correction per pool can be used to adequately correct for ununiform inflation.

To further illustrate the potential of SNP-set GWAS, the coordinates of interogressed Neanderthal haplotype blocks were used as the grouping structure, and each block was subsequently tested for association with brain shape. In total 6 blocks exceeded the most stringent significance threshold, suggesting that SNPs within these blocks could potentially affect the brains of present-day humans.

Abbreviations

AD	Alzheimer's disease
AMD	age-related macular degeneration
BMI	body mass index
bp	base pair
CCA	canonical correlation analysis
CNS	central nervous system
df	degrees of freedom
FDR	false discovery rate
GO	gene ontology
GPA	Generalized Procrustes analysis
GW	genome-wide
GWAS	genome-wide association study
HLA	human leukocyte antigen
IBS	identical by state
kb	kilobases
LD	linkage disequilibrium
lincRNA	long intergenic non-coding RNA
MAF	minor allele frequency
MDD	major depressive disorder
microRNA	miRNA
MRI	magnetic resonance imaging
PC	principal component
PCA	principal component analysis
PD	Parkinson's disease
QT	quantitative trait
RA	rheumatoid arthritis
SNP	single nucleotide polymorphism
SW	study-wide
vGWAS	voxel-based genome-wide association study
VIF	variance inflation factor

List of Figures

List of Tables

Table of Contents

1 STUDY OF LITERATURE

1.1 Genotype-phenotype associations

Finding and identifying the genes involved in common diseases has been a challenge for genetic epidemiologists for decades. Finding causal genes for monogenic disorders or traits can be done via linkage analysis in pedigrees of affected families (Guo & Lange, 2000). This approach makes use of genetic markers to point to specific loci in the genome. If a certain marker is persistently co-inherited with the disease allele, both are likely in linkage disequilibrium (LD). This means that there have been relatively few recombination events between the markers. Since the probability of recombination is proportional to the distance between two markers, the persistent co-inheritance of two markers suggests close genomic proximity. Although this strategy has proven to be useful for monogenic diseases, most diseases, or traits in general, however, are highly polygenic in nature, which makes linkage analysis less efficient (Guo & Lange, 2000).

Single nucleotide polymorphisms (SNPs) are variations in the DNA sequence consisting of a single nucleotide alteration that are common in the population. Here ‘common’ means that the SNP has a minimal minor allele frequency (MAF) of 1%, or in other words that the less common variant occurs in at least 1% of the population (Kruglyak & Nickerson, 2001). Any two human genomes are approximately 99.9% identical, with most of the remaining variation due to SNPs (Kruglyak & Nickerson, 2001). It is estimated that the human genome has around 11 million SNPs (H. Shen et al., 2013).

The functional implications of SNPs can vary according to their location in the genome (Haraksingh & Snyder, 2013). For example, if a SNP is located in a gene, it has the potential to change the amino acid sequence of the corresponding protein and create a structural variant with altered properties. SNPs located within a coding region of a gene (i.e., an exon) are classified as coding variants. This category of SNPs includes missense mutations, which alter a codon (i.e., 3 consecutive base pairs that encode a single amino acid) such that a different amino acid is incorporated in the corresponding protein; and nonsense mutations, which change a codon into a stop codon, leading to early transcription termination and thus a truncated protein. Missense mutations in codons of interaction domains, i.e., protein-substrate (e.g., active site), protein-protein (e.g., cleavage site), or protein-DNA (e.g., DNA binding domain) interaction domains, especially impact a proteins function. Non-coding variants include SNPs in regulatory domains of genes, as well as intronic and intergenic SNPs. SNPs in regulatory domains have the potential to alter gene regulation and as such affect protein abundance. For intronic and intergenic SNPs, the exact mechanisms of action are often less well understood, yet they are the most represented category of SNPs in known genotype-phenotype associations (Zou et al., 2020).

Technological advances in genotyping have made collecting SNP data low cost and high throughput to the point where over 800,000 SNPs can be genotyped on a single genotyping array (Bycroft et al., 2018). These arrays contain allele-specific DNA probes which hybridize with fragmented donor DNA bound to a fluorescent label. A detection system then measures and interprets the fluorescent signal corresponding to each SNP allele. After quality control of

the collected SNP signals, and optionally, imputation of even more SNPs to a reference panel, genotypes are ready to be used for genome-wide association studies (GWAS).

The advances in genotyping technology have been at least in part driven by GWAS, which try to find genotype-phenotype associations between hundreds of thousands to millions of SNPs and a complex disease or trait. The statistical framework of GWAS consists of a model of inheritance, usually the additive genetic model, where SNPs are encoded by allelic dosage (the number of times the reference allele is present in an individual's genome), as well as a regression model. Given that diseases are generally dichotomous phenotypes, a logistic regression model is often applied on genomic data of cases and controls, i.e., affected and unaffected individuals respectively. For complex quantitative traits a univariate regression model is primarily used. In both cases SNPs are regressed individually against the phenotype of interest which requires a stringent threshold of significance and adequate correction for the number of tests as to correctly control for false discoveries. Often a genome-wide significance threshold of $P < 5 \times 10^{-8}$ is used in GWAS on European populations, which has been found to be an adequate choice in terms of false discovery rate regardless of the number of SNPs (Dudbridge & Gusnanto, 2008). However, the use of such stringent thresholds may hinder the discovery of trait-relevant loci.

GWAS have revolutionized the field of complex genetics, and since the first GWAS on age-related macular degeneration (AMD) in 2005 (Klein et al., 2005), over 5000 studies have been published identifying over 250,000 significant associations ($P < 5 \times 10^{-8}$) with diverse phenotypes (Buniello et al., 2019). Figure 1 shows the exponential increase in GWAS publications since 2007 based on data from the NHGRI-EBI GWAS Catalog (Mills & Rahal, 2019). In addition, large, rich datasets are now available to researchers, e.g., the UK Biobank project, which offers information on many phenotypes and health related variables for approximately 500,000 individuals (Bycroft et al., 2018).

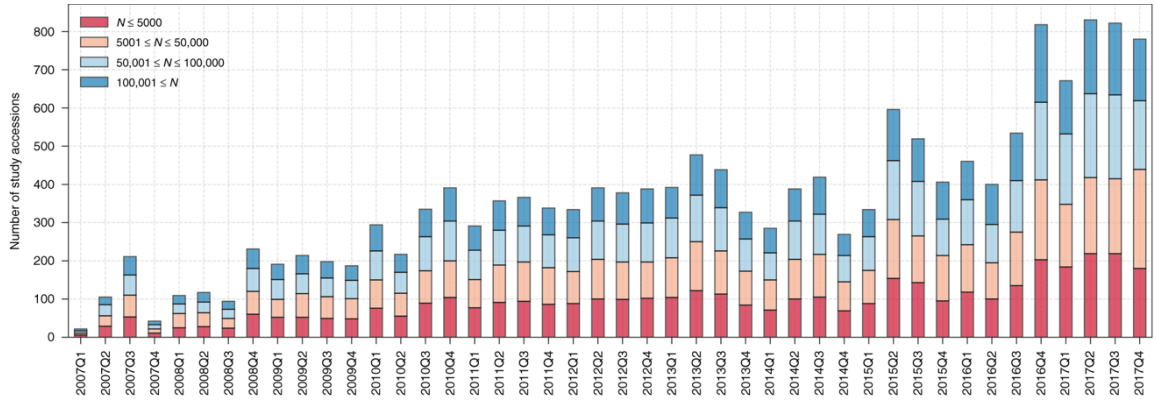


Figure 1 Trend in the number of GWAS publication (2007 – 2017) (Mills & Rahal, 2019). Bars represent the number of GWAS per quartal between 2007 and 2017. Colors indicate the number of individuals in the study: ≤ 5000 (red); 5001 – 50,000 (tan); 50,001 – 100,000 (light blue); and ≥ 100,001 (dark blue). A growing trend is observed for both small and large studies.

GWAS have been successful in identifying many risk loci for a plethora of complex diseases and traits, including inflammatory bowel disease (de Lange et al., 2017), schizophrenia (Li et al., 2017), cancer (Sud et al., 2017), educational attainment (Lee et al., 2018), and more. Often, disease-associated loci implicate genes with unknown function or genes previously thought to be unrelated to the trait (Hirschhorn, 2009). Those genes especially can give new valuable insights in biological mechanisms underlying the disease (Hirschhorn, 2009). For example, the

role of autophagy in Crohn's disease was only shed light upon after the discovery of two risk alleles in genes *ATG16L1* and *IRGM* through GWAS (Murthy et al., 2014). The *rs2241880* SNP causes a missense mutation near a caspase-3 cleavage domain in *ATG16L1*, which significantly sensitizes it to caspase-3-mediated processing and diminishes autophagy under cellular stress; in turn leading to reduced intracellular bacterial clearance and increased production of inflammatory cytokines (Murthy et al., 2014).

Though most GWAS have been conducted on univariate traits, in recent years the approach has been successfully applied to multivariate complex phenotypes, including the face and the brain (Claes et al., 2018; Naqvi et al., 2021), which are hard to fully describe univariately. While for diseases, the effect sizes of disease-associated alleles can be used in the calculation of a polygenic risk score that measures the inherent genetic risk of developing the disease; for faces and brains, the effect sizes can be used for the prediction of face and brain shape from one's DNA. Interestingly, combining information from both face and brain GWAS has led to support for the hypothesis that face and brain mutually shape each other during early embryogenesis through both structural effects and paracrine signaling (Naqvi et al., 2021). Many shared brain-face loci included genes that encode transcription factors involved in neural crest formation and/or craniofacial skeletal development.

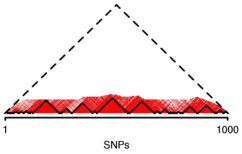
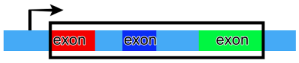
Beside generating valuable insights in molecular mechanisms, GWAS have also revealed insights in the genetic lay-out of complex traits. It appears that the vast majority of loci affecting complex phenotypes are common variants with weak effect sizes, and hence large samples sizes are essential to detect these effects (Génin, 2020). Still, in the case of human height, although numerous large scale GWAS have been conducted, using data of several hundreds of thousands of individuals, common variants identified so far only explain ~25% of the estimated ~80% of genetic heritability (Marouli et al., 2017). The remaining ~55%, or so-called missing heritability has not been detected by GWAS yet. It is suggested that at least some of the missing heritability can be explained by even smaller effects from many more common SNPs, or possibly larger effects from uncommon ($0.5\% < MAF < 5\%$) and rare variants ($MAF < 0.5\%$), as well as gene-gene interactions, all of which pose statistical challenges (Manolio et al., 2009). More recently, it was proposed that the genetic lay-out of common diseases is omnigenic, rather than polygenic (Boyle et al., 2017). It is argued that if gene-regulatory networks are sufficiently interconnected, all genes within disease-relevant cells can affect the function of core disease-related pathways (Boyle et al., 2017). Either way, continuing to look for common variants with small effect sizes is probably a good way forward, preferably with additional, new approaches (Tam et al., 2019).

1.2 SNP-set association testing

One approach to detect loci with small effects on a phenotype is to group SNPs in SNP-sets, and jointly analyze their effect. The idea behind this technique is that small effect sizes are combined, and that the combined influence on the phenotype is more easily detectable. Another major advantage is the decrease in the number of statistical tests, referred to as multiple testing burden, which allows SNP-set GWAS to apply a less stringent significance threshold in comparison with single SNP testing. Both these properties combined make SNP-set GWAS a promising approach to find new, previously undiscovered loci affecting complex traits and disease.

Dividing the genome in relevant blocks, i.e., determining the SNP grouping structure can be done in numerous ways. Most popularly, SNPs are grouped based on genes (Buil et al., 2009; Liu et al., 2010; Tang & Ferreira, 2012; Wu et al., 2010), haplotype blocks (Guinot et al., 2018; Hamazaki & Iwata, 2020), or a sliding window (Braz et al., 2019). These three methods all have their benefits as well as drawbacks. So far, to the best of my knowledge, no paper has been published directly comparing different grouping methods on the same data set. An overview of these grouping methods is given in Table 1.

Table 1 Overview of grouping strategies. Left to right: Sliding window – Haplotype blocks – Genes. Pros and cons are listed for each method.

		
Sliding Window	**Haplotype Blocks**	**Genes**
Pros: Versatile, window size easily adjustable More uniform group size	**Pros:** Captures LD well Somewhat flexible: algorithm, parameters LD blocks estimated on data (if large dataset)	**Pros:** Interpretable units Functional grouping Captures LD well
Cons: Naivest approach	**Cons:** Some SNPs not in block Less uniform group size (bias)	**Cons:** No one reference list, possible redudancy Not all SNPs in genes Less uniform group size (bias)

1.2.1 Gene-based grouping

Gene-based GWAS consider associations between a trait and all SNPs within a single gene. The idea is that if a gene contributes to a trait or disease, then not only large, but also many small alterations in the amino acid sequence or regulatory domains of the gene have the potential to affect the respective genetic pathways and hence affect the phenotype. Some of these effects are so small that they are very hard to detect, even with large sample sizes. If a trait-relevant gene has only SNPs with very mild effects, individual SNP analysis will likely fail to detect this gene as no strong signal will be generated. Yet, the gene may play an essential role in the phenotype and discovering its association to the phenotype could lead to valuable biological insights. Moreover, the results of a gene-based GWAS immediately point to target genes, hinting at relevant pathways.

Intuitively it makes sense to use genes as units for association testing, as they are also inherited as a unit and intragenic SNPs are strongly correlated with one another. As a main advantage over individual SNP GWAS, testing associations for the estimated 20,000 to 25,000 genes in the human genome requires a far less stringent significance threshold compared to testing 0.5 to 10 million individual SNPs (Salzberg, 2018).

One caveat to using gene-based SNP sets, however, is that it excludes intergenic SNPs from the association analysis. Considering that approximately 90% of the identified trait-associated SNPs are non-coding, of which about half have been detected in intergenic regions, this is a major drawback (Zou et al., 2020). Alternative gene-based grouping strategies have been proposed that partially solve this issue. For example, certain gene annotation lists, such as the UCSC RefSeq gene list include known long intergenic non-coding RNAs (lincRNAs) which are

known to harbor many of the intergenic trait-associated SNPs (Zou et al., 2020). Additionally, SNPs within a gene's close proximity can be included in the SNP-set. A window of 15kb upstream and downstream of each gene has been used in other gene-based GWAS (Tang & Ferreira, 2012), which comprises 90% of SNPs affecting gene expression levels (Pickrell et al., 2010). Other points of attention related to gene-based grouping include the fact that no single reference list exists, and that most gene lists have some level of redundancy (i.e., overlapping gene annotations) that should be handled appropriately as to not unnecessarily inflate the number of statistical tests (Salzberg, 2018). Lastly, the substantial differences in size between genes are a possible source of bias depending on the association model.

1.2.2 Haplotype block-based grouping

Haplotype block-based SNP grouping is based on the idea of dividing the genome in blocks of strongly correlated SNPs, or haplotype blocks. These are discrete regions of high LD spanning a few to several hundred thousand kilobases for which little evidence for historical recombination events has been detected, resulting in the non-random assortment of alleles in such region (Wall & Pritchard, 2003). Generally, haplotype blocks correspond to only around 2 – 4 common haplotypes. Haplotype blocks are on average 5 – 20 kb in size, however, several very large blocks exist, such as an 804 kb block on chromosome 22 (Dawson et al., 2002). Such large blocks suggest recombination cold spots, or regions in which recombination rarely occurs. Conversely, there are also recombination hot spots, where recombination frequently occurs. SNPs around such regions are in near complete linkage equilibrium (i.e., the occurrence of one SNP is independent of another SNP), and as a consequence these regions lack discrete haplotype blocks (Wall & Pritchard, 2003).

Haplotype blocks are widely used in SNP-set GWAS because they capture the inherent correlational structure of the human genome, and unlike gene-based grouping, they cover most of the human genome. Coordinates of haplotype blocks may be obtained from the HapMap Project, which has created a genome-wide map of LD and haplotype blocks (Frazer et al., 2007). Alternatively, haplotype blocks can also be estimated based on the data itself using a measure of LD such as D' or r^2, using publicly available software packages such as PLINK 1.9 (Purcell et al., 2007) or Haploview (Barrett et al., 2005). Different algorithms exist, with parameters that can be adjusted, allowing for some flexibility and tuning possibilities. Due to the inherent property of LD calculations to give noisy results, the resulting blocks are moderately influenced by both the data and the algorithm (Kim et al., 2018). Because haplotype blocks capture the inherent correlation structure of the human genome, they are an excellent candidate for SNP-set GWAS, reducing the dimensionality of the genotype data, and thus the number of statistical tests in a well-informed manner. However, some SNPs do not correspond to any block, and as with gene-based GWAS, heterogeneous group sizes are a potential source of bias.

1.2.3 Sliding window-based grouping

A third popular grouping strategy involves a sliding window of a fixed size. Instead of looking at the inherent genetic structure of the human genome, this approach naively scans each chromosome and merges SNPs into a group if they fit within the window. A major advantage is that it guarantees that each SNP is assigned to a group. It is also much more flexible than other methods in that the size of the window can be freely chosen. In addition, the fixed window

size results in a more uniform distribution of SNPs between groups, which could provide statistical benefits depending on the choice of model. To date, no optimal size has been determined, neither in terms of kilobases nor in terms of SNP count, however, studying the influence of varying the window size could lead to new insights in SNP-set association methods in general.

1.2.4 The current state of SNP-set GWAS

A first class of methods for SNP-set association studies used results from single-SNP analysis, and aimed to combine *P*-values of SNPs within a group using an appropriate aggregation method. An example is the *set-based test* in PLINK (Purcell et al., 2007). The algorithm first converts *P*-values of individual SNPs into a χ^2 test statistic with 1 degree of freedom (df) and corrects these values by dividing by the genomic inflation factor, lambda. The score for each group is set to the average χ^2 statistic of the group's most significant SNPs (up to five). Permutation testing is used to determine the empirical null distribution of scores. In this way, LD structure, group size, and other confounding factors are accounted for. Although conceptually simple, and flexible in terms of grouping structure, the need for permutation testing comes with high computational requirements, restricting it from being used on a genome-wide level.

Based on the same idea of combining results from single-SNP analysis, Liu et al. (2010) proposed a new gene-based association method, that does not rely on permutation testing to determine group significance. Instead, their method makes use of Monte-Carlo simulations, which is much faster. For each gene, test scores are simulated based on reference LD data from e.g., the HapMap Project. The empirical gene-based *P*-value is the proportion of simulated test statistics that exceeds the observed gene-based test statistic. This test statistic is the sum of all χ^2 statistics within a gene, which were obtained from converting *P*-values into χ^2 statistics with 1 df. The method was demonstrated on a melanoma dataset for ~1300 cases and controls, achieving similar results as the PLINK set-based test in only a fraction of the computation time. Both methods could detect signals from loci for which single-SNP analysis failed. In addition, since only summary level data is required, i.e., single-SNP *P*-values, this type of approach can be easily applied to already published, and freely available SNP-by-SNP GWAS data as a complementary analysis.

Buil et al. (2009) proposed a new gene-based method for GWAS and applied their method on rheumatoid arthritis (RA) data for 868 cases and 1194 controls. The method consists of three steps. First, the genetic similarity between individuals is estimated based on the SNP genotypes in a given gene. Gower distance (Gower, 1971) is used as a similarity measure, computed as the average of the SNP-wise dissimilarity (0: same base pair; 1: different base pair) over a gene. Second, individuals are clustered into groups based on genetic similarity of the gene. Spectral clustering was used to divide individuals in three groups. Lastly, a logistic regression model was used to find for each gene if any of the three clusters was significantly associated with the disease. Genome-wide gene-based associations were compared to SNP-by-SNP associations. Both approaches detected significant hits in the human leukocyte antigen (HLA) region, known to be strongly associated with RA. Outside this region, both methods detected different genes, making both methods complementary.

Another prominent class of SNP-set methods are kernel-based methods. Wu et al. (2010) proposed a gene-based logistic kernel machine test. Kernel-based methods are popular for

SNP-set analysis because of their inherent property to model non-linear SNP effects, as well as SNP-SNP interaction (epistasis). In addition, confounders and population structure can be included in the model and are naturally adjusted for. Linear kernels had overall good performance, and (weighted) identical-by-state (IBS) kernels were well-suited to capture epistatic effects, but lost power when the underlying signal was linear. The effective degrees of freedom for each gene are estimated by accounting for LD between SNPs. The model was applied to breast cancer data, as well as simulated data, and was found to have good power and low false discovery rate (FDR).

Tang and Ferreira (2012) proposed a gene-based test of association using canonical correlation analysis (CCA). CCA aims to find associations between two sets of multivariate variables by maximizing the correlation between their linear combinations (Hotelling, 1936). Earlier, Ferreira and Purcell (2009) proposed CCA to find associations between a single SNP and multiple phenotypes simultaneously. In the case of a single phenotype or individual SNP analysis, CCA is equivalent to multiple regression, and inversed multiple regression respectively. Tang and Ferreira (2012) tested the performance of their CCA model on both simulated data and white blood cell count data of 1061 individuals, and both for individual phenotypes as all phenotypes simultaneously. They conclude that CCA provides a robust and computationally efficient method for finding multivariate associations and that its power is comparable to permutation-based tests. Although CCA is naturally able to find association between multiple SNPs and multiple phenotypes, it is unable to take confounders into account. As such, they should be adjusted for prior to association testing, e.g., by partial least square regression (Naqvi et al., 2021).

Most SNP-set GWAS approaches fall into one of above discussed categories. Many publications focus on gene-based grouping of SNPs despite the major disadvantage of not including intergenic SNPs. Hence, the full potential of SNP-set GWAS is yet to be further explored by the use of other grouping methods. Moreover, most SNP-set GWAS methods available can be adapted to implement any grouping method of choice. A special interest goes out to bi-multivariate association models, such as CCA, as they can naturally deal with multivariate phenotypes as well as sets of SNPs. Those methods especially are promising candidates for finding previously unknown genotype-phenotype associations in highly complex phenotypes such as the brain or face, which are hard to fully describe univariately (Naqvi et al., 2021).

1.3 Studying the human brain through brain imaging genomics

The brain is arguably the most complex and least-well understood human organ. Early on, insights on the functionality of brain were obtained from patients with brain injuries or having received lobectomy (removal of a part of the brain). Nowadays, non-invasive imaging technologies, such as magnetic resonance imaging (MRI) are available to study brain structure and function. In addition, brain and brain-related phenotypes can be related to genotypic data through association studies.

The effects of several neurological and psychiatric disorders can be seen on MRI data, including Alzheimer's disease (AD), Parkinson's disease (PD), autism, schizophrenia, bipolar disorder, and major depressive disorder (MDD) (Elliott et al., 2018). Thus, MRI data can provide intermediate or endophenotypes to assess the genetic architecture of such disorders.

In particular, GWAS provide an excellent tool to find loci responsible for brain-related disorders and to help identify the underlying processes. Furthermore, longitudinal studies using brain imaging data can help reveal genes and pathways involved in brain development.

1.3.1 Anatomy of the human brain

The brain and spinal cord together make up the central nervous system (CNS). Brain structure is documented in numerous brain atlases, amongst which the Desikan-Killiany atlas (Desikan et al., 2006). Functionally, the brain comprises of three main components: the cerebrum, cerebellum (little brain) and brainstem (Figure 2) (Nowinski, 2011).

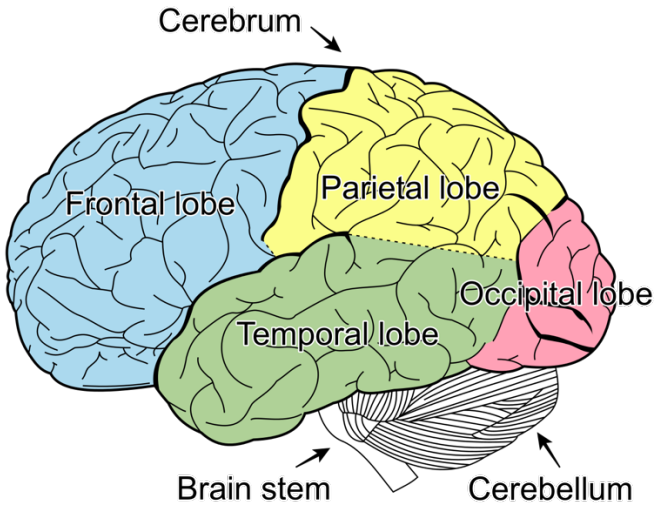


Figure 2 The three main components of the brain, and lobes of cerebrum (excluding the limbic lobe) viewed laterally (Henry Gray (1918) *Anatomy of the Human Body*). The three main components of the brain: cerebrum, cerebellum, and brain stem are indicated with arrows. Lobes are indicated in color: frontal (blue), parietal (yellow), temporal (green), and occipital lobe (red).

The cerebrum, the largest part of the human brain comprises left and right hemispheres and the diencephalon, which is the interbrain between the brainstem and the cerebrum. The cerebral hemispheres are composed of outer gray matter, termed the cerebral cortex; and inner gray matter, encompassing deep gray nuclei. The gray matter in the cortex has many neuronal cell bodies present, while the white matter is made up predominantly of nerve fibers (axons). The large number of neurons in the gray matter enable individuals to process information, while the axons in the inner white matter transmit signals to other parts of the cortex, making complex thoughts and high-level functions possible. The structure of the cortex consists of folds, termed gyri, that are separated by grooves, termed sulci, or fissures, which are deep sulci. These features increase the surface area of the cortex and hence the number of neurons, thereby also increasing the brain's processing power (Nowinski, 2011).

Fissures segment each brain hemisphere into five lobes (Figure 2): the frontal, parietal, occipital, temporal and limbic lobe. Although this classification was initially purely anatomical, each lobe was later shown to be responsible for specific functions:

– The frontal cortex, i.e., the outer gray matter of the frontal lobe is involved in many movement-related processes, including skeletal movement, ocular movement, speech and the expression of emotions (Astafiev et al., 2003; Kotz & Paulmann, 2011). Other functions of the frontal cortex include language processing, social behaviors, short-

term memory, attention, planning and decision making (Kotz & Paulmann, 2011). Many of these functions are related to dopaminergic neurons and pathways which have the largest presence in the frontal cortex (Astafiev et al., 2003).

- The parietal cortex' main function is the localization of sensory information from different parts of the body, including the processing of touch (Culham & Valyear, 2006). In addition, it is involved is spatial processing of visual information, such as shape, size and (relative) positional information (Astafiev et al., 2003; Culham & Valyear, 2006). Problems related to the parietal cortex include, spatial disorientation, dyslexia and a reduced ability to solve mathematical problems (Astafiev et al., 2003; Culham & Valyear, 2006).
- The occipital cortex is the visual processing center of the brain, containing the functionally significant primary visual cortex. This part of the cortex is involved in color differentiation and movement tracking (Fraser et al., 2011). Damage to the occipital cortex may cause visual field defects, that are complete or partial, and cortical blindness in the case of bilateral lesions (Fraser et al., 2011).
- The temporal cortex is involved in long-term memory formation. It contains the primary auditory cortex and transforms auditory information into interpretable units such as words (Lech & Suchan, 2013). In addition, high-level visual processing takes place in the temporal cortex, such as recognition of faces (Lech & Suchan, 2013). Clinical problems associated with the temporal cortex include dyslexia, amnesia, visual agnosia (impaired identification of familiar objects), word deafness and complex hallucinations (Lech & Suchan, 2013). Hallucinations in schizophrenia patients, often auditory, are linked to alterations in the temporal cortex (Jardri et al., 2011).
- The limbic cortex, mainly studied in the context of the limbic system, handles emotional response and adaptive behavior, as well as motivation and will (Heimer & Van Hoesen, 2006). In addition, the limbic cortex plays a role in self-awareness and the sense of ego (Heimer & Van Hoesen, 2006). The limbic lobe has several connections to the hypothalamus, which serves to maintain homeostasis (Heimer & Van Hoesen, 2006).

The cerebellum is a separate brain structure near the back of the brain, and under the cerebral hemispheres (Mauk et al., 2000). Just like the cerebrum, the cerebellum is composed of two hemispheres, and has a tightly folded cortex made up of gray matter (Mauk et al., 2000). It plays a major role in motor control, and contributes to precision, coordination, and timing of movement, as well as motor learning (Mauk et al., 2000). Cerebellar damage may result in disorders in fine movement, equilibrium and posture (Mauk et al., 2000).

The brainstem is subdivided into midbrain, pons, and medulla (Parvizi & Damasio, 2001). Given its position between the spinal cord and the rest of the brain, it naturally plays a role in conduction, i.e., information from the body traverses the brainstem and vice versa (Parvizi & Damasio, 2001). In addition, it has integrative functions in the body such as cardiovascular system control and respiratory control; processing pain and temperature signals; and regulating wakefulness, consciousness, and attention (Parvizi & Damasio, 2001).

1.3.2 Approaches and directions in GWAS studies of the brain

Brain imaging genomics aims to examine the associations between genetic markers, most notably SNPs, and imaging-derived quantitative traits (QTs). Research approaches to date differ greatly, both in terms of phenotypes extracted from imaging data, as well as in the choice whether or not to simultaneously analyze multiple markers, phenotypes, or both.

Early association studies mostly used univariate models to study genotype-phenotype associations, and thus extracted a single phenotype from the imaging data or analyzed phenotypes separately. Typically, single voxels (i.e., 3D pixels) (Stein et al., 2010); or cortical thickness (L. Shen et al., 2010), surface area, density (L. Shen et al., 2010), or volume measurements (L. Shen et al., 2010) were extracted from MRI data and were tested for associations with individual SNPs using a univariate regression model. This short review of brain imaging genomics methods (focusing on GWAS) aims to illustrate the importance of adequate phenotype extraction methods and the advantage of simultaneously analyzing multiple phenotypes.

Stein et al. (2010) proposed a voxel-based GWAS (vGWAS) approach to find associations between 448,293 SNPs and 31,622 voxels from calibrated MRI images of 740 elderly individuals, some of which had Alzheimer's disease. Per voxel, a linear regression approach was used to detect associated SNPs genome-wide. Then, the best *P*-value was taken for each voxel, and an adequate multiple testing threshold was determined using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995), mathematically known to be less stringent than the often-used Bonferroni method. No SNPs passed the threshold, which can be attributed to the relatively low sample size for such a study design and the extreme number of tests (448,293 x 31,622) performed. Furthermore, the heavy computational requirements for this approach make it unattractive at best, and unfeasible otherwise, especially for the larger sample sizes in recent studies.

Shen et al. (2010) investigated the genetic effects on 56 volumetric and cortical thickness measurements and 86 local gray matter densities across the brains of 818 individuals. Linear regression was applied to each of the 142 phenotypes separately. Results per phenotype were reported at two levels of significance ($P < 10^{-6}$ and $P < 10^{-7}$), which were somewhat arbitrarily chosen. A handful of SNPs were found to have associations with the phenotypes, including SNPs already known to be related.

Despite the success of recent, large-scale GWAS on average cortical thickness (Grasby et al., 2020), such single-measurement QTs are not able to capture the full complexity of cortical features and thus, in a way, leave much on the table in terms of the discovery of new brain-related loci (L. Shen et al., 2010). On the other hand, voxel-wise approaches include local information on the entire brain or cortex with the idea that as little as possible information is lost (Stein et al., 2010). Yet, the high number of tests required in such a setting requires a very high significance threshold in order to control for false discoveries. Both approaches can be thought of as different ends of the spectrum: one represents the brain by just a single feature, the other considers as many as possible. Ideally, phenotype extraction approaches find a balance between the amount of information extracted from imaging data and the dimensionality of the resulting phenotype. Logically, this translates to finding an optimal number of maximally informative features. Alternatively, phenotypes can be analyzed simultaneously using a multivariate testing approach, which offers a flexible association framework whilst reducing the number of statistical tests.

Naqvi et al. (2021) performed a GWAS on 19,644 individuals from the UK Biobank using phenotypes extracted from MRI scans via a state-of-the-art, data-driven phenotyping method developed for complex, multidimensional traits (Claes et al., 2018). Their method segments the brain in a hierarchical global-to-local manner based on similarity-based clustering of spatially dense 3D vertices of the mid-cortical surface mesh. Principal component analysis (PCA) on each segment results in multivariate phenotypes in the shape space, describing the

between-individual variance. CCA was used to find associations between a single SNP and the multivariate phenotype for each segment. It does this by taking a linear combination of the multidimensional phenotype, which can be considered as extracting a latent phenotype in the shape space. A total of 472 loci were detected affecting brain shape at different hierarchical levels. Their phenotyping method stands out from other phenotyping methods in that is able to capture the complex morphological features of the cortex, unlike single-measurement phenotypes like cortical thickness.

Besides varying greatly in their strategy of phenotype extraction, brain imaging genomic studies also vary in whether or not they analyze multiple phenotypes, SNPs, or both simultaneously. CCA and its extensions have been playing a prominent role recently in brain imaging genomics (L. Shen & Thompson, 2020), mostly because they naturally provide a framework to analyze highly multivariate phenotypes characteristic of complex traits such as cortical shape. In addition, they provide a framework for bi-multivariate association testing, i.e., finding associations between multiple phenotypes and a set of SNPs. As mentioned before, this is an effective strategy to 1) reduce the number of statistical tests, and 2) combine small contributions from individual SNPs, making such an approach attractive for finding new loci affecting brain shape. Yet, most studies applying this idea restrict themselves to only a small number of candidate genes, or only several regions of interest on the brain (Du et al., 2020; J. Kim et al., 2016). As such, the characteristics of bi-multivariate, genome-wide, brain-wide association testing are yet to be fully explored.

1.4 Research approach and aims

This work aims to explore the potential and features of bi-multivariate, genome-wide, brain-wide association testing on mid-cortical brain shape, extracted from MRI data using a state-of-the-art approach by Claes et al. (2018). This is done on MRI scans and ~10 million SNPs for 19,643 individuals from the UK Biobank. The effect of different grouping strategies will be evaluated and compared. Concretely, SNPs will be grouped based on 1) a sliding window, 2) haplotype blocks calculated on the dataset using PLINK 1.9, and 3) UCSC RefSeq genes. Moreover, recent work from Naqvi et al. (2021), who used the same UK Biobank data in a per-SNP GWAS setting, allows for comparison between both approaches in terms of the number of (new) loci. In addition, using a sliding window grouping approach with different window sizes, the effect of group size will be examined.

2 MATERIALS AND METHODS

So far, bi-multivariate GWAS are not widely adopted as an approach to association testing. Hence, in contrast to regular GWAS, many software tools, notably PLINK 1.9 (Purcell et al., 2007), as well as other analysis methods are lacking or are not yet adapted for a bi-multivariate analysis. As a result, most analysis were performed using newly developed scripts in MATLAB 2020b.

2.1 Data and preprocessing

The data was originally obtained from the UK Biobank, which encompasses ~500,000 British volunteers. T1-weighted MRI images, as well as imaging variables, individual covariates (age, weight, blood pressure, etc.), and genotype data were acquired from the UK Biobank v1.5 release (2018). This cohort was composed of 21,780 individuals, of which 51.6% female, and 48.4% male, with a mean age of 60 (range 40 – 70), a mean body mass index (BMI) of 26.6, and a predominantly white British ancestry (97.1%). Genotyping was done on the Applied Biosystems™ UK Biobank Axiom™ Array from ThermoFisher Scientific at 825,927 markers with DNA extracted from blood. SNP genotypes were imputed to the Haplotype Reference Consortium and merged UK10K and 1000G panels.

Both genotype and phenotype data were obtained in preprocessed form from Naqvi et al. (2021) and were used as such. The only exceptions to this are that the total set of 9,705,931filtered SNPs was pruned to remove SNPs with a variance inflation factor (VIF) larger than 2 before association testing (unless mentioned otherwise), and that one individual was omitted from the dataset because they opted out of the UK Biobank. The resulting dataset consisted of 19,643 individuals and 9,705,931 SNPs, 1,319,838 of which were used in association testing. Data preprocessing as done by Naqvi et al. (2021) is described in the subsequent paragraphs.

2.1.1 SNP-data quality control

PCA was used to select European individuals only. First, SNPs in LD (PLINK 1.9: 50 variant window-size, 5 variant step size, 0.2 r^2) from the 1000G (Phase 3) data were excluded from the dataset. Individuals were clustered using a k-nearest neighbors algorithm based on the first 25 reference ancestry principal components (PCs). This way a 1000G super population label was placed on each cluster, and only individuals with a EURO label were selected for further analysis. Next, indels and multi-allelic SNPs were removed. SNPs were filtered based on genotyping rate (< 50%), MAF (< 1%), and Hardy-Weinberg equilibrium ($P < 10^{-6}$). SNPs were subsequently filtered for LD (PLINK 50 variant window-size, 5 variant step size, 0.2 r^2), and related individuals were removed (identity by descent > 0.125). The resulting dataset contained 9,705,931 filtered SNPs. Before association testing, SNPs were pruned based on VIF (> 2) resulting in a pruned set of 1,319,838 SNPs.

2.1.2 Phenotype extraction and preprocessing

In a first step, the cortical surface was segmented and reconstructed into a 3D mesh from MRI data using the *recon-all* command in FreeSurfer v6.0.0 (Dale et al., 1999). The Connectivity Informatics Technology Initiative file format (CIFTIFY) was used to convert the FreeSurfer output to a Human Connectome Project (HCP) file format (Dickie et al., 2019). Cortical meshes were down sampled to a lower resolution (32,492 3D vertices spaced ~2 mm apart). The mid-cortical surface of both brain hemispheres was obtained from the resulting file by removing the subcortical vertices, using the Conte69 atlas (resulting in 29,759 3D vertices). This is the surface at the midpoint of the cortex, equidistant from the pial surface (i.e., the outer cortical surface) and the white matter interface. As a quality control measure, images were checked for mesh artefacts. Next, brain shapes were symmetrized.

For each of the individuals with processed brain data, the following covariates were used to control for during association testing: age, age squared, weight, height, systolic and diastolic blood pressure, the first 20 genetic PCs to control for population stratification, as well imaging-specific covariates. Individuals with extreme outlier covariates (> 6 standard deviations) were removed from the dataset. Partial least squares regression (*plsregress* in MATLAB 2020b) was used to correct brain shape for the selected covariates.

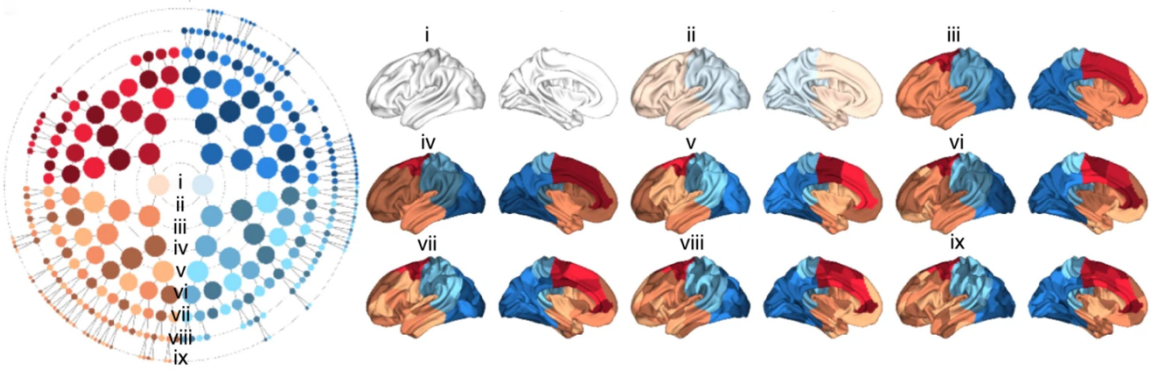


Figure 3 Hierarchical global-to-local segmentation of mid-cortical surface. Lower-case Roman numerals indicate the hierarchical level corresponding to circles in the polar dendrogram.

Next, cortical surfaces were segmented in a global-to-local hierarchical manner using a data-driven approach (Claes et al., 2018) (Figure 3). First a squared similarity matrix was calculated for the 29,759 3D vertices based on the RV coefficient (Robert & Escoufier, 1976) to quantify pairwise structural connections. A Laplacian transformation on this matrix was used to enhance the pairwise similarities. An eigenvalue decomposition was performed on the resulting matrix and *k*-means++ clustering was used in the transformed space to group highly correlated vertices together that divide the cortical surface into separate segments. This was done in a bifurcating hierarchical manner, i.e., vertices in a segment were grouped into two clusters, resulting in the subsequent bifurcation of the segment into two lower tier hierarchical segments. Hierarchical bifurcation was done over 9 levels, with 1, 2, 4, 8, 16, 32, 64, 128, and 256 nonoverlapping segments in levels 1, 2, 3, 4, 5, 6, 7, 8 and 9. Segments with less than 1% of the total vertex count were removed, resulting in a total of 285 hierarchical segments across 9 levels. Finally, for each of the 285 brain segments a generalized Procrustes analysis (GPA) was done to generate a shape space independent of other segments. To reduce the dimensionality, a PCA was done on the subsequent shape, and PCs were retained such that 80% of the variability of each segment was explained.

2.2 SNP grouping

Prior to association analysis, SNPs were grouped using one of three methods: based on 1) a sliding window, 2) haplotype blocks, and 3) genes. Grouping was done for each chromosome separately, so cross-chromosomal groups are not considered. The following approach was used to group SNPs:

Let a_k and b_k be the starting and ending coordinates for group S_k respectively, with $a_k < b_k$, and let x_i be the coordinate of SNP_i, then $\forall\, i,k: SNP_i \in S_k \Leftrightarrow\ a_k \leq x_i\ \leq b_k$. In other words, all SNPs within the boundaries (boundaries included) of a group are assigned membership to the group.

2.2.1 Sliding window-based grouping

For a fixed and positive window size w, expressed in kilobases (kb), the boundaries for a group S_k are $\left[x_1^k, x_1^k + w\right]$ where x_1^k is the coordinate of the first SNP not in groups $S_1 \ldots S_{k-1}$, assuming SNPs are in order of ascending coordinates on the chromosome. Graphically this approach can be seen as a sliding window moving over the chromosome, from start to end. The lower boundary of the window is at the coordinate of the first SNP currently not in any prior group. The upper boundary is one window size further. SNPs within the boundaries are assigned group membership. In this approach, to start a new group, the window 'jumps' to the next unassigned SNP and leaves small gaps where no SNPs from the dataset are located. This reduces the number of groups, and hence the number of tests compared to a no-jump approach.

2.2.2 Haplotype-based grouping

Haplotypes were calculated on complete set of 9,705,931 filtered SNPs for 19,643 individuals using the *blocks* function in PLINK 1.9. Default settings were used, expect for:

blocks-max-kb 500, which restricts blocks sizes to a maximum of 500 kb.

blocks-strong-lowci 0.5, which considers SNPs to be in strong LD if the lower bound of the 90% confidence interval for D' is larger than 0.5. This is the lowest value PLINK 1.9 takes and relaxes the LD threshold for SNPs in a block, resulting in slightly larger block sizes and thus fewer tests.

Block boundaries were considered group boundaries, and all SNPs were assigned a group based on their coordinates. Groups without any SNPs and SNPs not falling into any group were omitted from the analysis.

2.2.3 Gene-based grouping

A list of known RefSeq genes (hg19/GRCh37) was downloaded from the UCSC ftp servers. The list was filtered for *transcripts* only, and only genes on autosomal chromosomes were included. This means that microRNAs (miRNAs) and lincRNAs were also included in the analysis, partially representing intergenic regions. To reduce redundancy, any entry in the list that was completely enclosed within another entry was removed. In addition, per gene ID, entries were merged, which corresponds to taking the longest possible isoform, and the new

lower and upper boundaries were set as the minimal start position and maximal end position respectively. Different isoforms were not analyzed separately because they would cause redundancy and inflate the number of statistical tests. The boundaries were extended by 15 kb to include 90% of regulatory SNPs (Pickrell et al., 2010). All SNPs were assigned a group based on their coordinates, and groups without any SNPs, as well as SNPs not falling into any group were omitted from the analysis. In total 23,040 genes were analyzed.

2.3 SNP-set GWAS

Canonical correlation analysis (CCA) (Hotelling, 1936) was used to find associations between a set of p SNPs, represented by the matrix X_{nxp} and q phenotypes, represented by the matrix Y_{nxq}, for n individuals $(n = 19{,}643)$. SNPs were encoded as allelic dosage of the reference allele (0, 1, or 2) under the additive genetic model.

CCA calculates the canonical weight vectors a_i and b_i that maximize the correlation between the canonical variates $u_i = X\ a_i$ and $v_i = Y\ b_i$, and such that u_i and v_i are unit length. The Pearson correlation coefficient between u_i and v_i, denoted as $\rho_i = corr(u_i, v_i)$ is called the i^{th} canonical correlation. In addition, all vectors a_i are orthogonal, as well as vectors b_i. Concretely,

$$(a_i, b_i) = \underset{a_i, b_i}{\operatorname{argmax}} \rho_i = \underset{a_i, b_i}{\operatorname{argmax}} corr(X\ a_i, Y\ b_i) \qquad s.t. \quad a_i^T a_j = 0\ and\ b_i^T b_j = 0 \quad \forall\ i \neq j$$

The number of canonical correlations is equal to the minimal dimensionality between X and Y:

$$i_{max} = \min\{p, q\}.$$

Hypothesis testing was done using sequential testing. H_0: $k-1$ *canonical correlations are non-zero,* was tested against H_1: *at least* k *canonical correlations are non-zero*, for $k = 1 \ldots \min\{p, q\}$. The Bartlett-Lawley statistic (Glynn & Muirhead, 1978), L_k was applied,

$$L_k = -\left(n - k + 1 - \frac{1}{2}(p + q + 1) + \sum_{j=1}^{k-1} \rho_j^{-2}\right) ln\left(\prod_{j=k}^{\min\{p,q\}} (1 - \rho_j^2)\right)$$

The null distribution of L_k is asymptotically equal to a χ^2 distribution with $(p - k + 1)(q - k + 1)$ df and approximately equal for n sufficiently large.

It suffices that at least one canonical correlation is significantly different from zero for an association to be significant. Consequently, testing H_0: *no canonical correlations are significant* corresponds to taking $k = 1$. This simplifies the expression for L, and the resulting L_1 statistic with pq df, denoted as:

$$L = -\left(n - \frac{1}{2}(p + q + 1)\right) ln\left(\prod_{j=k}^{\min\{p,q\}} (1 - \rho_j^2)\right)$$

Worth noting is that the significance of this test should be interpreted as the significance of the entire decomposition of canonical correlations, and not just of the first canonical correlation. This property has important implications for determining generalizability of an association.

Other methods for hypothesis testing exist, most notably Rao's *F*-approximation. In practice, the *P*-value for an association was calculated in MATLAB 2020b by calling $canoncorr(X, Y)$, which implements both methods of hypothesis testing, but the approximate χ^2 test was chosen since genomic control factors were later calculated using χ^2 statistics.

In total 285 GWAS were conducted, one for each multivariate phenotype describing a brain segment, which magnified the multiple testing burden. Hence, associations are reported at two levels of significance. Genome-wide significance was declared for associations with a *P*-value lower than $0.05 \,/\, N$, with N the total number of SNP-based groups. A more stringent, study-wide threshold was determined as $0.05 \,/\, (N \times 285)$. These thresholds were obtained by applying the Bonferroni correction for the number of genome-wide and study-wide tests respectively. Note that brain segments overlap due to their hierarchal structure, and that therefore the tests across the 285 GWAS are not fully independent. As such, permutation testing (i.e., phenotypes are permuted and then tested for association, repeated e.g., 10,000 times) can be performed to estimate the number of independent tests. Although this would result is a slightly less stringent study-wide threshold, this was not done here.

Genome-wide significant SNP-set signals were merged into a single locus if their closest SNPs were less than 250 kb apart. Better ways of merging signals into independent loci exist, taking into account LD structure, but based on visual inspection of the Manhattan plots, this approach was sufficiently accurate. Gene-based signals were not merged.

2.4 Assessing generalizability through cross-validation

Cross-validation is a widely used approach in machine learning model evaluation and to assess the generalizability of a relation between sets of variables. K-fold cross-validation splits the data into k folds. The model's coefficients are learned on the data in all folds but fold k (i.e., the training dataset), and are subsequently applied to the data in fold k (i.e., the validation dataset). If the relation learned on the training data still holds on the validation data, the relation is considered generalizable. Here, 3-fold cross-validation is used, which keeps the validation dataset large enough (i.e., $n \gg p, q$, with p and q, the number of SNPs and dimensionality of the phenotype respectively, and n, the number of individuals). Canonical weights were learned on 2/3 of the data, and then applied to the other 1/3 of the data. The canonical correlations were calculated from the corresponding set of canonical covariates as the Pearson correlation coefficient. This was done three times, each time taking another 1/3 of the data as the validation set.

The method by which significance of a relation is normally determined in CCA cannot be applied to the correlations in the validation set (i.e., testing the entire set of canonical correlations at once using a Bartlett-Lawley test statistic (Glynn & Muirhead, 1978)). Generally, generalizability for CCA is arbitrarily determined based on the value of the canonical correlations calculated on the training and validation data, and without the use of significance testing (Dinga et al., 2019). Though such an approach is feasible for only a few tests, robust statistical methods for high-throughput screening of generalizability are lacking (Wang et al., 2020). Hence, a different approach was taken. To test whether an association between a set of SNPs and a multivariate phenotype was generalizable, the first set of canonical covariates was calculated for the validation data were tested for correlation directly. The *F*-statistic from a linear regression between the covariates was calculated. The *P*-value for the median *F*-

statistic over three folds served to indicate whether an association was generalizable. Since this method should not be considered a rigorous statistical test, extreme *P*-values only (i.e., > 0.1 or < 10^{-8}) were used to give an indication on whether or not a relation might be generalizable.

2.5 Genomic control factor calculation

The genomic control factor, λ_{GC} (Devlin & Roeder, 1999), based on the mean test statistic (Reich & Goldstein, 2001) was calculated for each of the 285 GWAS corresponding to a multivariate cortical phenotype. An implementation of λ_{GC} for bi-multivariate association testing is proposed and derived based on the idea that λ_{GC} is the ratio of the mean test statistic in the presence of population stratification over the mean test statistic in the absence of population stratification. A group test statistic, L_i from a bi-multivariate GWAS is asymptotically χ^2 distributed with $p_i q$ degrees of freedom, where p_i is the number of SNPs in a given group, and q is the dimensionality of the multivariate phenotype. As a result, the test statistics do not share the same null distribution, and hence an adapted method for λ_{GC} is necessary.

For λ_{GC}, defined as the observed mean test statistic over the expected mean test statistic it follows that $\lambda_{GC}$is equal to the observed sum of test statistics over the expected sum of test statistics by applying the linearity of expectation property for a constant factor $1/N$, where N is the total number of groups:

$$\lambda_{GC} = \frac{\frac{1}{N}\sum_{i=1}^{N} L_i}{\mathcal{E}\left[\frac{1}{N}\sum_{i=1}^{N} L_i\right]} = \frac{\frac{1}{N}\sum_{i=1}^{N} L_i}{\frac{1}{N}\mathcal{E}\left[\sum_{i=1}^{N} L_i\right]} = \frac{\sum_{i=1}^{N} L_i}{\mathcal{E}\left[\sum_{i=1}^{N} L_i\right]}$$

For large n, and $n \gg p, q$, with n the number of individuals, L_i is approximately distributed as χ^2 with $p_i q$ degrees of freedom. Thus, the following approximation is made:

$$\lambda_{GC} = \frac{\sum_{i=1}^{N} L_i}{\mathcal{E}\left[\sum_{i=1}^{N} \chi^2_{p_i q}\right]}$$

After applying the linearity of expectation property on the expected sum, it follows that:

$$\lambda_{GC} = \frac{\sum_{i=1}^{N} L_i}{\sum_{i=1}^{N} \mathcal{E}\left[\chi^2_{p_i q}\right]}$$

Since the expected value for a χ^2_k statistic with k degrees of freedom equals k, it follows that:

$$\lambda_{GC} = \frac{\sum_{i=1}^{N} L_i}{\sum_{i=1}^{N} p_i q} = \frac{\sum_{i=1}^{N} L_i}{q\sum_{i=1}^{N} p_i}$$

This implementation of λ_{GC}, just like its original implementation (Devlin & Roeder, 1999) empirically measures how much the observed mean test statistic deviates from the expected mean test statistics in the absence of population stratification, and thus is a useful diagnostic measure. Nonetheless, unlike the original implementation by Devlin & Roeder, it does not automatically follow that χ^2_k is inflated to $\lambda_{GC}\chi^2_k$ (for any k degrees of freedom) under the influence of population stratification; or in the case of the approximate χ^2_k statistic, L^k, that L^k is inflated to $\lambda_{GC}L^k$, only that the sum of test statistics is inflated by a factor λ_{GC}. This implicates

that λ_{GC} as proposed here does not automatically allow for correction of an inflated test statistic.

To overcome this issue, tests are pooled per number of degrees of freedom, k_j for each pool, j or until the size of the pool, N_j was at least 10 (which was only the case for the largest groups). As such, λ_{GC}^{j} is defined as the genomic control factor for pool j:

$$\lambda_{GC}^{j} = \frac{\frac{1}{N_j}\sum_{i=1}^{N_j} L_i^{k_j}}{\mathcal{E}\left[\chi_{k_j}^2\right]}$$

This ensures that the initial statement remains valid, i.e., that $\chi_{k_j}^2$ is inflated to $\lambda_{GC}^{j}\chi_{k_j}^2$, or that L^{k_j} is inflated to $\lambda_{GC}^{j}L^{k_j}$. As a result, correction for population stratification can be accomplished per pool as:

$$\chi_{k_j,corr}^2 = \frac{\chi_{k_j}^2}{\lambda_{GC}^{j}} \; or \; L_{corr}^{k_j} = \frac{L^{k_j}}{\lambda_{GC}^{j}}$$

2.6 Overlap with per-SNP GWAS

Summary statistics were obtained from (Naqvi et al., 2021), which report *P*-values for single SNP associations between the same phenotypes and genotypes of the same individuals analyzed in this work. Phenotypes were extracted from UK Biobank MRI images using exactly the same procedure.

Analogous to merging SNP-set signals into independent peaks, genome-wide significant per-SNP loci were merged into a single locus if two SNPs were closer than 250 kb apart. Although better peak calling method exist based also on LD measures, the accuracy was sufficient based on visual inspection of the Manhattan plot.

Per-SNP peaks were ranked based on peak strength before scanning for overlap with SNP-set peaks. Peak strength was calculated based on 1) the number of genome-wide significant SNPs in the peak, and 2) the distribution of genome-wide significant *P*-values in the peak. Concretely, peak strength, s for peak, π containing m genome-wide significant SNPs was calculated as:

$$s = -^{10}log\left(\sqrt[m]{\prod_{i \in \pi} P_i}\right)m$$

Overlap between per-SNP loci and SNP-set loci was considered if there was at least a 1 bp overlap between the loci.

2.7 Functional annotation and enrichment

To study the functional enrichment of genes in genomic proximity to the genome-wide significant loci detected by haplotype-based and window-based GWAS, the coordinates of the

most significant SNP-set for each locus were given as input to GREAT (Genomic Regions Enrichment of Annotations Tool) v4.0.4 (McLean et al., 2010), which was run with default settings. GREAT looks for enrichment in *cis*-regulatory regions of genes and for enriched biological pathways corresponding to those genes. Gene ontology (GO) terms with binomial and hypergoniometric FDR *q*-values below 0.05 were considered to be enriched in the GWAS results.

Study-wide significant genes detected in the gene-based GWAS were looked-up the GWAS Catalog to see whether they had known associations with brain-related phenotypes detected in other studies.

2.8 Influence of window size

To investigate the effects of group size on the number of significant associations, multiple window-based GWAS were carried out with window sizes ranging from 5 kb to 200 kb (5 – 10 – 20 – 50 – 100 – 200). For each GWAS the number of genome-wide and study-wide significant loci was determined, as well as the number of genome-wide significant loci that did not overlap with per-SNP GWAS loci from Naqvi et al. (2021), determined in the same way as described in section *2.6*.

2.9 Neanderthal interogressed haplotype block associations

Genomic coordinates for a total of 6004 Neanderthal interogressed haplotype blocks were retrieved from Vernot et al. (2016). These blocks were tested for associations with each of 285 multivariate cortical phenotypes in two different ways. 1) A SNP-set association analysis was done, where the Neanderthal haplotype blocks define the groups, and SNPs located inside a Neanderthal block were assigned group membership. This way, only SNPs within Neanderthal blocks were included in the analysis. 2) The results of the regular haplotype-based GWAS were scanned for overlap with Neanderthal haplotype blocks, and associations with the phenotype were based on *P*-values from the regular blocks.

3 RESULTS

3.1 Effect of pruning SNPs on false discovery rate and comparison with per-SNP association testing

Because CCA always finds the maximum correlation between linear combinations of both a multivariate phenotype and a set of SNPs, even when in reality no underlying correlation exists, it has a natural tendency to overfit the data, especially in the case of highly correlated variables. This in turn leads to false discoveries. In addition, the degree of overfitting is magnified when the dimensionality of the data is large in comparison to the number of individuals. To overcome this issue, SNPs were pruned, and removed from the dataset if the VIF was larger than two. The benefit is twofold as it also decreases the group sizes. To illustrate the effect of pruning, multivariate gene-based association analysis was performed for full brain shape (segment 1) with chromosome 17 (Figure 4). This chromosome was chosen because a very large peak was detected in the individual SNP analysis by Naqvi et al. (2021), and thus is expected to also be detected in a gene-based association analysis (i.e., a positive control).

Before pruning, a total of 5 genes passed the genome-wide significance threshold: 2 genes overlapping with the large per-SNP peak (Figure 4, indicated with a green circle); 3 genes detected in a desert region (i.e., a region where no per-SNP signal was produced) (Figure 4, indicated with a red circle). After pruning, the two genes overlapping with the large per-SNP peak remained significant, while the genes in desert regions had *P*-values close to 1. A 3-fold cross-validation approach was used to verify that the green genes were indeed true discoveries, and the red genes were false discoveries. In the unpruned dataset, the median test-set *P*-values of the two green genes were 1.83×10^{-12} and 2.00×10^{-13} respectively, while those of the three red genes were 0.493, 0.261, and 0.680 respectively. After pruning based on VIF, the two green genes remained significant, while the red genes were no longer significant. These results indicate that pruning is necessary to reduce the false discovery rate in bi-multivariate GWAS based on CCA, which is in accordance with findings from Tang & Ferreira (2012).

Additionally, Figure 4 illustrates the reduced testing burden achieved by grouping SNPs. The basal gene-based *P*-values fluctuate at far higher levels (i.e., lower levels on the $-^{10}\log$-scale) than those from the individual SNP analysis. The individual SNP analysis detected more genome-wide significant loci ($P < 5 \times 10^{-8}$), but its strongest signal was reproduced by the gene-based analysis.

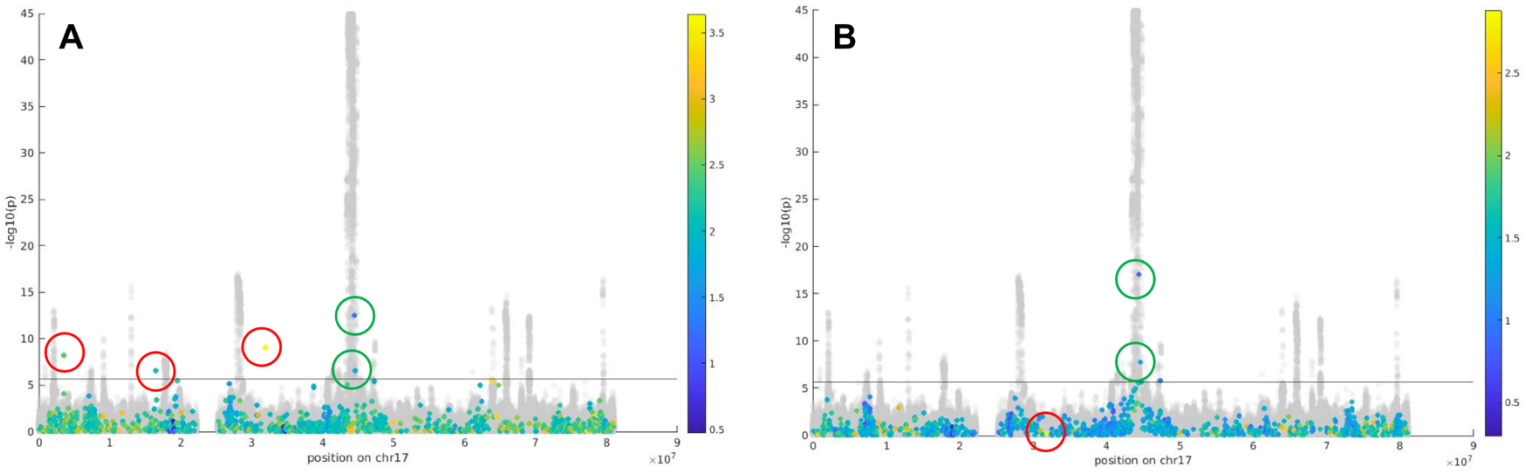


Figure 4 Multivariate gene-based association analysis results for the main brain segment with chromosome 17. The horizontal line indicates the Bonferroni corrected genome-wide threshold for 23,040 gene-based tests ($P < 2.17 \times 10^{-6}$). Colored dots correspond to genes, where the color indicates logarithmically the number of SNPs (i.e., log10(#SNPs)). Grey dots correspond to individual SNP *P*-values. **A** shows the results for the unpruned set of SNPs. **B** shows the results for the pruned set (VIF < 2). Green circles indicate SNP-sets that overlap with the central per-SNP peak. These SNP-sets remain significant after pruning. Red circles indicate SNP-sets located in per-SNP deserts. These SNP-sets are no longer significant after pruning.

3.2 Comparative bi-multivariate genome-wide SNP-set association testing

To test and compare the performance of SNP-set association testing using a CCA model formulation, genome-wide associations were tested across all 285 hierarchical brain segments, and for three SNP-grouping methods: based on 1) genes, derived from the UCSC known RefSeq gene list; 2) haplotype blocks, estimated on the unpruned SNP data itself using PLINK 1.9; and 3) based on a sliding window of 20 kb (Figure 5). Associations are reported at two levels of significance, genome-wide and study-wide significance. Table 2 summarizes the thresholds and results corresponding to each grouping method. Quantile-quantile plots (QQ-plots) are given by Supplementary figure 1.

Table 2 Summary of thresholds and detected associations for SNP-set GWAS based on genes, haplotype blocks, and a sliding window (20kb). Genome-wide (GW) and study-wide (SW) signals refer to individual SNP-sets. Loci refer to regions resulting from the merging of nearby individual SNP-set signals into independent, larger groups.

	Nr. of groups	GW threshold	SW threshold	GW signals	SW signals	GW loci	SW loci
genes	23,040	2.17e-06	7,61e-09	120	21	/	/
haplotype blocks	239,244	2.09e-07	7,33e-10	338	147	124	49
sliding window (20kb)	117,697	4.25e-07	1,49e-09	363	150	124	31

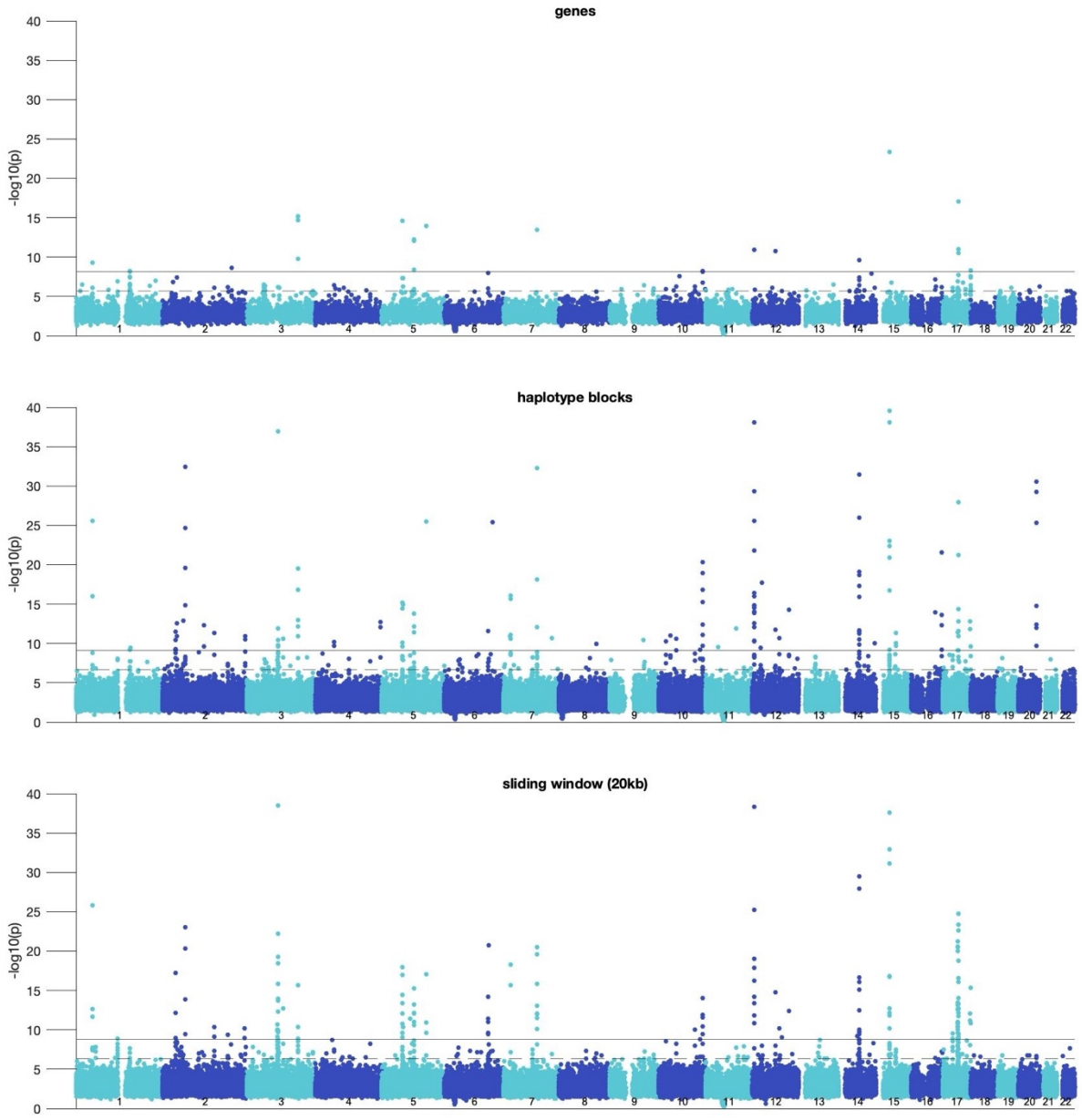


Figure 5 Manhattan plots for bi-multivariate GWAS based on genes, haplotype blocks, and a sliding window (20 kb) respectively. SNP-sets are represented by the minimal *P*-value obtained across all 285 phenotypes. The Bonferroni corrected genome-wide significance threshold is indicated by a dashed horizontal line, and the Bonferroni corrected study-wide significance threshold is indicated by a full horizontal line.

Grouping SNPs effectively reduced the genome-wide significance threshold, compared to the conventional threshold of $P < 5 \times 10^{-8}$ often used in individual SNP GWAS. All methods reduced the threshold by at least a factor 10 (i.e., 1 log point), and the gene-based GWAS with a factor > 200. Despite having the least stringent significance thresholds, both genome-wide and study-wide, gene-based GWAS detected the fewest number of associations, even after signals generated by the use of other grouping methods were merged based on genomic position (i.e., loci). Interestingly, although haplotype-based and window-based GWAS both detect 124 genome-wide significant loci, haplotype-based GWAS detected 49 study-wide significant loci, whereas window-based GWAS detected only 31. Interestingly, haplotype-based GWAS detected a large study-wide significant peak on chromosome 20 ($P < 10^{-30}$), which was not detected by any other grouping method. Gene-based GWAS, in comparison with the other

methods detected far fewer loci. In addition, many of the associated loci detected by haplotype-based and window-based GWAS that were not detected by gene-based GWAS do overlap with genes.

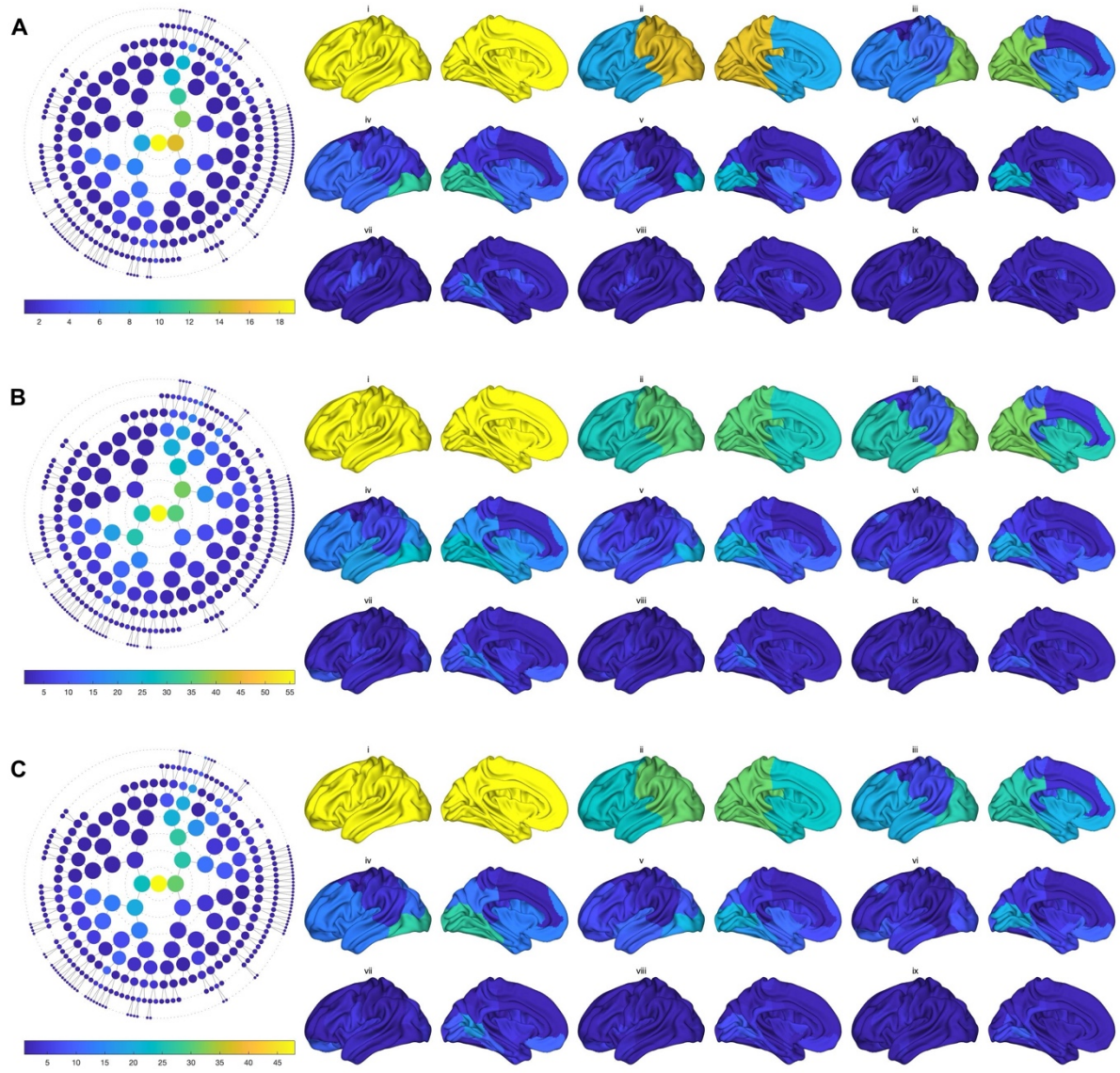


Figure 6 Number of genome-wide significant associations per hierarchical brain segment. The number of genome-wide significant loci from **A** gene-based GWAS, **B** haplotype-based GWAS, and **C** window-based GWAS is represented by colors. Scales are set relative to the maximum number of associations for each SNP-set GWAS method in order to compare trends between different methods. Lower-case Roman numerals indicate the hierarchical level corresponding to circles in the polar dendrogram (with *i* the center, and *ix* the outer layer).

The number of genome-wide significant associations for each of 285 hierarchical brain segment is shown in Figure 6. The relative number of associations per segment was similar for each SNP-set GWAS method. Most associations were detected for the larger segments in hierarchical levels *i*, *ii* and *iii*. In the lower hierarchical tiers, *iv*, *v*, and *vi*, few loci were associated with the corresponding brain segments, however, a relatively high number of loci were found to be associated with the occipital lobe. At even lower hierarchical tiers only very few associated loci were detected. This is in accordance with findings from Naqvi et al. (2021).

3.3 Inflation of test statistics

An adapted version of genomic control factor suited for bi-multivariate GWAS was applied to the results of gene-based, haplotype-based, and window-based GWAS for each multivariate phenotype. The maximum genomic control factors across all multivariate phenotypes were 1.014, 1.019, and 1.017 for gene-based, haplotype-based, and window-based respectively (Supplementary figure 2). Hence, the mean test statistic is only slightly inflated, and population stratification was not a problem. In spite of that, these results so far only investigated the inflation of the mean test statistic, however, there is no guarantee of uniformity of the genomic inflation factor over test statistics with different null distributions as generated by CCA. To investigate the inflation of test statistics across different group sizes (and thus different null distributions), test statistics, L_i generated by CCA were sorted and subsequently pooled.

Figure 7 illustrates how the test statistic is inflated for larger degrees of freedom in the haplotype-based and the window-based GWAS, resulting from larger group sizes in segments 1 and 7, which have 437 and 170 dimensions respectively. In the lower-dimensional segments, 44 and 172, with 90 and 10 dimensions respectively, the inflation of the test statistic was very small and only observed at the highest degrees of freedom. For the gene-based GWAS, the test statistic was only slightly inflated in segment 1 (full brain), but not in the other segments under investigation. In addition, across most segments and in all SNP-set GWAS variants, the smallest groups (i.e., pools with the lowest df) had deflated test statistics (i.e., a genomic control factor < 1). The segments chosen for this analysis are indicated on Supplementary figure 3.

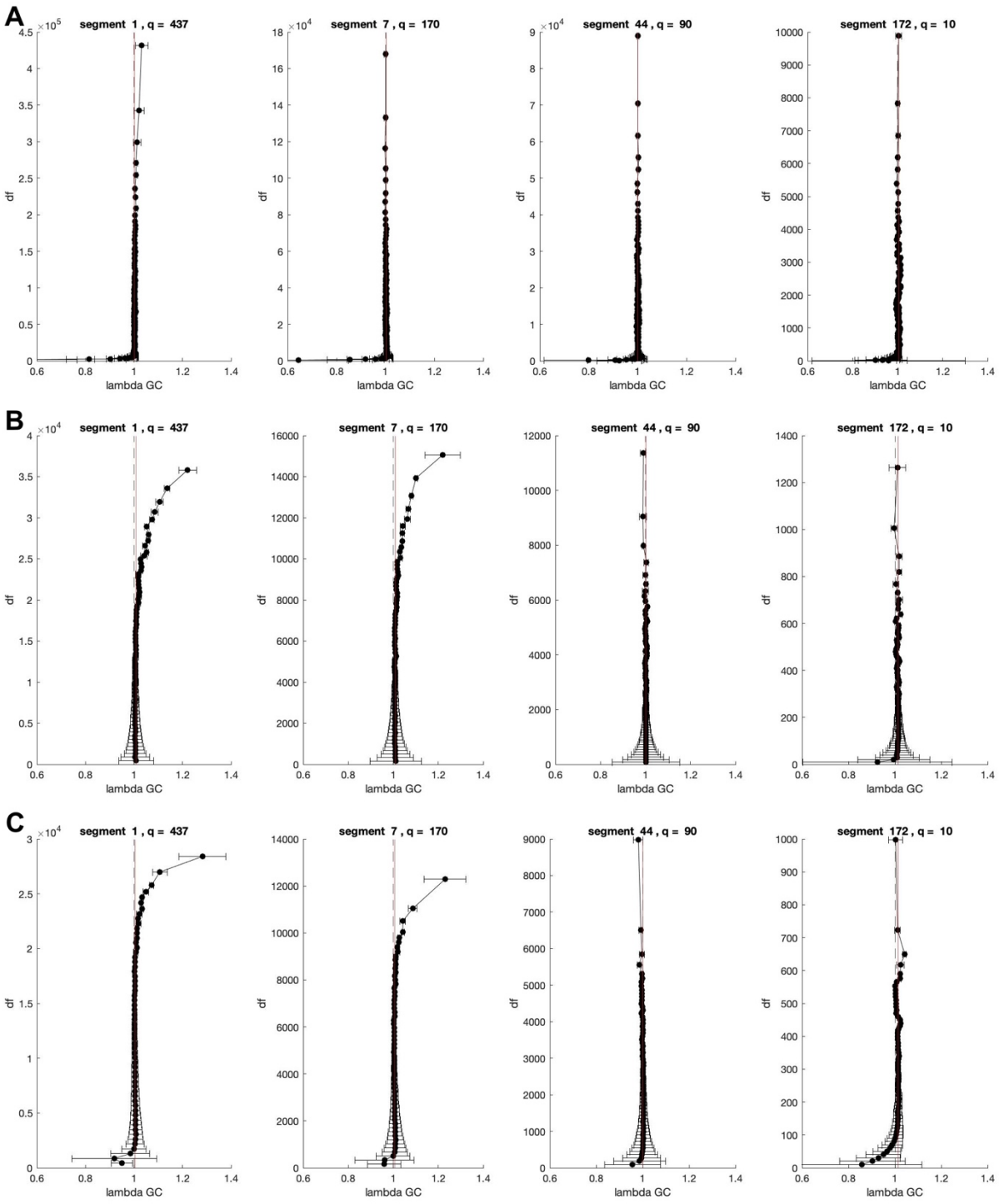


Figure 7 Genomic inflation factor for pooled test statistics. The genomic inflation factor was calculated on test statistics, pooled per df or until the pool size was 10 for segments 1, 7, 40, and 172 for **A** gene-based, **B** haplotype-based, and **C** window-based GWAS. Segment 1 represents the highest dimensional phenotype, segment 172 the lowest. Other segments were selected in between. Error bars indicate the standard deviation on lambda. Dashed vertical lines indicate 1. Vertical red lines indicate the median lambda over all pools.

Furthermore, the idea of a genomic control factor, λ_{GC} is that a test statistic, χ_k^2 with k degrees of freedom is inflated to $\lambda_{GC}\chi_k^2$ under the influence of population stratification or systematic bias. As such, genomic control factors calculated on pooled test statistics can be used to correct the corresponding test statistics. This correction is illustrated by Figure 8 on the *P*-values of the window-based GWAS. Worth noting is that the inflated test statistics correspond

to less than 100 tests for each multivariate phenotype, therefore the vast majority of test statistics were not inflated. Correction of the test statistic leads only to a minor loss in power. Note that no correction was made to deflated test statistics.

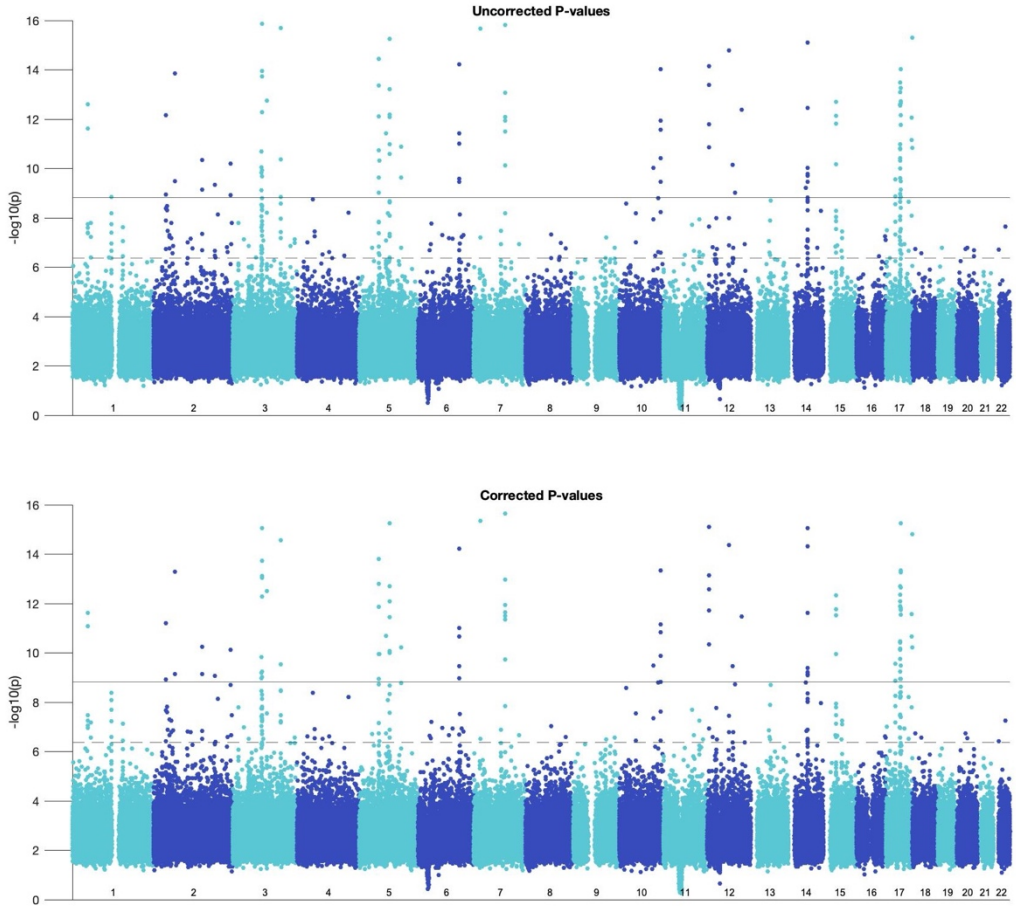


Figure 8 Manhattan plot for window-based (20kb) GWAS before and after the pool-based correction of test statistics. SNPs are represented by the minimal *P*-value obtained across all 285 phenotypes. Genome-wide ($P < 4.25 \times 10^{-7}$) and study-wide significance ($P < 1.49 \times 10^{-9}$) are indicated by a dashed and full horizontal line respectively.

To further investigate whether the inflation of test statistics for larger SNP-sets is due to the number of SNPs simultaneously tested for association, test statistics from different sized window-based GWAS (5 – 200 kb) were pooled and analyzed for inflation (Figure 9). This was done for the main brain segment (i.e., segment 1) since the degree of inflation of test statistics was highest for it. The purpose of this analysis is to investigate whether there is a bias for large group sizes due to inherent properties of CCA, which could in turn lead to an increased FDR. For the largest window sizes, i.e., 100 and 200 kb, no inflation of the test statistic was observed despite having the largest group sizes in terms of SNP-count. This shows that simultaneous testing of up to at least 350 SNPs in itself does not lead to an inflated test statistic, and that there is therefore no bias for larger group sizes based on the number of SNPs alone. On the other hand, the highest degree of test statistic inflation was observed for the smallest window sizes, i.e., 5, 10, and 20 kb despite having the smallest group sizes in terms of SNP-count. In addition, the number of SNPs required per window to inflate the test statistic is different for every window size. Just like earlier, it was observed that inflation of the test statistic affected fewer than 100 SNP-sets who were clear outliers in terms of size (Supplementary figure 4).

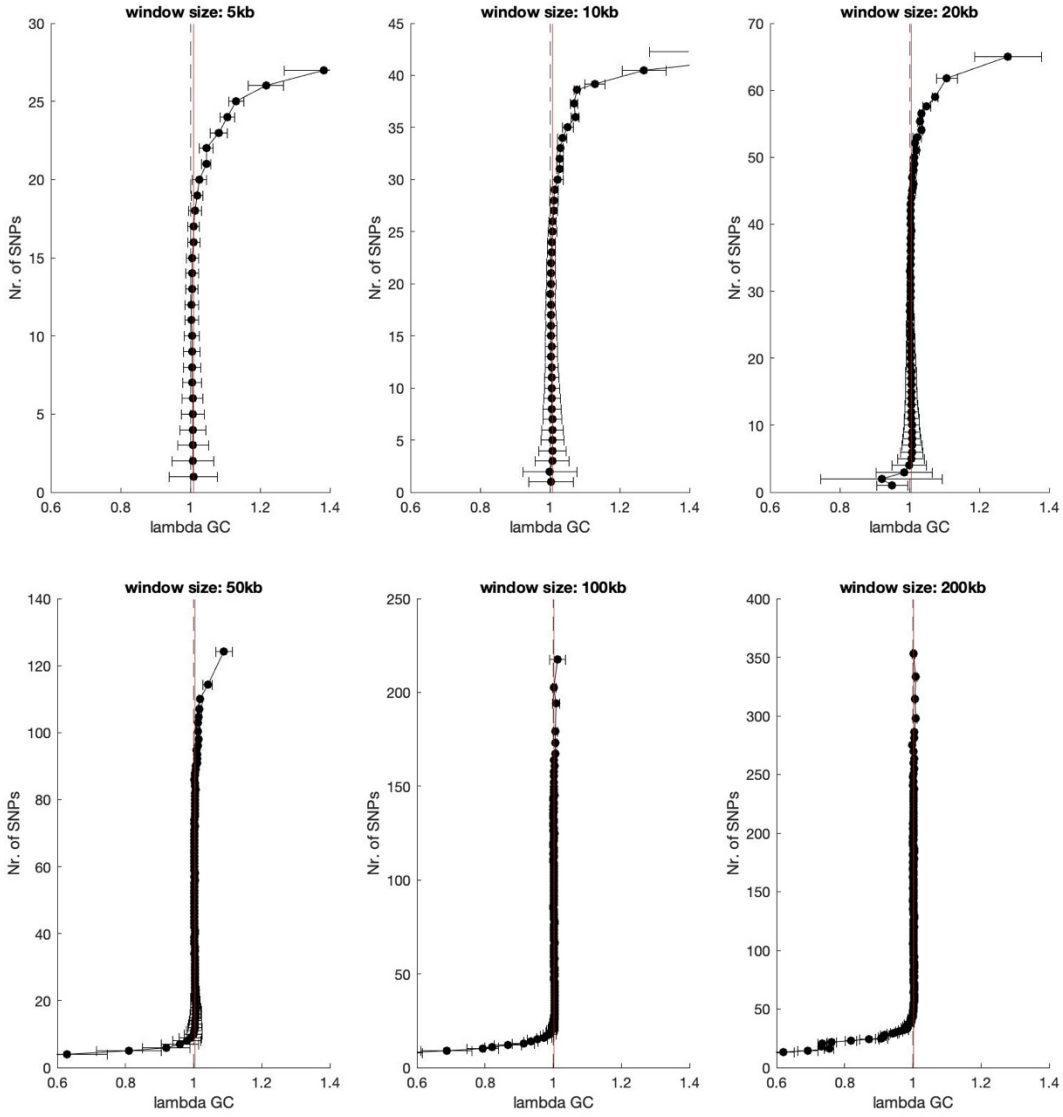


Figure 9 Genomic inflation factor for pooled test statistics. The genomic inflation factor was calculated on test statistics, pooled per df or until the pool size was 10 for segment 1 using test statistics from window-based GWAS of sizes: 5 – 10 – 20 – 50 – 100 – 200 kb. Error bars indicate the standard deviation on lambda. Dashed vertical lines indicate 1. Vertical red lines indicate the median lambda over all pools.

3.4 Functional annotation, enrichment, and known associations

GREAT (McLean et al., 2010) was used to look for biological pathways enriched for genes in the genomic proximity of genome-wide significant loci detected by haplotype-based and window-based GWAS. Genes associated with loci detected by haplotype-based GWAS were enriched in brain-related and morphogenesis-related processes (binomial and hypergoniometric FDR *q*-value < 0.05) (Table 3). No significant enrichment was found in human phenotypes, but enrichment was found for mouse phenotypes (abnormal neurocranium morphology, abnormal craniofacial development, and abnormal embryo morphology). No enrichment was found for results from window-based GWAS.

Table 3 Biological processes enriched for genes in genomic proximity of genome-wide significant loci detected by haplotype-based GWAS. Biological processes were enriched in brain-related and morphogenesis-related GO terms.

Brain-related GO terms	Morphogenesis-related GO terms
– craniofacial suture morphogenesis	– embryonic morphogenesis
– negative regulation of neuron projection development	– embryonic organ development
– cranial suture morphogenesis	– chordate embryonic development
– regulation of neuron differentiation	– negative regulation of developmental growth
– regulation of neurogenesis	– regulation of developmental growth
– forebrain cell migration	– animal organ morphogenesis
– telencephalon development	
– negative regulation of axon extension	
– telencephalon cell migration	
– cerebral cortex cell migration	

Each of the 21 study-wide significant ($P < 7.61 \times 10^{-9}$) genes, detected by gene-based GWAS, were screened for known associations with brain-related traits, or associations already detected by GWAS (Table 4). *LOC440982* had no known associations with brain-related traits, but is only 5 kb apart from *ZIC1*, which is known to be associated with brain morphology (Grasby et al., 2020; Zhao et al., 2019). All other genes had known brain-related phenotype associations, most notably cortical surface area and brain volume measurements. Thus, gene-based GWAS did not find any study-wide significant loci not previously found to be associated with the brain.

Table 4 Associated traits in for study-wide significant ($P < 7.61 \times 10^{-9}$) genes detected by gene-based GWAS.

Chromosome	Gene ID	Nr. of SNPs	*P*-value	Known associated traits
1	*LINC01389*	26	4.99e-10	Brain volume measurement (GWAS) (Zhao et al., 2019)
1	*EFNA4*	16	6.09e-09	Craniosynostosis (Merrill et al., 2006)
2	*HSPD1*	9	2.13e-09	Schizophrenia (GWAS) (Lam et al., 2019), Major depression and alcohol dependence (GWAS) (Zhou et al., 2017)
3	*ZIC4*	17	6.36e-16	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020)
3	*ZIC1*	15	1.97e-15	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020)
3	*LOC440982*	44	1.58e-10	/
5	*SMIM15*	10	2.39e-15	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020)
5	*NR2F1-AS1*	42	6.18e-13	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020), educational attainment (GWAS) (Donati et al., 2021)
5	*NR2F1*	17	8.91e-13	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020)
5	*FAM172A*	95	4.45e-09	Neuroticism (GWAS) (Luciano et al., 2018), cortical surface area (GWAS) (Grasby et al., 2020; van der Meer et al., 2020)
5	*MIR4460*	23	1.18e-14	Cortical volume (GWAS) (Hofer et al., 2020)
7	*SEM1*	76	3.75e-14	Cortical volume (GWAS) (Hofer et al., 2020), cortical surface area (GWAS) (Grasby et al., 2020; van der Meer et al., 2020)
10	*EEF1AKMT2*	14	5.53e-09	Cortical thickness (GWAS) (Grasby et al., 2020), cortical surface area (GWAS) (Grasby et al., 2020; van der Meer et al., 2020)
12	*LINC02443*	11	1.19e-11	Cortical surface area (GWAS) (Grasby et al., 2020; van der Meer et al., 2020)
12	*LOC100507065*	49	1.89e-11	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020)
14	*DAAM1*	107	2.28e-10	Cortical volume (GWAS) (Hofer et al., 2020), cortical surface area (GWAS) (Grasby et al., 2020)
15	*C15orf54*	22	4.09e-24	Brain volume measurement (GWAS) (Zhao et al., 2019), cortical surface area (GWAS) (Grasby et al., 2020)
17	*NSFP1*	7	9.51e-18	Dyslexia (Veerappa et al., 2014)
17	*ARL17A*	18	3.33e-11	Intelligence (GWAS) (Davies et al., 2018)
17	*NSF*	34	1.02e-11	Cortical volume (GWAS) (Hofer et al., 2020), cortical surface area (GWAS) (Grasby et al., 2020)
17	*NPLOC4*	64	5.07e-09	Mathematical ability (GWAS) (Lee et al., 2018)

3.5 Overlap with individual SNP GWAS

GWAS summary data were obtained from Naqvi et al. (2021), and individual SNPs were merged into a single loci if they were closer than 250 kb, similar to how sets of SNPs were merged. This resulted in 443 genome-wide significant peaks. Peaks were subsequently ranked. Figure 10 shows the Manhattan plot for the individual SNP *P*-values from Naqvi et al. (2021) and the rank of each peak is indicated.

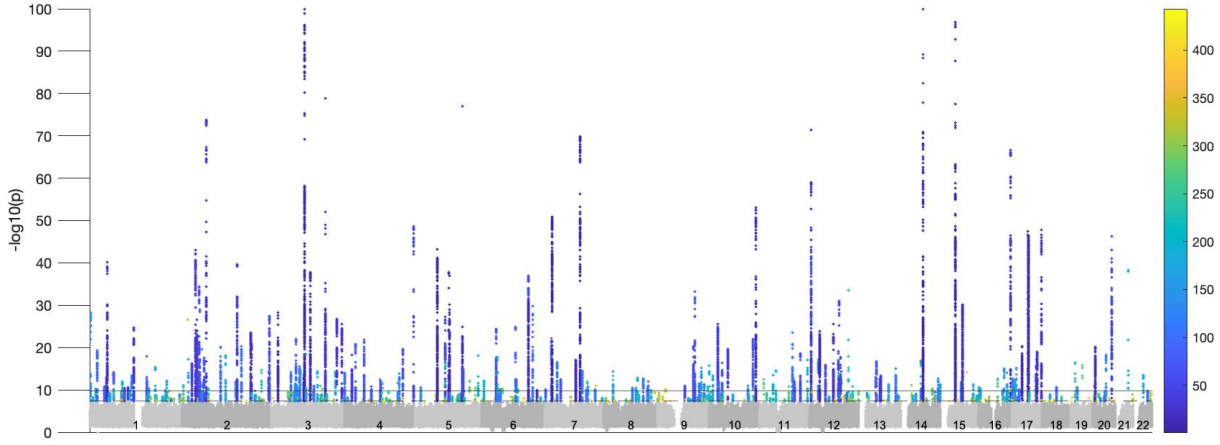


Figure 10 Manhattan plot for individual SNP *P*-values from Naqvi et al. (2021). SNPs are represented by the minimal *P*-value obtained across all 285 phenotypes. Genome-wide ($P < 5 \times 10^{-8}$) and study-wide significance ($P < 1.75 \times 10^{-10}$) are indicated by a dashed and full horizontal line respectively. SNPs are colored per peak and the color corresponds to the rank assigned for peak strength (strongest peak gets rank 1). SNPs above the genome-wide significance threshold are colored gray.

Per-SNP and SNP-set GWAS peaks were considered to be overlapping if there was at least a 1 bp overlap between the peak's respective genomic coordinates. Figure 11 shows that stronger per-SNP GWAS peaks are also more likely to also be detected in a SNP-set GWAS. This finding is consistent for all three SNP-set methods, i.e., gene-based, haplotype-based, and window-based. Gene-based GWAS detected the fewest number of associations, and hence overlapped with the fewest number of per-SNP GWAS peaks. Of the 50 strongest peaks, 50, 80, and 84% were detected in gene-based, haplotype-based, and window-based GWAS respectively, while peaks ranked 250 or higher were only rarely detected in any of the SNP-set GWAS.

In addition, SNP-set GWAS resulted in genome-wide significant loci that were not detected by per-SNP GWAS (Naqvi et al., 2021): 21, 31, and 41 for gene-base, haplotype-based, and window-based GWAS respectively. Table 5 lists genome-wide significant sets of SNPs detected through window-based GWAS (20 kb) that did not contain any SNPs with genome-wide significance in the per-SNP GWAS and which are located at least 1 Mb away from any significant individual SNPs. Some SNP-sets were over 10 Mb away from any significant individual SNPs, illustrating that SNP-set GWAS is an effective approach to find associations in regions where per-SNP GWAS fails to find any. Worth noting is that these 'new' associations (i.e., not detected by per-SNP GWAS) have small to medium group sizes (< 20 SNPs) and are therefore not affected by inflation of their test statistic as illustrated earlier in *3.3* (Figure 7C).

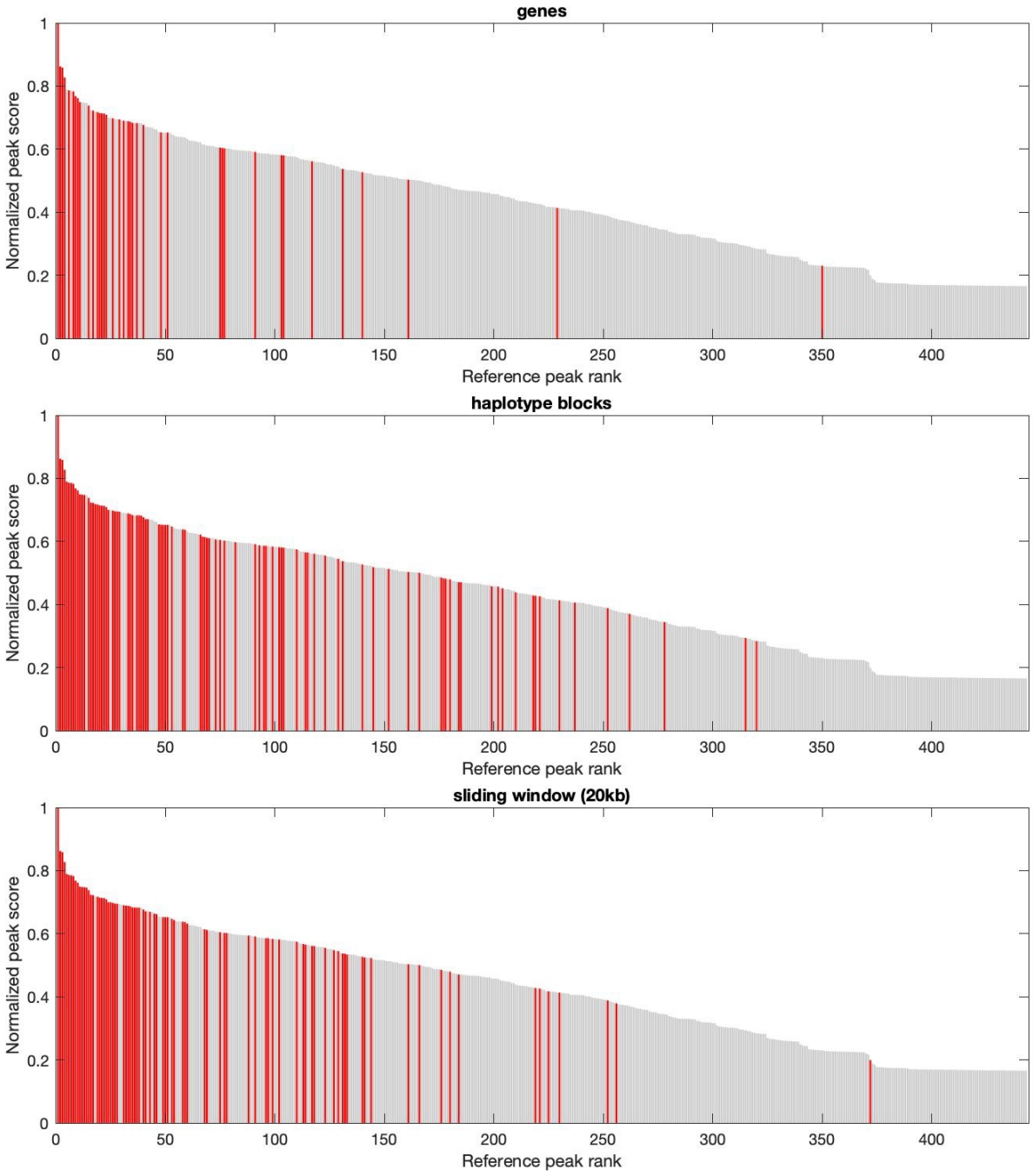


Figure 11 Detection of per-SNP GWAS loci by SNP-set GWAS. Vertical lines correspond to peaks from per-SNP GWAS (Naqvi et al., 2021) ranked by peak strength, represented the height of each line. Peaks are colored red if they overlap with any peak from gene-based, haplotype-based, or window-based GWAS respectively.

Table 5 Genome-wide significant sets of SNPs detected through window-based GWAS (20kb) that did not contain any SNPs with genome-wide significance in the per-SNP GWAS (Naqvi et al., 2021) and which are located at least 1 Mb away from any significant individual SNPs. Coordinates of the first and last SNP in the set are given, as well as the number of SNPs in the set and its corresponding *P*-value. Genome-wide and study-wide significance are declared at $P < 4.25 \times 10^{-7}$ and 1.49×10^{-9} respectively.

Chromosome	Position first SNP (bp)	Position last SNP (bp)	Nr. of SNPs	*P*-value	Distance to nearest significant SNP (bp)
2	43,598,125	43,617,931	7	3.29e-08	1,291,869
2	113,376,046	113,390,332	15	6.99e-08	5,620,268
2	225,936,885	225,954,460	14	1.52e-07	3,554,517
2	228,193,036	228,211,836	9	2.40e-07	5,810,668
2	240,897,186	240,915,079	13	1.63e-08	3,176,050
3	34,853,919	34,867,820	8	7.25e-08	10,897,312
4	47,387,943	47,404,621	11	1.83e-09	1,215,315
4	77,040,695	77,058,983	7	3.09e-07	1,408,521
4	107,367,936	107,384,291	6	4.19e-07	2,049,352
5	38,673,224	38,693,118	11	1.65e-07	12,999,380
6	74,177,337	74,197,300	11	4.92e-08	4,517,308
6	106,171,462	106,189,866	11	6.96e-08	7,181,198
8	103,645,715	103,661,730	8	3.77e-07	4,731,950
10	111,218,808	111,234,402	10	3.38e-07	5,545,968
10	120,502,340	120,521,312	8	1.58e-09	1,221,860
11	65,473,798	65,492,211	8	3.18e-07	3,414,357
11	94,402,958	94,421,449	7	2.13e-07	1,045,159
11	121,501,406	121,514,843	5	3.15e-07	8,751,277
12	17,010,248	17,029,231	8	1.53e-07	12,885,076
12	114,030,068	114,049,977	18	1.14e-07	1,133,228
13	55,399,131	55,419,062	9	8.91e-08	1,338,251
14	88,809,657	88,824,831	13	1.62e-07	3,046,841
15	42,521,388	42,537,718	11	1.33e-07	2,500,876
17	57,281,195	57,298,806	3	1.19e-07	6,604,314
18	8,091,578	8,106,321	4	1.80e-07	1,846,240
18	26,728,624	26,748,359	7	2.76e-07	3,217,602
20	25,939,119	25,949,415	5	1.86e-07	3,966,539
22	17,745,930	17,760,895	18	1.98e-07	10,121,145

3.6 Influence of group size

To investigate the effects of group size on the number of significant associations, multiple window-based GWAS were carried out with window sizes ranging from 5 kb to 200 kb

(Manhattan plots in Supplementary figure 5). For each GWAS the number of genome-wide and study-wide significant loci was determined, as well as the number of genome-wide significant loci that did not overlap with per-SNP GWAS loci from Naqvi et al. (2021), determined in the same way as in section *3.5* (Figure 12). With increasing window size, both the number of genome-wide and study-wide significant loci decreased. The number of genome-wide significant loci started decreasing rapidly for window sizes larger than 20 kb, and the number of study-wide significant loci already for widow sizes larger than 10 kb. Between 5 and 200 kb windows, the number of genome-wide and study-wide significant loci dropped by > 3-fold (145 to 44) and > 5-fold (58 to 11) respectively. The number of genome-wide significant loci that did not overlap with any per-SNP GWAS loci from Naqvi et al. (2021) increased for window sizes of 20 and 50 kb, and then decreased for even larger window sizes. A window of 20 kb resulted in a high number of SNP-set exclusive loci, as well as total number of loci and is therefore a good choice in the current GWAS setting (i.e., for the current phenotypes and model).

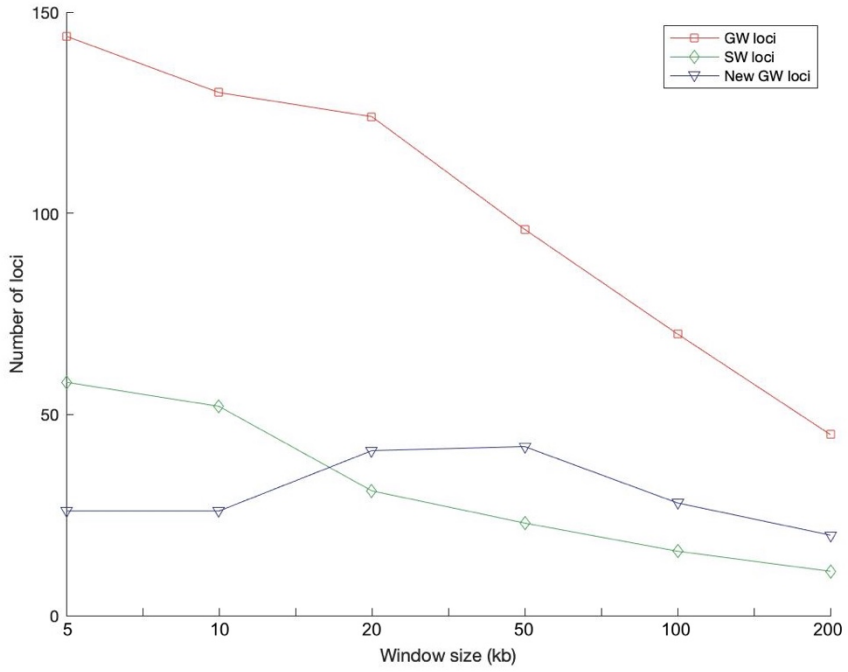


Figure 12 The number of genome-wide significant (GW loci – red), study-wide significant (SW loci – green), and genome-wide significant loci not overlapping with individual SNP loci (New GW loci – blue) from window-based GWAS. Window sizes are: 5 – 10 – 20 – 50 – 100 – 200 kb.

3.7 Neanderthal interogressed haplotype blocks

As a first means to explore the idea that Neanderthal interogressed haplotype blocks have the potential to affect brain morphology in present-day humans, the coordinates of 6004 interogressed Neanderthal blocks were obtained from Vernot et al. (2016) and subsequently used as group boundaries in a SNP-set GWAS (Figure 13A). As such, alleles located within the boundaries of these Neanderthal blocks are simultaneously tested for association with cortical brain morphology. In total, 44 and 6 Neanderthal blocks (Table 6) corresponding to 38 and 6 independent loci reached genome-wide and study-wide significance respectively.

Furthermore, the interogressed Neanderthal blocks were projected onto the results from the haplotype-based GWAS done in section *3.2*, and associated haplotype blocks overlapping with Neanderthal blocks were marked (Figure 13B). Neanderthal blocks overlapped with 49 and 23 genome-wide and study-wide significant loci respectively, which is over 3 times more study-wide significant loci. In addition, the Manhattan plot shows how some peaks entirely overlap with Neanderthal blocks (e.g., on chromosome 14), and how some do not (e.g., on chromosome 7). In addition, the study-wide significant peaks from the Neanderthal haplotype-based GWAS (Figure 13A) line up with study-wide significant peaks from the regular haplotype-based GWAS (Figure 13B). The signal on chromosome 17 is only ~100 kb apart from an associated locus and is therefore lacking on Figure 13B.

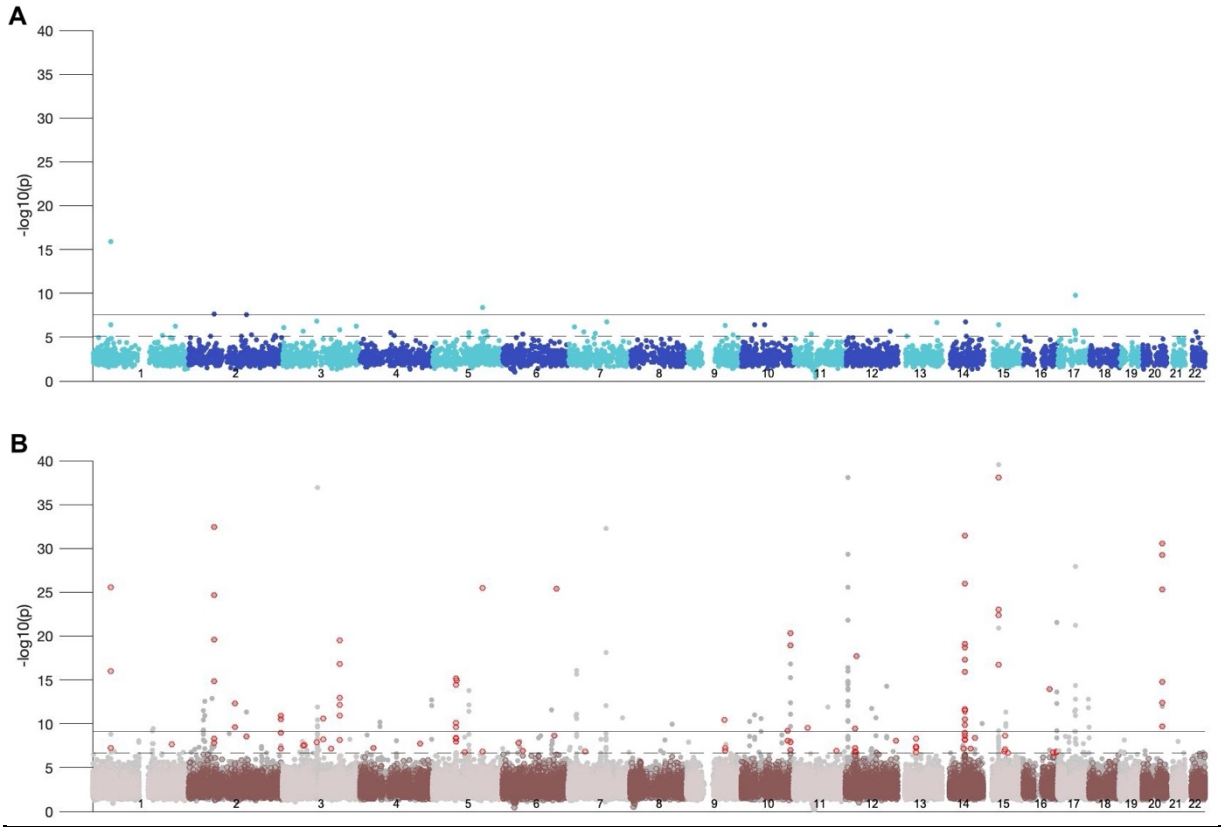


Figure 13 (A) Manhattan plots for SNP-set GWAS based on interogressed Neanderthal haplotype blocks. SNPs are represented by the minimal *P*-value obtained across all 285 phenotypes. Genome-wide ($P < 8.33 \times 10^{-6}$) and study-wide significance ($P < 2.92 \times 10^{-8}$) are indicated by a dashed and full horizontal line respectively. One block on chromosome 15 (39,569,380 – 39,643,906) had a *P*-value of 4.27×10^{-43} and is therefore not shown on the plot. **(B) SNP-set GWAS based on haplotype blocks calculated on 19,643 individuals from the UK Biobank marked by overlap with Neanderthal blocks.** SNPs are represented by the minimal *P*-value obtained across all 285 phenotypes. Genome-wide ($P < 2.09 \times 10^{-7}$) and study-wide significance ($P < 7.33 \times 10^{-10}$) are indicated by a dashed and full horizontal line respectively. SNP-sets are indicated by gray dots, and those that overlap with a Neanderthal block are indicated by a circle, or a red circle if they reached genome-wide significance.

Table 6 Study-wide significant ($P < 2.92 \times 10^{-8}$) Neanderthal interogressed haplotype blocks.

Chromosome	Start posistion (bp)	End position (bp)	Nr. of SNPs	*P*-value
1	47,931,769	47,974,535	21	1.26e-16
2	65,613,438	66,025,838	241	2.26e-08
2	150,003,486	150,049,352	14	2.87e-08
5	128,721,245	128,844,403	56	4.14e-09
15	39,569,380	39,643,906	40	4.27e-43
17	44,186,478	44,252,598	20	1.58e-10

4 DISCUSSION

This thesis explored the potential of bi-multivariate GWAS on mid-cortical brain shape by evaluating and comparing the performance and features of different grouping strategies. In addition, the recent GWAS conducted by Naqvi et al. (2021) on the same UK Biobank dataset allowed for comparison between a per-SNP approach to GWAS and different SNP-set approaches. Furthermore, it was demonstrated that SNP-set GWAS was successful in finding associations with genes and genetic loci known to be associated with the brain or known to be involved in the regulation of brain-related or morphology-related processes. Due to the reduced number of statistical tests, not only was SNP-set GWAS able to apply less stringent significance thresholds relative to a per-SNP GWAS, but the computation time was also drastically reduced. The subsequent paragraphs discuss the specific findings and insights gained in this thesis.

4.1 Influence of group size on the power of SNP-set GWAS

The haplotype-based and window-based (20 kb) GWAS detected associations within genes that the gene-based GWAS failed to detect despite having the least stringent significance thresholds. Hence, the smaller number of associations detected by gene-based GWAS is not only due to omitting intergenic SNPs from the analysis. The added 15 kb windows upstream and downstream of each gene alone make gene-based SNP-sets larger than the average haplotype block and 20 kb window. Tang & Ferreira (2012) have already demonstrated based on simulated data that increased group sizes lead to a decrease in power. Here, it was also demonstrated that larger group sizes result in fewer associations by conducting window-based GWAS with different window sizes, which in addition demonstrates the use of window-based GWAS as a flexible means to study the properties and features of SNP-set GWAS in general. These results could explain why gene-based GWAS was the least powerful SNP-set GWAS approach in this thesis. In addition, although the haplotype-based and window-based GWAS detected the same number of genome-wide significant loci, the haplotype-based GWAS was able to detect almost 60% more study-wide significant loci. It would make sense that the incorporation of more knowledge on correlational structures between SNPs (i.e., LD) results in increased power, however, considering that both grouping methods (i.e., haplotype-based and window-based grouping) result in different group size distributions, the better detection power of haplotype-based GWAS might be due to differences in group size alone, so further research is needed. These results show that group size and therefore grouping structure are influential variables when it comes to statistical power in CCA-based SNP-set GWAS.

4.2 SNP-set GWAS for the discovery of novel trait-associated loci

Because SNP-set GWAS tests multiple SNPs simultaneously, small effects of different SNPs are potentially combined, which makes them easier to detect. Hence, SNP-set GWAS has the potential to find associations that are very hard to detect in per-SNP GWAS. A recent GWAS

conducted by Naqvi et al. (2021) on the same individuals from the UK Biobank, and using the same phenotype extraction method from MRI images was used as the per-SNP GWAS reference to the SNP-set GWAS conducted in this work. Notably, their GWAS reported 472 and 242 genome-wide and study-wide significant loci respectively, which is several times more than any of the SNP-set methods here. However, the value of SNP-set GWAS should not be measured in terms of the number of associations, as it was never meant to replace per-SNP GWAS. Gene-based, haplotype-based, and window-based (20 kb) GWAS all detected associations that per-SNP GWAS failed to detect. This suggests that those loci harbor SNPs with effects too small to have been detected by per-SNP GWAS, and that SNP-set GWAS is an effective way to detect new associations in those regions by simultaneously analyzing multiple SNPs. Remarkably, none of the associations detected by window-based GWAS which were not detected by per-SNP GWAS reached the stringent study-wide significance threshold. This suggests that even when combined, the effects on the phenotype are still small. Nonetheless, SNP-set GWAS is a promising approach to discover novel genotype-phenotype associations. Loci detected by GWAS typically only explain a modest part of the total heritability of a trait, which is often estimated based on twin and family studies. The majority of this missing heritability is suggested to be accounted for by SNPs with very small effect sizes and rare SNPs (Marouli et al., 2017; Yang et al., 2010). Large cohorts, such as the UK Biobank have already boosted the discovery of new trait-related loci by providing researchers with large sample sizes. However, even large-scale studies to date fail to explain much of the missing heritability (Grasby et al., 2020). This thesis demonstrates that SNP-set GWAS is able to detect additional loci that per-SNP GWAS fails to detect, without the need for even larger sample sizes. As such it could be useful to further excavate the missing heritability. Since data, and especially imaging data is expensive, SNP-set GWAS is a useful complementary method to regular GWAS.

4.3 Effects of SNPs in SNP-set and per-SNP GWAS

This thesis proposes a simple measure to determine peak strength of GWAS signals which takes into account both the number of significant SNPs and the individual *P*-values contained within the peak. This measure of peak strength allows to rank the signals from a GWAS. Such ranking can be used to compare the results from different GWAS, specifically whether the strong signals of one GWAS are reproduced in the other GWAS. Here, peaks calculated on the summary data from the per-SNP GWAS of Naqvi et al. (2021) were used as the reference peaks. Only the strongest peaks had overlap with SNP-set GWAS peaks, suggesting that SNP-set GWAS is a more conservative version of GWAS for loci where individual SNP effects cannot be effectively combined.

An obvious requirement for individual SNP effects to be combined is that the genomic region which is tested for association harbors at least two SNPs that affect the phenotype. In addition, SNPs that are not related to the phenotype potentially ‘dilute’ the joint effect of a set of SNPs. This could explain why it is possible that a single SNP can be significantly associated with the phenotype, but when that SNP is simultaneously analyzed with other SNPs in a group, the group is not. Furthermore, the requirement of two or more SNPs with separate effects could help explain why an optimal range of window sizes was observed in terms of SNP-set unique loci (i.e., loci not detected by per-SNP GWAS). Small window sizes of 5 or 10 kb are statistically less likely to contain more than one effect SNP. Although at least two SNPs are required for

individual SNP effects to be combined, it could be that in reality more than two are necessary for SNP-sets to reach significance. On the other hand, the decrease in observed SNP-set unique loci for larger window sizes is likely due the negative effect of group size on power.

4.4 Limitations, suggestions and model extensions: towards more powerful SNP-set association testing

A main limitation of the gene-based GWAS is that it omits intergenic regions from the analysis. This was partially solved by extending the group boundaries by a window of 15 kb upstream and downstream of the gene. As such 90% of SNPs influencing the expression level of the gene are included its corresponding set (Pickrell et al., 2010). In addition, lincRNAs were given their respective set. This still left many intergenic SNPs out of the analysis, even though a substantial number of trait-associated SNPs has been detected in intergenic regions (Zou et al., 2020). A possible solution is to assign each SNP group membership of the nearest gene. As such, all SNPs across the genome would be assigned a group. This would, however, increase the size of gene-based SNP-sets even more, which based on results from this thesis would further decrease the power of such a gene-based GWAS. Better solutions that include every SNP in the analysis are possible, such as the use of a hybrid grouping strategy. In such a grouping approach, SNPs would still be grouped based on genes, with the addition that intergenic SNPs would be grouped based on LD structure, similar to haplotype blocks. In addition, large genes could also be subdivided to achieve a finer grouping structure. In this way, it would be possible that gene-based methods become competitive with haplotype-based and window-based methods, while still maintaining their gene-oriented nature.

Since window-based grouping methods are conceptually simple, there is relatively little room for improvement of these methods beyond varying the window size which was already demonstrated here. On the other hand, haplotype-based grouping methods still remain somewhat unexplored. Haplotype blocks could for example also be calculated on reference panels, or reference panels merged with the data in order to achieve better haplotype block estimation. Different algorithms exist for the calculation of such blocks (S. A. Kim et al., 2018), some of which might be better suited in the context of SNP-set GWAS because they avoid very large blocks for example. Moreover, LD blocks can be calculated through hierarchical SNP aggregation, and machine learning approaches can subsequently be applied to determine the optimal number of groups (Guinot et al., 2018). Whether such methods would result in a substantial improvement in power is unknown. Therefore, it cannot be concluded that more knowledge-integrated methods such as haplotype-based approaches have an edge over naïve methods such as sliding window approaches until the potential of haplotype-based methods in the context of SNP-set association testing is further explored.

Because SNPs needed to be pruned prior to SNP-set association analysis in order to avoid overfitting, the pruning algorithm offers another possibility for improvement. The degree of pruning potentially affects the power of the GWAS. In addition, in gene-based GWAS SNPs could be preferentially excluded from the analysis based on codon information. For example, SNPs in the third position of a codon are least likely to result in amino acid changes and are thus least likely to cause structural protein variants which could affect the phenotype. Still, because the abundance of tRNAs differs between corresponding codons and because some tRNAs are far less abundant than others, synonymous mutations can slow down translation

which affects protein abundance and therefore potentially also the phenotype (Kimchi-Sarfaty et al., 2007). Further research could point out whether the effects of different pruning algorithms are substantial.

Multivariate SNP-set GWAS is not limited to CCA as a test of association. Other bi-multivariate methods such as partial least squares regression (PLSR) (Wold et al., 1984) have been proposed (Bjørnstad et al., 2004; Mitteroecker et al., 2016). PLSR aims to find associations between two sets of variables by maximizing the variance between their linear combinations, similar to CCA, which maximizes their correlation. In addition, extensions to CCA have been proposed for association testing, notably sparse and (regularized) kernel CCA (Du et al., 2020; Le Floch et al., 2012). Sparse CCA (SCCA) applies a penalty to the canonical covariates while maximizing the canonical correlations, which forces the contributions of most SNPs within a set to zero. Such a penalty, often an ℓ_2-norm penalty can be applied to the phenotype as well resulting in a form of feature selection. The resulting sparse solution offers better interpretability and generalizability since the effects of irrelevant SNPs or features are nullified. Furthermore, kernel CCA (KCCA) is able to capture nonlinear genotype-phenotype associations by maximizing the canonical correlations between two sets of variables in a feature space (Yamanishi et al., 2003). Doing so allows for modeling complex relations between a group of SNPs and the phenotype as well as model epistatic effects within the SNP-set (Wu et al., 2010). As such, kernel methods could be an especially useful extension to the CCA-based SNP-set tests of association here proposed for detecting novel loci and further excavate the missing heritability of diverse complex traits.

4.5 Genomic control in SNP-set GWAS

Test statistics generated by CCA are asymptotically χ^2 null distributed with degrees of freedom depending on the number of SNPs in the set. Hence, a SNP-set approach to GWAS generates differently null distributed test statistics because sets of SNPs vary in size. This thesis proposes an adapted implementation of the genomic control factor based on the mean test statistic to overcome the difference in degrees of freedom. The proposed implementation of the genomic control factor was demonstrated on different SNP-set GWAS and showed only a minute inflation of the mean test statistic (< 2% across all brain segments and all SNP-set GWAS) illustrating that population stratification was not of concern.

In addition, this thesis demonstrates how pooling of the test statistics can be an effective strategy to investigate uniformity of inflation across differently null distributed test statistics. Through this approach, it was shown that for narrow window sizes (i.e., the spanned genomic distance), the largest groups (in terms of SNP-count) had inflated test statistics, thus suggesting an effect of larger group sizes. Interestingly, for wider window sizes, the largest groups did not have inflated test statistics despite having a far greater number of SNPs in comparison with the largest groups corresponding to the narrower windows. Since the CCA model is unaware of the window size, only of the number of SNPs per group, these results show that it is not just the number of simultaneously tested SNPs that cause inflation of the test statistic. Rather, they suggest that the inflation is actually caused by a relatively large number of SNPs within a narrow genomic window, i.e., high local SNP density and potentially its implications on the degree of correlation between SNPs. Despite the inflation of test statistics for the densest groups, it was illustrated that the novel associations (i.e., associated

SNP-sets that did not harbor any significant individual SNPs from the per-SNP GWAS (Naqvi et al., 2021) and were at least 1 Mb away from significant individual SNPs) all had small to moderate group sizes. This is important, because it means that the inflation of test statistics did not lie at the basis of SNP-set unique associations. It was demonstrated that pruning of the SNP data to about $1/9^{th}$ its density was necessary to avoid magnifying false discovery rate. This further suggests that extremely dense SNP-sets better be avoided. Worth noting is that inflation of the test statistic was observed in fewer than the 100 largest groups for every window size. Apart from those groups, no inflation of the test statistic was observed. This thesis proposes and demonstrates a correction for the inflation factor per pool of test statistics. It was illustrated that this correction had only a minor effect on the *P*-values. To avoid potentially large corrections, it would be better to further prune SNPs in the densest SNP-sets before association testing. Based on these results, I recommend that pooling test statistics and calculating the inflation of the mean test statistic per pool becomes a standard quality control method for SNP-set GWAS. Moreover, it was shown that with increasing window size, the test statistic was deflated more for the smallest SNP-sets. This suggests that sparse SNP-sets should probably be avoided as they lead to decreased power. One solution could be to replace each sparse set by its two densest subsets.

4.6 Generalizability of bi-multivariate associations

The relations found by CCA are known to not generalize well on unseen data (Le Floch et al., 2012). Cross-validation is a known strategy to assess their generalizability, which applies the canonical weights learned on the training data to unseen validation data (Dinga et al., 2019). If the corresponding canonical correlations are of similar strength in the validation data, the association is generalizable. The threshold to determine 'similar strength' is often arbitrary. Although this strategy is moderately helpful when the number of associations to inspect is small, it is not practical in a high throughput context such as in GWAS, which could benefit from a more formal test of generalizability. This thesis assessed generalizability of associations for diagnostic purposes through linear regression of canonical covariates in the validation set. Such an approach was useful to distinguish obvious false positive from true positive associations. Due to substantial fluctuations in the *F*-statistic between different replicates, a median *F*-statistic over the different replicates was used to make the test more robust. Still, this approach should in its current form be limited to assess extreme cases only as its general performance is not well understood. Moreover, it has been suggested that the first canonical correlation is not necessarily the most generalizable one, and instead the second canonical correlation sometimes is (Le Floch et al., 2012).

In general, it is hard to determine which part of the CCA decomposition captures the generalizable trend (Wang et al., 2020). This problem is still largely unanswered and has implications for prediction and reconstruction of a phenotype from genotypic data, which could be of interest for the craniofacial system, as well as the brain (Makowsky et al., 2011). For prediction, the relation between phenotype and genotype obtained through GWAS is applied to an individual's genotypic data. As such, prediction of a complex, multivariate trait is similar to calculating a polygenic risk score for a dichotomous disease. Note that the problem simplifies if only one of the data views is multidimensional, in which case CCA becomes essentially equivalent to univariate multiple regression or inversed univariate multiple regression (Hotelling, 1936).

Another issue related to prediction is the effect sizes for bi-multivariate associations from CCA. Estimates for the variance explained by a scored set of SNPs or locus are overoptimistic due to the inherent property of CCA to find maximally correlated sets of canonical covariates. As such, more realistic estimates of the effect size can be obtained via cross-validation methods, where the variance explained is assessed on unseen data (Gianola et al., 2014; Makowsky et al., 2011). Whether bi-multivariate GWAS results are better or worse suited for prediction in comparison to per-SNP multivariate GWAS is therefore unsure until more research is done.

4.7 Bi-multivariate association testing to study Neanderthal influences on present-day humans

Fairly recently the genomes of several Neanderthals were sequenced, and subsequent analysis revealed that present-day non-African humans have a substantial amount of Neanderthal DNA in their genomes (Green et al., 2010). This DNA, present as haplotype blocks (i.e., interogressed haplotype blocks) results from interbreeding events between Neanderthals and our modern human ancestors (Prüfer et al., 2014). From paleontological discoveries it was learned that the Neanderthal craniofacial system looks different compared to that of modern humans (Ponce de León & Zollikofer, 2001), and it was hence hypothesized that it influences the underlying brain anatomy (Pearce et al., 2013). Interestingly, studies have suggested that some Neanderthal alleles were under positive selection pressure in modern humans, including alleles related to cognition and cranial morphology (Green et al., 2010). Potentially, these Neanderthal alleles affect the craniofacial system, brain morphology, and/or cognitive functions in present-day humans. This thesis has shown how SNP-set association testing can be used to study specific regions of interest in the genome such as interogressed Neanderthal haplotype blocks. The results show that SNPs located within the boundaries of Neanderthal interogressed haplotype blocks affect cortical brain shape and suggest that the Neanderthal alleles could influence brain morphology. This SNP-set GWAS conducted based on the coordinates of Neanderthal interogressed haplotype blocks was among the first steps towards 1) understanding how Neanderthal DNA affects the brain of present-day humans, and 2) what the cognitive effects are of Neanderthal alleles.

A few recent publications explored similar ideas by applying a per-SNP GWAS approach, however, they found only 1 (Gregory et al., 2017) and 3 (Gunz et al., 2019) brain-associated loci, whereas the SNP-set approach in this thesis found 6 after applying the stringent study-wide significance threshold. This suggests that SNP-set approaches are possibly beneficial to study the effects of ancient alleles in modern genomes.

Effects of trait-associated alleles found through GWAS based on interogressed DNA from ancient hominins could be combined into a polygenic score. Such a score could be calculated on the genomes of present-day humans to estimate and model the effects of ancient alleles on their craniofacial system for example. Nonetheless, polygenic scores based on a single GWAS are unreliable, so associations should first be validated in independent GWAS.

5 CONCLUSION

This thesis has explored the features and performance of bi-multivariate association testing using the CCA framework. It was demonstrated that SNP-set GWAS is a useful approach for finding both known and novel genotype-phenotype associations with complex multivariate traits such as cortical brain shape. SNP-set GWAS was able to discover genotype-phenotype associations that per-SNP GWAS failed to find, and as such is a good complementary method to per-SNP GWAS and a promising approach to help further excavate the so-called missing heritability. In addition, the multiple testing burden was reduced, resulting in a less stringent threshold compared to per-SNP GWAS, and the fewer number of tests resulted in reduced computation time.

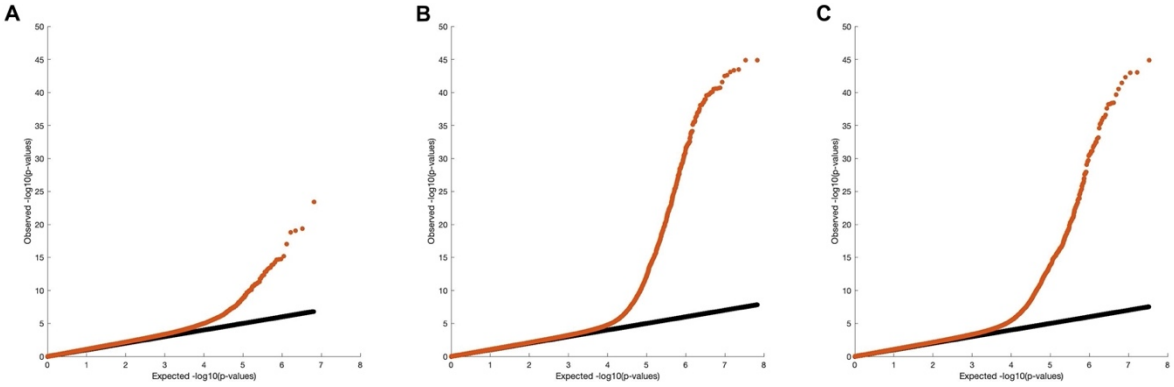
Different grouping strategies resulted in different GWAS results, illustrating that grouping structure plays a role in the detection of associations. Window-based GWAS was found to be a versatile approach to gain insights in the effect of group size. It was observed that larger group sizes result in fewer associations, which could explain in part why gene-based GWAS detected fewer associations compared to haplotype-based and window-based (20 kb) GWAS. In addition, it was found that an optimal range of window sizes exist which maximally harnesses the potential of SNP-set GWAS to find novel loci affecting the phenotype.

Results from this thesis suggest that SNP-set GWAS is a more conservative version of GWAS for groups in which individual effect cannot be effectively combined. It is suggested that irrelevant SNPs dilute the effects of trait-relevant SNPs.

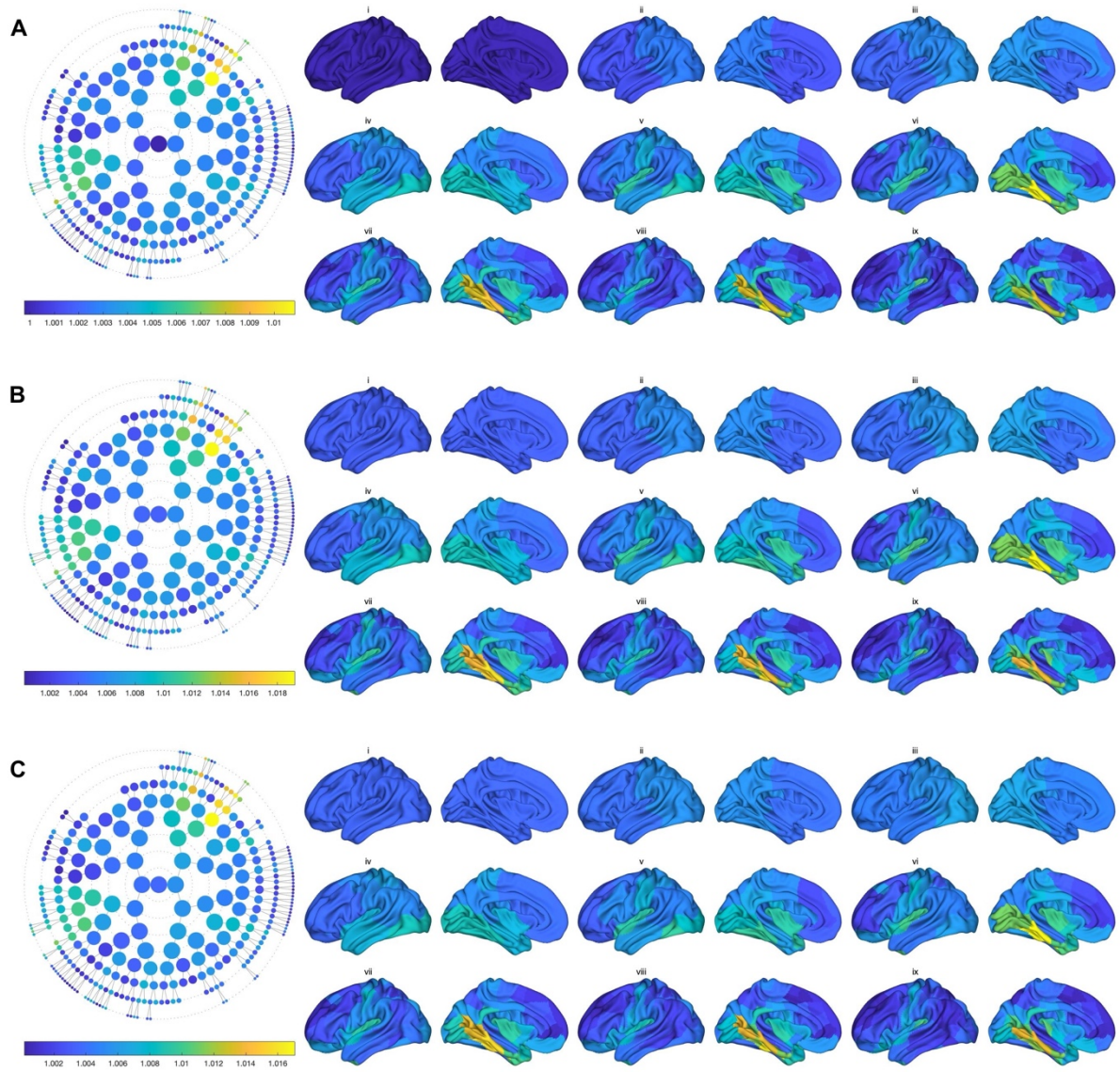
Furthermore, this thesis proposes an adapted version of the genomic control factor that can be used with the bi-multivariate CCA-based framework. Pooling test statistics per number of degrees of freedom and calculating the degree of test statistic inflation is proposed and recommended as a quality control measure to guarantee no bias for dense SNP-sets. The inflation factors per pool can subsequently be used for correction of the test statistics.

Looking forward, the power of SNP-set GWAS could be improved by optimizing the pre-association testing steps, notably pruning and grouping SNPs, as well as by considering model extensions such as kernel CCA.

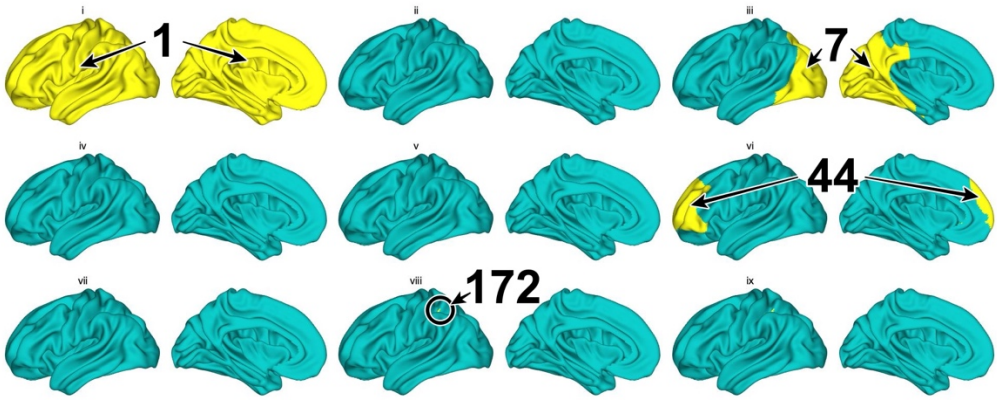
Supplementary material



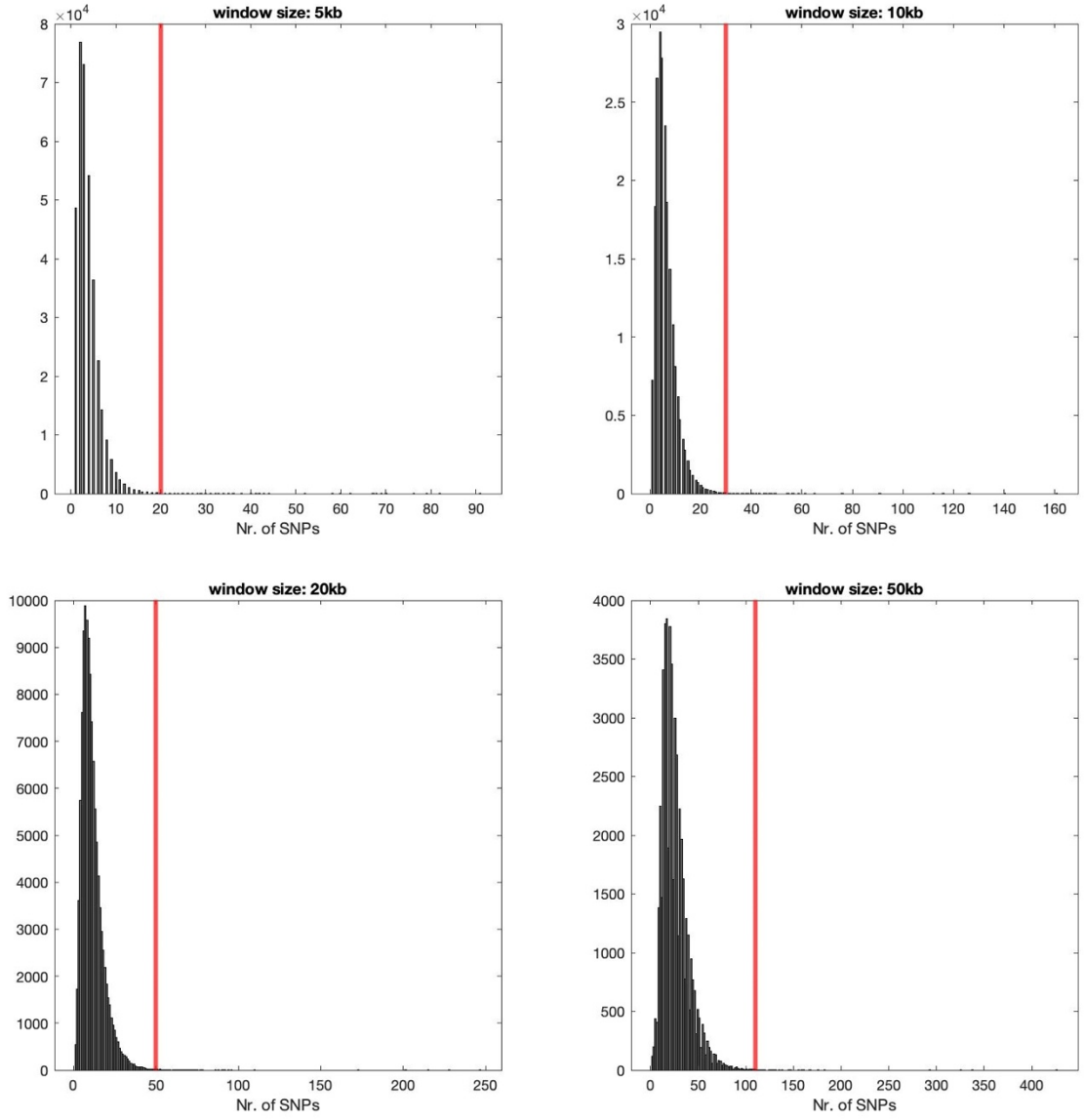
Supplementary figure 1 Quantile-quantile plots for SNP-set GWAS based on (A) genes, (B) haplotype blocks, and (C) a 20 kb sliding window.



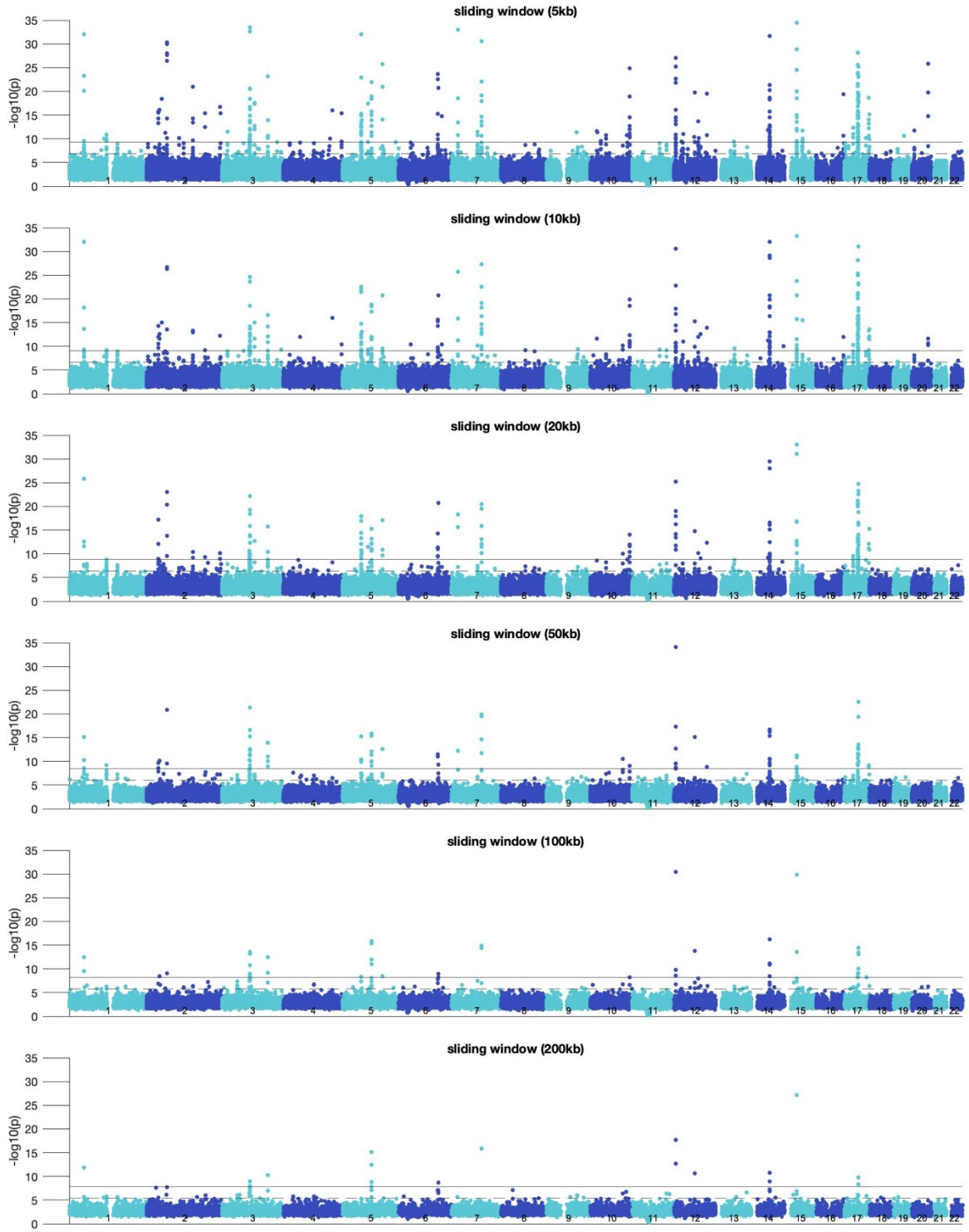
Supplementary figure 2 Genomic control factor, lambda, per hierarchical brain segment. The genomic control factors corresponding to **A** gene-based GWAS, **B** haplotype-based GWAS, and **C** window-based GWAS are represented by colors. Scales are set relative to the maximum number of associations for each SNP-set GWAS method in order to compare trends between different methods. Lower-case Roman numerals indicate the hierarchical level corresponding to circles in the polar dendrogram.



Supplementary figure 3 Brain segments 1 – 7 – 44 – 172, indicated in yellow. These segments have a dimensionality of 437, 170, 90, and 10 respectively.



Supplementary figure 4 Group size distributions for 5 – 10 – 20 – 50 kb widow-based GWAS. Groups with size right of the vertical red line were affected by inflated test statistics. These account for fewer than 100 groups per GWAS.



Supplementary figure 5 Manhattan plots for window-based GWAS. Window sizes are: 5 – 10 – 20 – 50 – 100 – 200 kb. SNP-sets are represented by the minimal *P*-value obtained across all 285 phenotypes. The Bonferroni corrected genome-wide significance threshold is indicated by a dashed horizontal line, and the Bonferroni corrected study-wide significance threshold is indicated by a full horizontal line/

References

Astafiev, S. V., Shulman, G. L., Stanley, C. M., Snyder, A. Z., Essen, D. C. V., & Corbetta, M. (2003). Functional Organization of Human Intraparietal and Frontal Cortex for Attending, Looking, and Pointing. *Journal of Neuroscience*, *23*(11), 4689–4699. https://doi.org/10.1523/JNEUROSCI.23-11-04689.2003

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bjørnstad, Å., Westad, F., & Martens, H. (2004). Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (PLSR). *Hereditas*, *141*(2), 149–165. https://doi.org/10.1111/j.1601-5223.2004.01816.x

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

Braz, C. U., Taylor, J. F., Bresolin, T., Espigolan, R., Feitosa, F. L. B., Carvalheiro, R., Baldi, F., de Albuquerque, L. G., & de Oliveira, H. N. (2019). Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. *BMC Genetics*, *20*. https://doi.org/10.1186/s12863-019-0713-4

Buil, A., Martinez-Perez, A., Perera-Lluna, A., Rib, L., Caminal, P., & Soria, J. M. (2009). A new gene-based association test for genome-wide association studies. *BMC Proceedings*, *3*(7), S130. https://doi.org/10.1186/1753-6561-3-S7-S130

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Claes, P., Roosenboom, J., White, J. D., Swigut, T., Sero, D., Li, J., Lee, M. K., Zaidi, A., Mattern, B. C., Liebowitz, C., Pearson, L., González, T., Leslie, E. J., Carlson, J. C., Orlova, E., Suetens, P., Vandermeulen, D., Feingold, E., Marazita, M. L., … Weinberg, S. M. (2018). Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics*, *50*(3), 414–423. https://doi.org/10.1038/s41588-018-0057-4

Culham, J. C., & Valyear, K. F. (2006). Human parietal cortex in action. *Current Opinion in Neurobiology*, *16*(2), 205–212. https://doi.org/10.1016/j.conb.2006.03.005

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., Hagenaars, S. P., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., Liewald, D. C. M., Okely, J. A., Ahola-Olli, A. V., Barnes, C. L. K., Bertram, L., Bis, J. C., Burdick, K. E., Christoforou, A., DeRosse, P., … Deary, I. J. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature Communications*, *9*(1), 2098. https://doi.org/10.1038/s41467-018-04362-x

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lõhmussaar, E., Zernant, J., Tõnisson, N., Remm, M., Mägi, R., … Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, *418*(6897), 544–548. https://doi.org/10.1038/nature00864

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-

Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261. https://doi.org/10.1038/ng.3760

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x

Dickie, E. W., Anticevic, A., Smith, D. E., Coalson, T. S., Manogaran, M., Calarco, N., Viviano, J. D., Glasser, M. F., Van Essen, D. C., & Voineskos, A. N. (2019). Ciftify: A framework for surface-based analysis of legacy MR acquisitions. *NeuroImage*, *197*, 818–826. https://doi.org/10.1016/j.neuroimage.2019.04.078

Dinga, R., Schmaal, L., Penninx, B. W. J. H., van Tol, M. J., Veltman, D. J., van Velzen, L., Mennes, M., van der Wee, N. J. A., & Marquand, A. F. (2019). Evaluating the evidence for biotypes of depression: Methodological replication and extension of Drysdale et al. (2017). *NeuroImage: Clinical*, *22*, 101796. https://doi.org/10.1016/j.nicl.2019.101796

Donati, G., Dumontheil, I., Pain, O., Asbury, K., & Meaburn, E. L. (2021). Evidence for specificity of polygenic contributions to attainment in English, maths and science during adolescence. *Scientific Reports*, *11*(1), 3851. https://doi.org/10.1038/s41598-021-82877-y

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Saykin, A. J., Guo, L., & Shen, L. (2020). Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach. *Medical Image Analysis*, *61*, 101656. https://doi.org/10.1016/j.media.2020.101656

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K., Douaud, G., Marchini, J., & Smith, S. (2018). Genome-wide association studies of brain structure and function in the UK Biobank. *BioRxiv*, 178806. https://doi.org/10.1101/178806

Ferreira, M. A. R., & Purcell, S. M. (2009). A multivariate test of association. *Bioinformatics*, *25*(1), 132–133. https://doi.org/10.1093/bioinformatics/btn563

Fraser, J. A., Newman, N. J., & Biousse, V. (2011). Chapter 8—Disorders of the optic tract, radiation, and occipital lobe. In C. Kennard & R. J. Leigh (Eds.), *Handbook of Clinical Neurology* (Vol. 102, pp. 205–221). Elsevier. https://doi.org/10.1016/B978-0-444-52903-9.00014-5

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., … The SNP Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

Génin, E. (2020). Missing heritability of complex diseases: Case solved? *Human Genetics*, *139*(1), 103–113. https://doi.org/10.1007/s00439-019-02034-4

Gianola, D., Weigel, K. A., Krämer, N., Stella, A., & Schön, C.-C. (2014). Enhancing genome-enabled prediction by bagging genomic BLUP. *PloS One*, *9*(4), e91693. https://doi.org/10.1371/journal.pone.0091693

Glynn, W. J., & Muirhead, R. J. (1978). Inference in canonical correlation analysis. *Journal of Multivariate Analysis*, *8*(3), 468–478. https://doi.org/10.1016/0047-259X(78)90067-2

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Grasby, K. L., Jahanshad, N., Painter, J. N., Colodro-Conde, L., Bralten, J., Hibar, D. P., Lind, P. A., Pizzagalli, F., Ching, C. R. K., McMahon, M. A. B., Shatokhina, N., Zsembik, L. C. P., Thomopoulos, S. I., Zhu, A. H., Strike, L. T., Agartz, I., Alhusaini, S., Almeida, M. A. A., Alnæs, D., … Group, E. N. G. through M.-A. C. (ENIGMA)—Genetics working. (2020). The genetic architecture of the human cerebral cortex. *Science*, *367*(6484). https://doi.org/10.1126/science.aay6690

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A Draft Sequence of the Neandertal Genome. *Science (New York, N.Y.)*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Gregory, M. D., Kippenhan, J. S., Eisenberg, D. P., Kohn, P. D., Dickinson, D., Mattay, V. S., Chen, Q., Weinberger, D. R., Saad, Z. S., & Berman, K. F. (2017). Neanderthal-Derived Genetic Variation Shapes Modern Human Cranium and Brain. *Scientific Reports*, *7*(1), 6308. https://doi.org/10.1038/s41598-017-06587-0

Guinot, F., Szafranski, M., Ambroise, C., & Samson, F. (2018). Learning the optimal scale for GWAS through hierarchical SNP aggregation. *BMC Bioinformatics*, *19*(1), 459. https://doi.org/10.1186/s12859-018-2475-9

Gunz, P., Tilot, A. K., Wittfeld, K., Teumer, A., Shapland, C. Y., van Erp, T. G. M., Dannemann, M., Vernot, B., Neubauer, S., Guadalupe, T., Fernández, G., Brunner, H. G., Enard, W., Fallon, J., Hosten, N., Völker, U., Profico, A., Di Vincenzo, F., Manzi, G., … Fisher, S. E. (2019). Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Current Biology*, *29*(1), 120-127.e5. https://doi.org/10.1016/j.cub.2018.10.065

Guo, S.-W., & Lange, K. (2000). Genetic Mapping of Complex Traits: Promises, Problems, and Prospects. *Theoretical Population Biology*, *57*(1), 1–11. https://doi.org/10.1006/tpbi.2000.1449

Hamazaki, K., & Iwata, H. (2020). RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLOS Computational Biology*, *16*(2), e1007663. https://doi.org/10.1371/journal.pcbi.1007663

Haraksingh, R. R., & Snyder, M. P. (2013). Impacts of Variation in the Human Genome on Gene Regulation. *Journal of Molecular Biology*, *425*(21), 3970–3977. https://doi.org/10.1016/j.jmb.2013.07.015

Heimer, L., & Van Hoesen, G. W. (2006). The limbic lobe and its output channels: Implications for emotional functions and adaptive behavior. *Neuroscience & Biobehavioral Reviews*, *30*(2), 126–147. https://doi.org/10.1016/j.neubiorev.2005.06.006

Hirschhorn, J. N. (2009). Genomewide Association Studies—Illuminating Biologic Pathways. *New England Journal of Medicine*, *360*(17), 1699–1701. https://doi.org/10.1056/NEJMp0808934

Hofer, E., Roshchupkin, G. V., Adams, H. H. H., Knol, M. J., Lin, H., Li, S., Zare, H., Ahmad, S., Armstrong, N. J., Satizabal, C. L., Bernard, M., Bis, J. C., Gillespie, N. A., Luciano, M., Mishra, A., Scholz, M., Teumer, A., Xia, R., Jian, X., … Seshadri, S. (2020). Genetic correlations and genome-wide associations of cortical structure in general population samples of 22,824 adults. *Nature Communications*, *11*(1), 4796. https://doi.org/10.1038/s41467-020-18367-y

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, *28*(3/4), 321–377. https://doi.org/10.2307/2333955

Jardri, R., Pouchet, A., Pins, D., & Thomas, P. (2011). Cortical Activations During Auditory Verbal Hallucinations in Schizophrenia: A Coordinate-Based Meta-Analysis. *American Journal of Psychiatry*, *168*(1), 73–81. https://doi.org/10.1176/appi.ajp.2010.09101522

Kim, J., Zhang, Y., Pan, W., & for the Alzheimer's Disease Neuroimaging Initiative. (2016). Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. *Genetics*, *203*(2), 715–731. https://doi.org/10.1534/genetics.115.186502

Kim, S. A., Cho, C.-S., Kim, S.-R., Bull, S. B., & Yoo, Y. J. (2018). A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, *34*(3), 388–397. https://doi.org/10.1093/bioinformatics/btx609

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., & Gottesman, M. M. (2007). A 'Silent' Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science*, *315*(5811), 525–528. https://doi.org/10.1126/science.1135308

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, *308*(5720), 385–389. https://doi.org/10.1126/science.1109557

Kotz, S. A., & Paulmann, S. (2011). Emotion, Language, and the Brain. *Language and Linguistics Compass*, *5*(3), 108–125. https://doi.org/10.1111/j.1749-818X.2010.00267.x

Kruglyak, L., & Nickerson, D. A. (2001). Variation is the spice of life. *Nature Genetics*, *27*(3), 234–236.

https://doi.org/10.1038/85776

Lam, M., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B. C., Liu, R., Zhou, W., Guan, L., Kamatani, Y., Kim, S.-W., Kubo, M., Kusumawardhani, A. A. A. A., Liu, C.-M., Ma, H., … Huang, H. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nature Genetics*, *51*(12), 1670–1678. https://doi.org/10.1038/s41588-019-0512-x

Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.-B., & Duchesnay, É. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage*, *63*(1), 11–24. https://doi.org/10.1016/j.neuroimage.2012.06.061

Lech, R. K., & Suchan, B. (2013). The medial temporal lobe: Memory and beyond. *Behavioural Brain Research*, *254*, 45–49. https://doi.org/10.1016/j.bbr.2013.06.009

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., Song, Z., Ji, W., Wang, M., Zhou, J., Chen, B., Liu, Y., Wang, J., Wang, P., Yang, P., … Shi, Y. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics*, *49*(11), 1576–1583. https://doi.org/10.1038/ng.3973

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., & Macgregor, S. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics*, *87*(1), 139–145. https://doi.org/10.1016/j.ajhg.2010.06.009

Luciano, M., Hagenaars, S. P., Davies, G., Hill, W. D., Clarke, T.-K., Shirali, M., Harris, S. E., Marioni, R. E., Liewald, D. C., Fawns-Ritchie, C., Adams, M. J., Howard, D. M., Lewis, C. M., Gale, C. R., McIntosh, A. M., & Deary, I. J. (2018). Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature Genetics*, *50*(1), 6–11. https://doi.org/10.1038/s41588-017-0013-8

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., & de los Campos, G. (2011). Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genetics*, *7*(4). https://doi.org/10.1371/journal.pgen.1002051

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., Rüeger, S., Thorleifsson, G., Justice, A. E., Lamparter, D., Stirrups, K. E., Turcot, V., Young, K. L., Winkler, T. W., Esko, T., … Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, *542*(7640), 186–190. https://doi.org/10.1038/nature21039

Mauk, M. D., Medina, J. F., Nores, W. L., & Ohyama, T. (2000). Cerebellar function: Coordination, learning or timing? *Current Biology*, *10*(14), R522–R525. https://doi.org/10.1016/S0960-9822(00)00584-4

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Merrill, A. E., Bochukova, E. G., Brugger, S. M., Ishii, M., Pilz, D. T., Wall, S. A., Lyons, K. M., Wilkie, A. O. M., & Maxson, R. E. (2006). Cell mixing at a neural crest-mesoderm boundary and deficient ephrin-Eph signaling in the pathogenesis of craniosynostosis. *Human Molecular Genetics*, *15*(8), 1319–1328. https://doi.org/10.1093/hmg/ddl052

Mills, M. C., & Rahal, C. (2019). A scientometric review of genome-wide association studies. *Communications Biology*, *2*(1), 1–11. https://doi.org/10.1038/s42003-018-0261-x

Mitteroecker, P., Cheverud, J. M., & Pavlicev, M. (2016). Multivariate Analysis of Genotype–Phenotype Association. *Genetics*, *202*(4), 1345–1363. https://doi.org/10.1534/genetics.115.181339

Murthy, A., Li, Y., Peng, I., Reichelt, M., Katakam, A. K., Noubade, R., Roose-Girma, M., DeVoss, J., Diehl, L., Graham, R. R., & van Lookeren Campagne, M. (2014). A Crohn's disease variant in Atg16l1 enhances its degradation by caspase 3. *Nature*, *506*(7489), 456–462. https://doi.org/10.1038/nature13044

Naqvi, S., Sleyp, Y., Hoskens, H., Indencleef, K., Spence, J. P., Bruffaerts, R., Radwan, A., Eller, R. J., Richmond, S., Shriver, M. D., Shaffer, J. R., Weinberg, S. M., Walsh, S., Thompson, J., Pritchard, J. K., Sunaert, S., Peeters, H., Wysocka, J., & Claes, P. (2021). Shared heritability of human face and brain shape. *Nature Genetics*, 1–10. https://doi.org/10.1038/s41588-021-00827-w

Nowinski, W. L. (2011). Introduction to Brain Anatomy. In K. Miller (Ed.), *Biomechanics of the Brain* (pp. 5–40). Springer. https://doi.org/10.1007/978-1-4419-9997-9_2

Parvizi, J., & Damasio, A. (2001). Consciousness and the brainstem. *Cognition*, *79*(1), 135–160. https://doi.org/10.1016/S0010-0277(00)00127-X

Pearce, E., Stringer, C., & Dunbar, R. I. M. (2013). New insights into differences in brain organization between Neanderthals and anatomically modern humans. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758). https://doi.org/10.1098/rspb.2013.0168

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

Ponce de León, M. S., & Zollikofer, C. P. E. (2001). Neanderthal cranial ontogeny and its implications for late hominid diversity. *Nature*, *412*(6846), 534–538. https://doi.org/10.1038/35087573

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., … Pääbo, S. (2014). The complete genome sequence of a Neandertal from the Altai Mountains. *Nature*, *505*(7481), 43–49. https://doi.org/10.1038/nature12886

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

Reich, D. E., & Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology*, *20*(1), 4–16. https://doi.org/10.1002/1098-2272(200101)20:1<4::AID-GEPI2>3.0.CO;2-T

Robert, P., & Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *25*(3), 257–265. https://doi.org/10.2307/2347233

Salzberg, S. L. (2018). Open questions: How many genes do we have? *BMC Biology*, *16*(1), 94. https://doi.org/10.1186/s12915-018-0564-x

Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., Zhao, F., Liao, L., Chen, J., Lin, Y., Tian, Q., Papasian, C. J., & Deng, H.-W. (2013). Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLOS ONE*, *8*(4), e59494. https://doi.org/10.1371/journal.pone.0059494

Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., Foroud, T., Pankratz, N., Moore, J. H., Sloan, C. D., Huentelman, M. J., Craig, D. W., DeChairo, B. M., Potkin, S. G., Jack, C. R., Weiner, M. W., & Saykin, A. J. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *NeuroImage*, *53*(3), 1051–1063. https://doi.org/10.1016/j.neuroimage.2010.01.042

Shen, L., & Thompson, P. M. (2020). Brain Imaging Genomics: Integrated Analysis and Machine Learning. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, *108*(1), 125–162. https://doi.org/10.1109/JPROC.2019.2947272

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., DeChairo, B. M., Potkin, S. G., Weiner, M. W., & M. Thompson, P. (2010). Voxelwise genome-wide association study (vGWAS). *NeuroImage*, *53*(3), 1160–1174. https://doi.org/10.1016/j.neuroimage.2010.02.032

Sud, A., Kinnersley, B., & Houlston, R. S. (2017). Genome-wide association studies of cancer: Current insights and future perspectives. *Nature Reviews Cancer*, *17*(11), 692–704. https://doi.org/10.1038/nrc.2017.82

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tang, C. S., & Ferreira, M. A. R. (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics*, *28*(6), 845–850. https://doi.org/10.1093/bioinformatics/bts051

van der Meer, D., Frei, O., Kaufmann, T., Shadrin, A. A., Devor, A., Smeland, O. B., Thompson, W. K., Fan, C. C., Holland, D., Westlye, L. T., Andreassen, O. A., & Dale, A. M. (2020). Understanding the genetic determinants of the brain with MOSTest. *Nature Communications*, *11*(1), 3512. https://doi.org/10.1038/s41467-020-17368-1

Veerappa, A. M., Saldanha, M., Padakannaya, P., & Ramachandra, N. B. (2014). Family based genome-wide copy number scan identifies complex rearrangements at 17q21.31 in dyslexics. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *165*(7), 572–580. https://doi.org/10.1002/ajmg.b.32260

Vernot, B., Tucci, S., Kelso, J., Schraiber, J. G., Wolf, A. B., Gittelman, R. M., Dannemann, M., Grote, S., McCoy, R. C., Norton, H., Scheinfeldt, L. B., Merriwether, D. A., Koki, G., Friedlaender, J. S., Wakefield, J., Pääbo, S., & Akey, J. M. (2016). Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science*, *352*(6282), 235–239. https://doi.org/10.1126/science.aad9416

Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *4*(8), 587–597. https://doi.org/10.1038/nrg1123

Wang, H.-T., Smallwood, J., Mourao-Miranda, J., Xia, C. H., Satterthwaite, T. D., Bassett, D. S., & Bzdok, D. (2020). Finding the needle in a high-dimensional haystack: Canonical correlation analysis for neuroscientists. *NeuroImage*, *216*, 116745. https://doi.org/10.1016/j.neuroimage.2020.116745

Wold, S., Ruhe, A., Wold, H., & Dunn, I., W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, *5*(3), 735–743. https://doi.org/10.1137/0905052

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics*, *86*(6), 929–942. https://doi.org/10.1016/j.ajhg.2010.05.002

Yamanishi, Y., Vert, J.-P., Nakaya, A., & Kanehisa, M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, *19*(suppl_1), i323–i330. https://doi.org/10.1093/bioinformatics/btg1045

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. https://doi.org/10.1038/ng.608

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., Alzheimer's Disease Neuroimaging Initiative, Pediatric Imaging, Neurocognition and Genetics, & Zhu, H. (2019). Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature Genetics*, *51*(11), 1637–1644. https://doi.org/10.1038/s41588-019-0516-6

Zhou, H., Polimanti, R., Yang, B.-Z., Wang, Q., Han, S., Sherva, R., Nuñez, Y. Z., Zhao, H., Farrer, L. A., Kranzler, H. R., & Gelernter, J. (2017). Genetic Risk Variants Associated With Comorbid Alcohol Dependence and Major Depression. *JAMA Psychiatry*, *74*(12), 1234–1241. https://doi.org/10.1001/jamapsychiatry.2017.3275

Zou, H., Wu, L.-X., Tan, L., Shang, F.-F., & Zhou, H.-H. (2020). Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.00347


1425