KU LEUVEN

Global-To-Local Segmentation and Genotypic Analysis Of Brain Shape Asymmetry

Vasileios Lemonidis

Thesis submitted for the degree of
Master of Science in Bioinformatics

Thesis supervisor:
Prof. Peter Claes
Dept. of Electrical Engineering
Dept. of Human Genetics

Prof. Isabelle Cleynen
Dept. of Human Genetics

Mentors:
MSc. Meng Yuan
MSc. Seppe Goovaerts

Academic year 2022

Preface

The text of the preface. A few paragraphs should follow.

The Author
1 January 2010

Contents

List of Figures

List of Tables

Abbreviations

3D three-dimensional

AD Alzheimer's disease

ADHD attention-deficit / hyperactivity-disorder

AI asymmetry index

ANOVA analysis of variance

ASD autism spectrum disorder

BD bipolar disorder

CCA canonical correlation analysis

CNS central neural system

D-V dorsal-ventral

DA directional asymmetry

DK Desikan-Killiany

DNA deoxyribonucleic acid

DOF degree of freedom

FA fluctuating asymmetry

FDR false discovery rate

GLM generalized linear model

GO gene ontology

GPA generalized Procrustes analysis

GRM genetic relationship matrix

GSEA gene set enrichment analysis

GWAS genome-wide association studies

HCP Human Connectome Project

HSC hierarchical spectral clustering

L-R left-right

LD linkage disequilibrium

LDSC LD score correlation

LDSR LD score regression

LFC language functional connectivity

MAF minor allele frequency

MDD major depressive disorder

ML machine learning

MRI magnetic resonance imaging

MSS mean sum of squares

mvGWAS multivariate genome-wide association study

NMI normalized mutual information

NPC neuroepithelial cell

OCD obsessive/compulsive disorder

PC principal component

PCA principal component analysis

PLSR partial least squares regression

R-C rostral-caudal

RGC radial glial cell

RNA ribonucleic acid

RSS residual sum of squares

SNP single nucleotide polymorphism

TF transcription factor

Abstract

The overall purpose of this thesis is to complement the existing bibliography on the detection and examination of the genetic associations of brain shape asymmetry. Asymmetry components are computed based on the brain magnetic resonance imaging (MRI) dataset provided by UK Biobank database. A data-driven approach is followed, where the brain surface is partitioned in an unsupervised manner, through hierarchical spectral clustering (HSC), a technique that allows for a coarse-to-fine segmentation. Aggregated asymmetry measurements are retrieved from the segments, whose genetic correlation is examined through a multivariate genome-wide association study (mvGWAS) statistical analysis. Recognized significant single nucleotide polymorphisms (SNPs) are then analyzed individually or in groups, through comparison with existing results and databases. The genetic overlap with neurodevelopmental disorders and traits, that have been reported to exhibit phenotypic associations with brain structure asymmetry, such as autism, Alzheimer's disease or intelligence, are examined. Functional annotations of variants associated with the genes where significant SNPs were detected are constructed, offering an insight into the functional reasoning behind the brain shape asymmetry existence.

Chapter 1

Introduction

1.1 Biomedical and anatomic principles

1.1.1 Bilateria lineage

Cerebral bilateral symmetry is a universal quality of organisms belonging to the Bilateria lineage [26, 27], the phylum incorporating all species with a single plane of symmetry, in contrast with their sister group, Cnidaria (Figure 1.1). Bilateral symmetry is a byproduct of the activity of two separate developmental processes. Those produce two axes of polarity [45], and therefore a symmetry plane. Firstly, the formation of primary body axis, that corresponds to the long anatomical dimension of the animal, called rostral-caudal (R-C) (i.e. head-to-tail), is primarily dictated by highly conserved controlled activation of HOX genes during cell differentiation. Secondly, the shaping of a secondary body axis, orthogonal to R-C, named dorsal-ventral (D-V) (i.e. back-to-front), is attributed to a variety of genes, such as the chromatin organizer CTCF, the left-right determination factor Nodal and central HOX genes [58]. The remaining axis, left-right (L-R), is the one along which the symmetry pattern is manifested. On account of the high biodiversity that bilateria group includes, only the subgroup of vertebrates is examined in the following literature study. In addition, any reference to symmetry or asymmetry from now on corresponds to the L-R direction, unless explicitly mentioned.

This study makes an effort to statistically identify the genetic origins of a complex structural phenotype. Hence, examining, based on existing research, the main brain developmental stages is essential to discern the processes that induce bilateral symmetry. An important vertebrates (and bilateria) common characteristic is the germ line **triploblasticity**: the embryo begins as a flat disk, through a process called **gastrulation**, with three distinct cell layers; **endoderm**, **mesoderm**, and **ectoderm** [41]. Of significance in the neural system formation is the ectoderm, which is initially equivalent to one of the flat disk sides. Under the context of this study, although a fact not directly connected to the brain cortex, it is necessary to mention that the perfect bilateral symmetry pattern appears to break even before gastrulation. In Xenopus (frog species) embryos, during fertilization and the initial 4-cell cleavage of the fertilized egg , the **cytoskeleton microtubules** appear to

asymmetrically localize the ion channels proteins, whose RNA has been passed on by the mother, with a preference for the right side of the complex [10]. Chicks embryos also exhibit a similar pattern. The occurrence of asymmetry at this extremely early time point underlines the significant role it has on the embryo development, species fitness, and, concomitantly, the conservation potential of this trait drivers [9]. Another cellular component that is considered to enhance asymmetry, during gastrulation, is the motile cilia, hair-like organelles on the cell surface with the ability to beat [53]. Their movement is by construction asymmetric, causing a leftward flow of extraembryonic fluid and, subsequently, asymmetric distribution of exogenously introduced proteins [99]. Both studied phenomena point to early initiation of asymmetric genes expression.

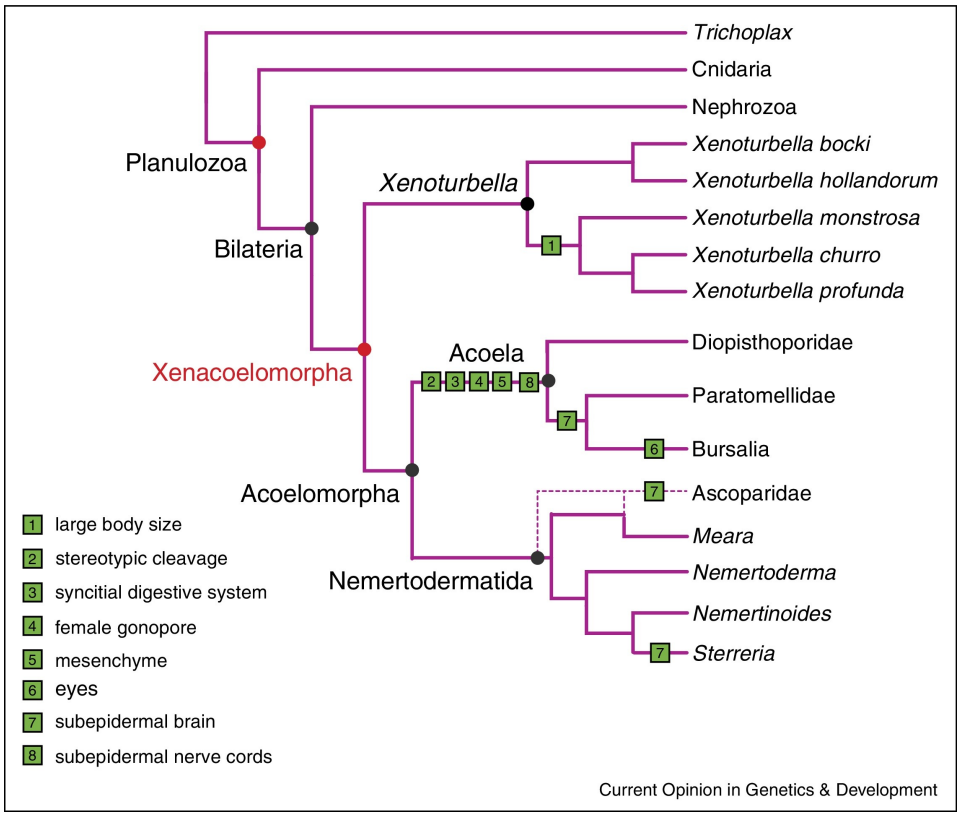


FIGURE 1.1: Species phylogenetic tree subset, displaying bilateria clade, its sister clade, Cnidaria, and the direct children[60]. Of great importance on the evolutionary studies of bilateral symmetry is the Xenacoelomorpha clade.

1.1.2 Symmetry during CNS formation

Shortly after gastrulation, the disk folds, in such a way that the central region of the ectoderm, called neural plate, forms a tube-like shape, the **neural tube**, which acts as the neural system precursor, under a process called **neurulation**. All bilateria have a central neural system (CNS), which entirely develops from the neural tube

walls [41]. The next pivotal step in the brain development, **differentiation**, leads to the creation of three distinct compartments along the R-C axis, at the neural tube rostral end, the **prosencephalon** (forebrain), which develops into the brain cerebrum, the **mesencephalon** (midbrain), and the **rhombencephalon** (hindbrain), that is later attached to the spinal cord in vertebrates. For the subsequent mechanisms and terminology to be compatible with human cerebrum related literature, the focus is shifted on mammals phylum and, spatially, on the prosencephalon. The differentiation proceeds, with two pairs of lumps extruding symmetrically from the prosencephalon, the **telencephalic** vesicles, the predecessors of cerebral region, and the optic vesicles, the precursors of optic nerves and retinas, while the central remaining, linking structure is called **diencephalon** [42]. The formed symmetry plane is called **midsagittal**. The telencephalic vesicles continue to grow, expanding also caudally and in parallel with the diencephalon, gradually assuming the form of the two hemispheres, while a new pair of vesicles appears on the rostral part of the diencephalon, giving rise to the **olfactory bulbs**. The neural tube shape also reacts to the changes, forming four distinct **ventricles** along the neural tube, with two of them, named **lateral ventricles**, being mirrored inside each of the telencephalic vesicles. The earliest stage where asymmetry is noted in an anatomic level inside the human brain is during the end of the first trimester of gestation [2]. Specifically, the choroid plexus, a specialized cell network that lies inside the ventricles, attached to the diencephalon, and produces most of the **cerebrospinal fluid** in the CNS, develops asymmetrically in each lateral ventricle. The cerebrospinal fluid is of great value for the developing brain, as the main source of nourishment, waste removal and protection [125]. Such an asymmetry manifestation in a macroscopic level, therefore, may be the progenitor of other forms of asymmetry at a later developmental stage [115], even at the brain surface. Cerebral bilateral symmetry therefore begins breaking down during fetal development, producing an asymmetric brain (Figure 1.2), and giving rise to partial functional disassociation, called **brain lateralization**. Lateralization becomes visible when examining organisms' behavior, with the most studied trait in humans being handedness and language [115, 27]. To better understand why and how the inner functions are related with the external brain cortex development, the underlying cellular processes of **neurogenesis** and **neuron migration**, active throughout differentiation, need to be identified, before introducing the reader to the anatomy of the fully grown brain. For this purpose, a further focus on primates phylum is needed, given the differences exhibited when comparing different mammals, such as rodents[89].

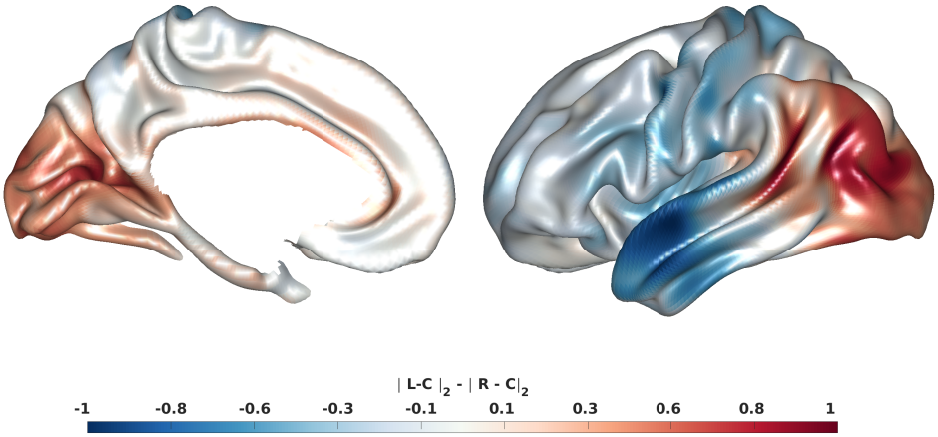


FIGURE 1.2: Illustration of human cerebrum brain asymmetry. Normalized differences of the distances of each hemisphere rescaled, rotated and averaged surface landmarks from the center of mass. See subsection 2.2.1 for more details on the preprocessing.

1.1.3 Neurogenesis, Neuronal Migration and Plasticity

The cells initially comprising the neural tube walls are named neuroepithelial cells (NPCs), and exhibit similar properties with stem cells, that is limited multipotency (i.e. they can differentiate into multiple cell types) and limited self-renewing (i.e. they can divide symmetrically into new NPCs a finite number of times), while also properties of epithelial cells, that is polarity (i.e. asymmetric cellular organization, with distinct basal and apical surfaces) and attachment (i.e. junctions tightly connect adjacent cells) [51]. This cells array is contained between the basal and apical laminae, lipid membranes lateral to each other, with the apical lamina facing the neural tube lumen [1], and the cells being radially distributed. During anatomical differentiation, around the 7th gestational week in humans [98], self-renewing is activated, leading to cells proliferation and CNS bilateral expansion, while attachment is hindered, gradually exchanging the NPCs with radial glial cells (RGCs), the fate-restricted progenitors of neurons, marking the initiation of **neurogenesis**[51]. A RGC acts as the main building block of the brain, from which a single neuron or a neural progenitor, that later divides symmetrically in neurons, is generated. RGCs' pivotal role does not end here. As it can be seen in Figure 1.3, RGCs are stretched during development, with processes connected to the surface of neural tube successor ventricles and to the outer cortical region surface, forming thread-like scaffolds. Newly formed neurons, generated from the RGCs main, oval body, which remains close to the ventricles, use this structure as a guide to move towards the outer region of the cortex, under a process named **neuronal migration**[103]. This type of movement actually implies that the newly formed neurons head towards the brain surface, building the brain in

an inside first, outside last fashion [89]. At later stages of human gestation, around week 19, studies have shown that a morphological transition happens, where the majority of RGCs stops being attached to the **pial surface**, the outer surface of the brain, limiting the migration ability of neurons [98] and affecting the way new layers are formed. Human neurogenesis extends to the third gestation trimester, being suppressed in case of premature birth [84]. Postnatal neurogenesis is therefore presumed to be quite limited for primates [39], despite the fact that the postnatal brain dramatically increases in size , with that attributed to a rapid increase in neurons connections and glial cells (i.e. cells that provide physical and metabolic support to neurons) number [38]. The environmental factors that may affect brain lateralization are mainly detected before or during birth, with epigenetics and birth complications appearing to be mostly correlated with handedness [115, 18]. However, the human brain exhibits high **plasticity**, namely the ability of intrinsic or extrinsic factors to change the neurons connectivity, setting aside the genetic predisposition, a property that has been proven to particularly affect the brain surface asymmetry in studies with monozygotic twins [140, 31]. In general, though, the more complex the phenomenon and the closer it is to humans, the higher the uncertainty and the greater the ethical implications. Only recently, non-invasive imaging and transcriptomic techniques have given out further details regarding the brain development sequence, with genetic studies indirectly identifying the landscape of the underlying genes that affect different brain regions formation and symmetry [18]. Moving on the literature study path and getting closer to the studied phenotype, the fully grown human brain cerebrum is subsequently anatomically described.

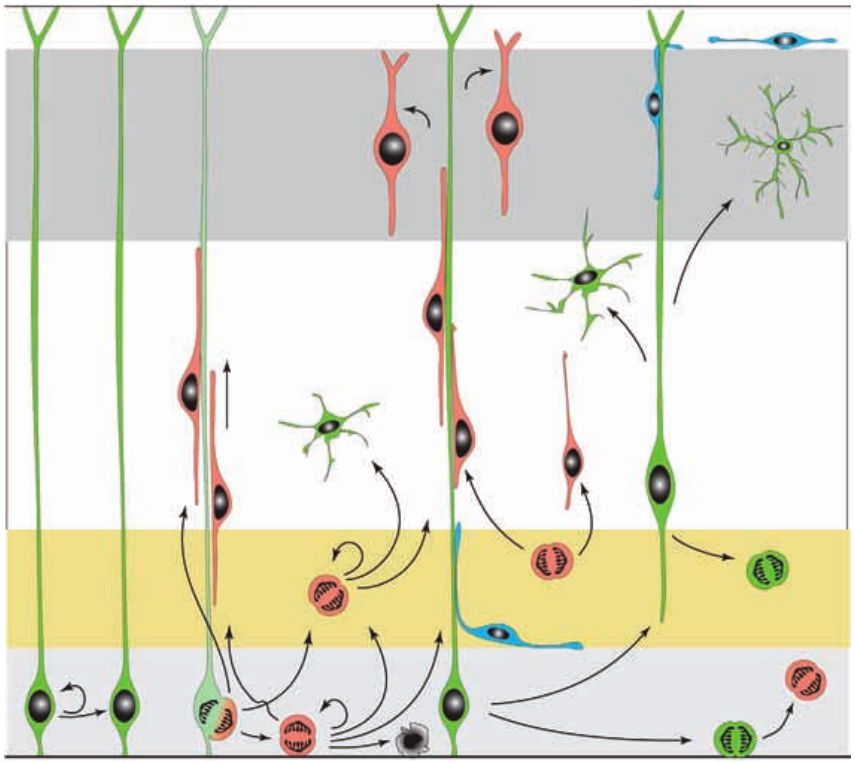
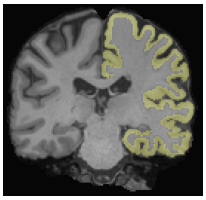


FIGURE 1.3: Illustration of a classical model of radial glial cells complex nonlinear division processes and neuronal migration [103]. From left to right: NPCs (green) originally divide symmetrically; During differentiation, NPCs become RGCs, which divide asymmetrically, generating neurons or neural progenitor cells (orange). Neural progenitor cells eventually divide symmetrically into neurons. The majority of neurons in humans is produced by neuronal progenitors. A part of the generated neurons migrate radially towards the cortical plate, by attaching on the RGCs projections; Eventually, after brain maturation, most RGCs in humans undergo apoptosis (i.e. cell death) or generate neurons-supporting cells, such as astrocytes.

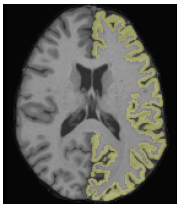
1.1.4 The adult human cerebrum anatomic and functional properties

Human cerebrum is the center of sensations and thinking. The following excerpt provides a summarized anatomic [40] and functional [44] perspective. As aforementioned, cerebrum is entirely produced from the telencephalon during fetal development, with the telencephalic vesicles ending up becoming the two hemispheres, that remain connected through what is known as the **Corpus callosum**. The side view of each hemisphere is named **lateral**, and the view of the inner side is called **medial**. The human cerebrum outer covering surface is called **cerebral cortex**, the region on which the current study focuses. The human cerebrum appears distinctly different from other organisms, mainly due to the **sulci** (i.e. grooves) and **gyri** (i.e. bumps), with them being the result of the tremendous expansion of the cerebral cortex surface

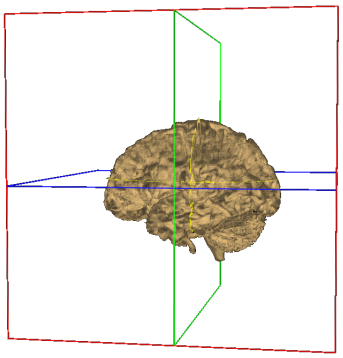
area during fetal development, folding and wrinkling in order to fit the skull. The precise pattern of gyri and sulci varies significantly across populations, rendering the brain surface unique per individual. Under a biopsy dissection or a magnetic resonance imaging (MRI) scan, the cerebrum appears to consist of two distinctly colored types of matter, implying changes in composition and consistency; the gray matter, at the outer part of the cerebrum, which contains the cell bodies, dendrites and the axon terminals, where all synapses are, and the white matter, at the inner part, made up of myelinated (i.e. biologically insulated) axons, which connect different parts of gray matter to each other (Figure 1.4). Protective layers on top of the gray matter, called **meninges**, ensure that the brain does not come in contact with the outer bone, with the one attached on and marking the outer borders of the gray matter named **pial surface**. In this study, the midthickness surface is examined, a term referring to the surface halfway between the pial and white matter surface.



(A) Coronal section



(B) Axial section



(C) Three-dimensional (3D) sections map. Green rectangle corresponds to the coronal section plane (A) and blue rectangle to the axial section plane (B).

FIGURE 1.4: Gray and white matter as seen from different sections, in an MRI screening, as retrieved from Freesurfer freeview routine. The gray matter is annotated with yellow color in the right hemisphere. Non brain regions have been removed.

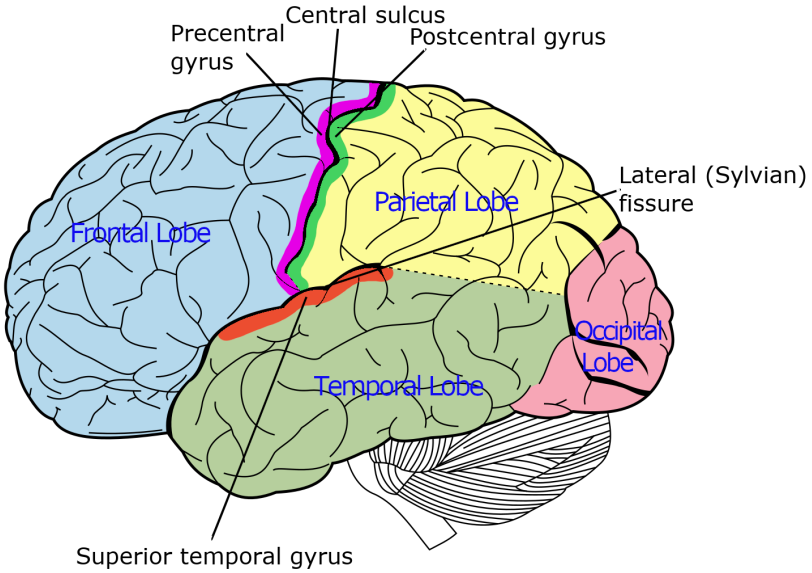
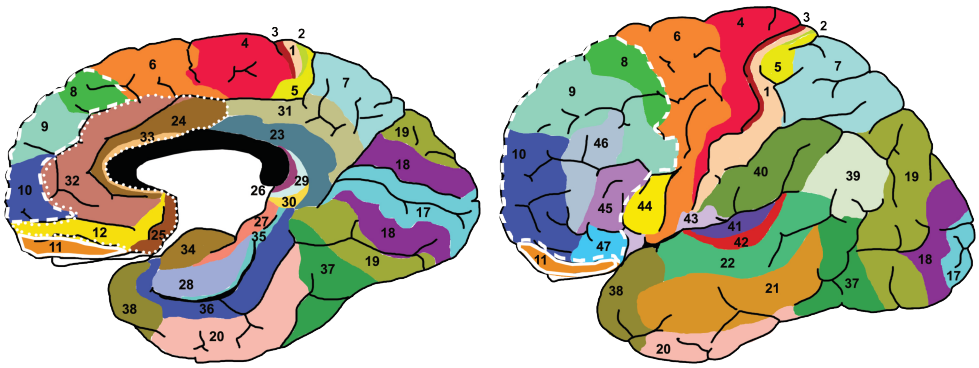


FIGURE 1.5: Cerebrum lobes (blue font) and main gyri, sulci and fissures approximate positions (black font).

Efforts of partitioning the brain have been numerous throughout the years of medicine, with diverse resolution and purpose. Crudely, the cerebrum hemisphere is divided into lobes, that are named, by convention, after the bones of the skull that lie over them (Figure 1.5). A more detailed approach is based on the identification of the functional processes that take place in each part of the cortex, with Korbinian Brodmann being the first person constructing a 52-partitions experimentally based approximation of the hemisphere [13] (Figure 1.6). Each partition is being represented by an identifier BA##. The main regions identified are:



(A) Medial surface (B) Lateral surface

FIGURE 1.6: Brodmann map of functional partitions.

- Sensory areas:

  - Somatosensory cortex (areas 1-3): the post-central gyrus (Figure 1.5). It is responsible for the body-wide sensory information processing, such as touch, temperature and pain.
  - Visual cortex (areas 17-19): occipital lobe surface. It constitutes the center of processing of visual information, as received from the optic nerve.
  - Auditory cortex (areas 41,42): rostral posterior part of the temporal lobe. It processes auditory information, identifying fundamental sound characteristics, such as frequency and loudness.
  - Gustatory cortex (area 43): An area behind the temporal lobe, responsible for taste signals processing.

- Motor areas, that are related to movement planning and manifestation:
  - Primary motor cortex (area 4): The precentral gyrus (Figure 1.5). It is the center of voluntary movements execution, generating the electrical signals required for the neural impulses to be transmitted to the body muscles.
  - Premotor cortex and supplementary motor area cortex (area 6): rostral part of the frontal lobe, anterior to the primary motor cortex. They are the center of motion planning and control, determining the sequence of movements required for a simple task to be performed.

- Association areas, which are related to perception, memory and thought processes:
  - Prefrontal cortex (areas 8-14,24,25,32,44-47): anterior part of the surface of the frontal lobe. It is centrally involved in cognitive control functions, spanning attention, salience detection, inhibitory control, working memory (i.e. short-term temporarily stored memory, related to a certain task), cognitive flexibility, empathy and pain processing [101]. Areas 44 and 45, referred to as **Broca's region**, are responsible for speech production. Human prefrontal cortex remains one of the least functionally demystified parts of the cortex , presenting difficulties in every level of study, as it exhibits a higher relative size, higher cellular type variety, more complicated neuronal migration and denser connectivity patterns than other animals.[20]
  - Inferior temporal cortex (areas 20,21): caudal part of the temporal lobe cortex. It is responsible for the aggregation of the processed visual information towards a meaningful interpretation, supporting object recognition.
  - Posterior parietal cortex (areas 5,7): posterior part of the parietal lobe surface. It processes sensory information produced from all six senses to construct a semantic representation of the person's surroundings, leading to motion planning and spatial reasoning.

– Cingulate gyrus (areas 23-24,28,33): an arch-like fold rostrally to corpus callosum. It is the conscious part of the **limbic system**, which is the center of emotions, instinct and reflex responses.

Recently, with the advance of imaging methods, maps have been manufactured, to automatically partition the MRI extracted three-dimensional (3D) cortical surface into 68 partitions, based on morphological characteristics. One such gyral-based atlas, Desikan-Killiany (DK), is derived from the changes in curvature under an expert-driven model of gyri locations [33] and provides automatic **cortical parcellation**, aligned to the Brodmann functional partitioning (Figure 1.7). This atlas is going to be used throughout the proceeding analysis for the quality control of applied segmentation techniques.

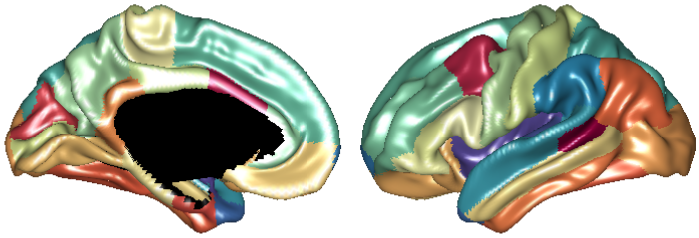


FIGURE 1.7: Desikan-Killiany atlas, mapped on the midthickness surface of the left hemisphere, with the medial (left) and the lateral (right) views displayed.[35] Different colors represent different partitions. The black region has not been mapped or is referring to sub-cortical parts.

1.1.5 Evolutionary studies

From an evolutionary perspective, it is extremely rare for the right conditions to occur, in order for any soft tissue specimen to be preserved, across a considerable amount of time. The only known way is through mineralization [102]. This fact renders a mammal's ancestor brain almost impossible to retrieve. Nevertheless, endocranial imprints have been used as a proxy to describe the relationships between hominids and their ancestors [11, 94]. The reason behind this phenotypic delegation is purely practical. The brain size and shape follow the container volume restrictions. Although such studies support the theory of propagating asymmetry among studied individuals, with the most evident signs in human skulls, little information about the surface shape can be retrieved, as only the convex hull shape of the brain can be delineated from such process. Through the association of brain asymmetry with DNA, a universal code among organisms, it becomes possible to deploy tools used by evolutionary geneticists, to identify the phylogenetic tree of this complex trait,

locating conserved regions among organisms and their predicted divergence in time, under a pleiotropic model [72].

1.1.6 Reported general human cortex symmetry traits

Although human cortex exhibits roughly symmetric structure, the symmetry is systematically suppressed, not only due to the environment, with plasticity playing a central role, but also because of genetic factors, as explained in the previous sections. An asymmetric pattern is manifested across adult individuals, irrelevantly of their upbringing, comprising, therefore, a characteristic of the human species, while general abnormalities in this pattern are related to the occurrence of mental disorders, such as autism or developmental language disorder [61, 75]. Some of the most prominent asymmetric traits across healthy individuals are the following:

- Yakovlevian torque (Figure 1.8): the right hemisphere prefrontal lobe and the left hemisphere occipital lobe tend to cross the midsagittal plane, extending towards the other hemisphere [76]. This creates a phenomenon of counter-clockwise warping, making the whole brain appear slightly leftwards rotated, while also making an impression on the inner part of the skull, called **petalia**. Increased left hemisphere occipital lobe extension, possibly caused by enlarged left lateral ventricle, is correlated with bipolar disorder[85]. Rising absence of the torque during aging is connected to schizophrenia and other mental disorders [107].

- Peri-Sylvian asymmetry: the left Sylvian (lateral) fissure is longer and sharper than the right one, while the right Sylvian fissure exhibits a more visible leftward curve, in the part where temporal lobe meets the parietal lobe, that is the auditory cortex, also called **planum temporale** [76]. The increased thickness of the right superior temporal lobe, that reduces the lateral fissure steepness, is attributed to increased white matter volume. Such trait has been reported to be gender-related, with males exhibiting greater asymmetry than females, as noted in previous studies, with steroid hormone receptor activity and steroid metabolic process related genes [55].

- Central sulcus asymmetry: the right hemisphere central sulcus is deeper and larger [76]. Larger asymmetry appears to be correlated with attention-deficit / hyperactivity-disorder (ADHD) [78].

- Motor areas asymmetry: the motor areas are generally larger on the left hemisphere.

Statistical modeling of the observed symmetry pattern can provide a hint on the significance of genetic and environmental factors contribution[69]. The current study focuses on the genetic component, which has been diversely investigated across literature.

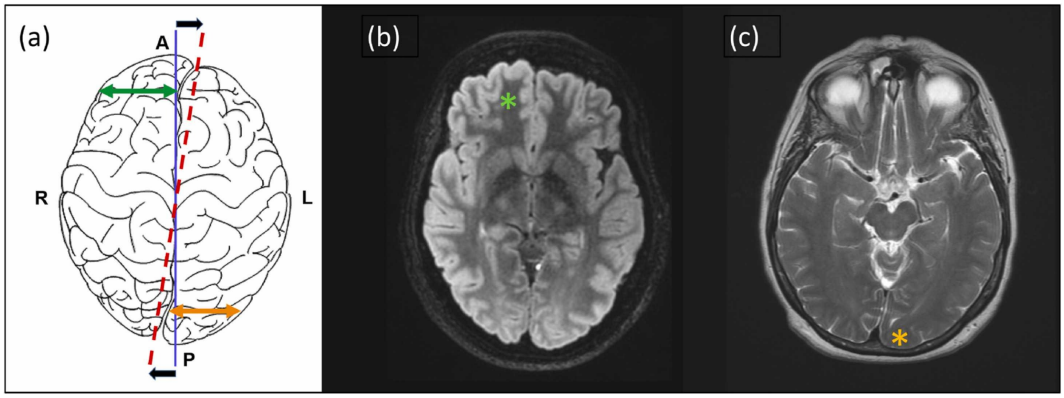


FIGURE 1.8: Yakovlevian torque schematically illustrated (a), along with its manifestation in different axial sections for a single individual (b,c) [76].

1.2 Genetics of multivariate quantitative traits

1.2.1 Multivariate genome-wide association study (mvGWAS)

Genome-wide association studies (GWAS) aim to relate genetic information, usually extracted from single nucleotide polymorphisms (SNPs) markers arrays, with a phenotypic trait. When the trait is dichotomous, measured by its presence or absence, then GWAS are applied on a case-control fashion, where two cohorts, an affected (case) and an unaffected (control), are compared. [127] In the present work, the focus is directed to quantitative traits, whose measurement takes continuous values.

1.2.2 Single nucleotide polymorphism (SNP) identity and importance

Single nucleotide polymorphisms (SNPs) are characterized by single nucleotide base-pairs positions where two or more different alleles (i.e. variants of nucleotide bases) are observed, with the second most frequent allele appearing with a frequency, called minor allele frequency (MAF), higher than 1%, a property that sets this term apart from the more general notion of a single-nucleotide variant. Being the most common type of polymorphism in the human genome, SNPs were popularized on account of their considerable effect on influencing transcription. Apart from the direct case, where a SNP belongs to an exon and the alternate allele corresponds to a non-synonymous (the translated amino-acid differs) or a nonsense (the codon stops translation) mutation [104], the majority of registered SNPs (88%) reside in non-coding regions. They can have an impact on the physio-chemical properties and conformation of docking positions for DNA-binding enzymes, such as transcription factors (TFs), causing binding affinity changes, influencing transcription regulation and, ultimately, altering biological pathways relevant to dependent phenotypic traits [97]. As a matter of fact, 31% of the known DNA elements, as reported by the

ENCODE project, the human genome encyclopedia, appear to be part of TFs binding domains [37].

1.2.3 Linkage disequilibrium (LD) effect on SNPs

The genetic information for each individual is only represented by a certain amount of single nucleotide polymorphisms (SNPs), called **tag** SNPs, based upon the principle of high linkage disequilibrium (LD) [133, 12]. LD, the non random association between alleles at different loci, is attributed to mutations, genetic drift and, concomitantly, selection [133] that has rendered certain combinations of alleles, named **haplotypes**, more beneficial for the survival and reproduction of a population than others, increasing therefore its **fitness**. These combinations are also more likely to occur topologically close, with recombination events (i.e. events that cause DNA strands to break and recombine, altering the haplotype) being less frequent the smaller the genetic distance [133]. Tag SNPs reduce the amount of information required to process pheno-to-geno associations, however the larger the effective population size(the part of the population that reproduces with viable offspring), the weaker the LD phenomenon [133].

1.2.4 Genetic association modeling

Phenotypic differences among individuals, described by the trait variance V_p, are the result of genetic variation V_g, known as **heritability**, environmentally induced variation V_e and developmental noise V_d (the deviations observed when environment and genetics are controlled), formally denoted as $V_p = V_g + V_e + V_d$ [134]. The genetic information content V_g is *approximated* by the amount of variation in tag SNPs that is translated to variation in the studied trait properties. The presence of a minor allele signifies divergence from the general population characteristics, and hence implies that information is contained in that SNP. The relationship of each SNP with the phenotypic trait is statistically represented by a certain genetic model. Under the assumption of an **additive model**, if a certain minor allele occurs in both DNA strands, i.e. the allele is homozygous at that locus, then its effect is double compared to the heterozygous case, independently of which strand is carrying it. This hypothesis requires no prior knowledge and makes no further assumptions regarding the alleles dynamics, that is the degree of dominance of each allelic variant. The described model, for a single quantitative trait (dependent variable) and a SNP with a single minor allele (assumed independent variable), assessed on a sample with size N, can be formulated using a univariate linear regression $y = \mu + \beta x + \epsilon$, with x the allele's occurrences number, y the phenotypic trait, β the SNP effect and ϵ the part of non controllable factors, referring to environment and developmental noise.

1.2.5 Contradicting no association

A SNP is considered to be significant, if its effect contradicts the null hypothesis H_0 of no association ($\beta = 0$). For the subsequent analysis, under a biological setting,

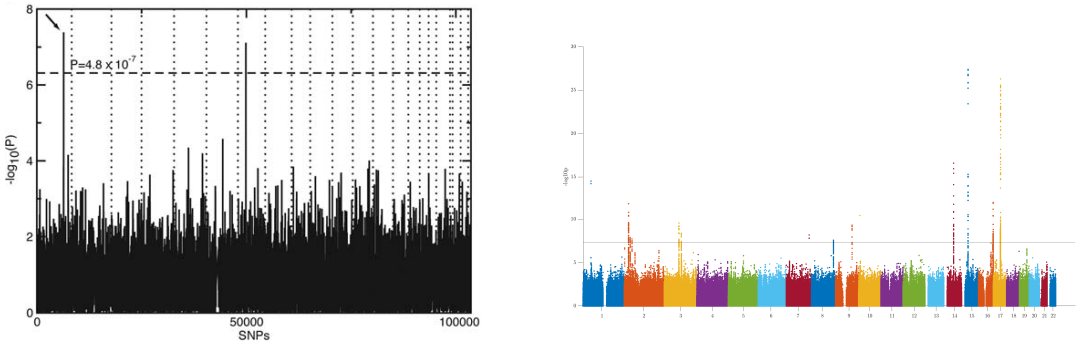
it is assumed that the phenotypic trait follows a normal distribution. Under the reduced model of H_0, the residual sum of squares (RSS) equals $RSS_R := \sum (y - \bar{y})^2$, with $\bar{y}$ the observed mean value of the phenotype, and the degrees of freedom (DOFs) being equal to $N - 1$. Under the full model of alternative hypothesis H_a, the RSS equals $RSS_F := \sum (y - \hat{y})^2$ with $\hat{y}$ the estimated trait, with $N - 2$ DOFs. In an analysis of variance (ANOVA) setting, the F-statistic $\frac{MSR}{MSE} := \frac{RSS_F - RSS_R}{RSS_R/(N-2)}$ is defined, as MSR and MSE follow a χ^2 distribution with 1 DOF and N-2 DOFs respectively under H_0, which can be used to contradict the null hypothesis.

An alternative to the aforementioned hypothesis test can be made by considering an H_0 that the coefficient β, scaled by the standard error (i.e. standard deviation), follows a standard normal distribution. The hypothesis then can be contradicted by comparing that quantity with the corresponding z-score. An advantage of the latter approach is that the distribution compared is two-sided, meaning that the effect can be given a positive or a negative sign. However, no consideration for the uncontrolled factors is made, a fact that could potentially influence the computed score and render it less interpretable, compared to the ANOVA case.

1.2.6 Measuring SNP significance

By assigning a minimal probability to the event that no association is observed, namely defining a p-value cutoff threshold, a SNP is found to be significant if the p-value from the corresponding F-test is less than the explicitly defined cutoff [3]. Greater sample size means lower MSE, larger F-statistic and, consequently, lower p-value. Thus, greater sample size increases the chance of discovering significant SNPs and low sample size raises the type I error of the detection, namely the presence of false negatives that actually confirm H_a.

The probability threshold is derived based on an empirical value, fixed to 0.05, corrected using the Bonferroni correction for multiple independent tests, hence $\frac{0.05}{N_t}$ with N_t the number of SNPs. The method is rather conservative, therefore usually cutoffs are computed by replacing the number of tests with the amount of independent common SNPs for a given population. Based on the findings of the International Hapmap Project, this amounts approximately between 200,000 to 1 million tag SNPs [12]. Therefore, the suggestive cutoff threshold 5×10^{-8} is used. The p-values are most commonly converted to values proportional to significance, by applying the $-\log_{10} p$ transformation. The LD phenomenon, being often locally observable, causes seemingly continuous p-value spikes to appear when plotting the data points, with the lead, or functional, SNP, that is the one with the greatest local significance (i.e. the lowest locally recorded p-value), being 'supported' by lesser significant SNP in its vicinity. The resulting scatter plot, with SNPs $-\log_{10} p$ values placed by bp position on x axis, resembles the Manhattan city landscape (1.9b).



(A) (B)

FIGURE 1.9: Examples of GWAS. In (A), the first recorded GWAS Manhattan plot [68] is displayed, where the SNP in chromosome 1 (black arrow) with the most significant effect, related to Complement H Factor Polymorphism, was identified to affect age-related macular degeneration disease propensity, one of the major causes of blindness among elderly. In (B), a GWAS scatter plot from the present work is shown, where the SNPs spikes signal, attributed to LD, is evident.

1.2.7 Major advantages and disadvantages of univariate GWAS

There are several benefits and limitations of such a study design. GWAS have been successful in revealing novel relationships between genes with known properties and a variety of observed phenotypic traits and clinical applications, presenting evidence of possible biological mechanisms related to genes with unknown function [123]. They also empower population-specific comparative studies, at the level of how ethnicity or other kinds of population stratification affect a certain trait, while accommodating the possibility to investigate the effect of an allele no matter how frequently it might appear in the studied sample [123]. On the other hand, it has been generally observed that each SNP can only explain a small part of the heritability of a certain trait, with a large amount of the signal hidden in gene-to-gene interactions, that are not captured in this method, and possibly in SNPs whose contribution has not been considered significant enough [123]. To remedy the latter issue, larger sample size is ideally required. Due to the fact that the largest amount of SNPs is located in intronic regions, it has also been difficult to assess the causality of those variants, which genes they affect and how they do it. Additionally, too many, possibly unrelated, hits may be involved in a GWAS result, due to LD. [123] Recent studies have deployed advanced post-GWAS statistical and machine learning (ML) approaches to account for identifying the causality and the functional behavior of significant SNP [96, 49].

1.2.8 Multivariate genome-wide association studies (mvGWAS)

When the analyzed phenotype is described by more than one measurements, as is the case in this work, where it is expressed as the combination of different 3D landmarks composing the cortex of each individual, GWAS require a methodological change to accommodate this fact, as univariate regression cannot be used as is. Also, a single

genetic locus may exhibit more than one minor allele. The goal, consequently, is to incorporate multi-allelic SNPs and, more importantly, multivariate phenotype, in a single hypothesis test per SNP, leading to what is called multivariate genome-wide association study (mvGWAS). In general, there is an abundance of strategies on how to perform a mvGWAS, ranging from direct methods, that approximate the inputs relation either in an unbiased manner or making certain educated guesses, to more complex techniques, that increase statistical power by transforming the inputs, at the expense of explanatory ability [48]. There are also methods that are based on the meta-analysis of outcomes from univariate studies, commonly used to juxtapose experiments from separate sources, for which the original data is missing, the experimental setup, mainly sample size, across studies varies, marking the studies data 'incompatible', or a single study is computationally intractable [127, 21]. These approaches combine the test-statistics produced from the individual studies and produce an estimate of the multiple trait test-statistic. Which approach performs best mainly lies on the dataset properties and the nature of the scientific question. Factors such as low sample size [117], genes pleiotropic effects [43] (i.e. a gene affects multiple phenotypically independent biological pathways) or within-study variability [128, 63] tend to handicap the statistical modeling and increase the type I and II errors of the corresponding hypothesis tests. In this study, canonical correlation analysis (CCA) was chosen due to the high capacity in efficiently reducing the inputs dimensionality while preserving most information regarding their correlation, and the same sample size across analyzed traits, an approach that has provided successful results in similar efforts of brain shape and face shape analysis [23, 93]. An additional strength of this method is that a single test is performed per SNP, not requiring further multiple test correction over the number of phenotypic traits, thus having increased statistical power than meta-analysis techniques.

1.2.9 Generalizing to genes - Functional association

The next natural step to perform, once significant variants have been identified for a specific phenotype, is to investigate how such an association is realized, ultimately supporting and extending the assembly of the complex relationships graph between genetics and actual observations. No matter if there is effect on regulatory elements or on the gene product itself, this kind of venture is largely obfuscated, given the little amount of knowledge that exists to fill the relational path, which may include a great amount of steps and interactions. The majority of SNPs reside in intronic regions [12]. A great number of them is also likely to exhibit cis-acting effects. Surprising is also the event that SNPs mapped on an exon of a known gene can be actually manifesting significant correlation with a trait, such as obesity [24], through the interaction with a different gene. Resolution of GWAS in discovering causal genes is also being limited by the number of studied individuals, the genotyping arrays technologies and the existence of LD [36], with the lead SNPs not necessarily being the functional source of the association signal. Even by ignoring the ambiguous relationships and by knowing the start and end of an association path, the probabilities are thin that the exact trajectory, implicating selective messenger RNA translation throughout

development, can be derived.

Nevertheless, active research is being performed, to approximate and model the underlying dynamics. Highly trait-specific and exhaustive wet-lab ablation studies, that include genome editing, can profoundly reduce the functional analysis complexity, particularly when little evolutionary and ethical barriers exist, and the analyzed trait has low dimensionality [105]. In the general case, though, modeling and approximations need to be performed, in order to deal with the task. Population stratification and ethnic group variation has shown little interference in common variants analyses [133], at least when considering GWAS on various diseases, a fact that points to highly frequent and widely distributed causal alleles and may support a form of simplification for the described, seemingly physically and computationally intractable task, possibly allowing for the unification of results from diverse ethnicity experiments. In addition, a gradual incorporation of multiple knowledge sources is taking place to support this undertake; findings from RNA expression profiling (such as RNAseq and scRNAseq), also known as eQTL analysis, [142, 145] or epigenetic regulation studies (such as ATACseq and CHIPseq) [28, 62], that could potentially provide snapshots of genes expression during development, are cross-tested against GWAS. Through identification of cell types by genetic expression profiles clustering and association between tissues and identified clusters, localized analyses can be more accurate in extracting SNP-to-gene relationships [17].

The key approximation required to combine information is to map the underlying data onto the same space. The most elementary jointly studied structure in these relational analyses is genes, due to the recurring need for actual, known, gene product concentrations to be detected. For the generalization from SNP to gene, the intermediate detection of lead SNPs needs to take place, as aforementioned, under a process called **fine-mapping** [114], which is tackled in various ways, from application of basic heuristics to regression and bayesian modeling, by incorporating priors fitted on the data. It is frequent practice to assume that lead SNPs reside in the captured and imputed data, therefore overlooking the event of no registration, when investigating rather common phenotypic traits. In such a case, a simple metric, utilized in the current study, to detect lead SNPs is to use the genetic relationship matrix (GRM), that is the symmetric matrix produced by computing r^2 for each SNP pair, to identify independent significant variants, using an arbitrary cutoff, over an explicitly defined genetic distance [139]. Once the lead SNPs have been pinpointed, they are matched to existing maps of annotated genes, after also considering surrounding non coding regions, such as the transcription start sites, as well as other information, retrieved from curated databases [137, 87]. A great part of the generalization is susceptible to underlying biases, with each gene having distinctly different size and possibly overlapping with others [138].

Functional association is then made possible under the framework of gene set enrichment analysis (GSEA) [121]. Being a well-established and extensively used method by a variety of different bioinformatics tools [19, 139, 79], with its first application dating back to 2003 [90], GSEA is realized by considering gene expression or epigenetic profiles, namely sets of significantly up and down-regulated genes, from different cell types or tissues, and statistically comparing the relation degree,

i.e. enrichment, with the genes identified in GWAS. This process can also be applied ubiquitously, from deriving connections with literature-specific gene annotated knowledge, such as gene ontology (GO) terms [7] and publications [122], to back-projecting the GWAS results to identify shared regulatory domains [64] or enriched TFs motifs [59]. Main disadvantage of GSEAs, other than the great number of assumptions required to retrieve GWAS gene sets, is the fact that the compared genes are lacking any prioritization or score, so their contribution is assumed uniform [138].

1.3 Related studies and contribution

The current work is meant to complement existing literature on the derivation of a statistical and genetic basis of cortical surface asymmetry on reportedly healthy individuals [116, 73, 74, 144]. However, the deployed methods are largely correlated with what has been done by Naqvi et al. [93] and Claes et al. [23]. Initially, a presentation of other association studies is given, and, afterwards, the contributions of this work are summarized.

1.3.1 The current landscape of cortical asymmetry GWAS

In the work of Sha et al. [116], DK atlas parcellation was used to extract widely-used asymmetry phenotypic features, called asymmetry indices (AIs) [52, 111, 73, 113, 146], that are calculated as the scaled difference of the contra-lateral segments areas, or volumes, depending on the nature of the research. The described quantity, for surface studies, is mathematically formulated as:

$$AI_i = \frac{S_{li} - S_{ri}}{S_{li} + S_{ri}}$$

with S_{li} and S_{ri} being the areas of the left and right hemispheric i-th segment of the atlas, respectively. Although the application of this method has brought forward plausible and promising results, regarding, for example, blood flow discrepancies [52], sex preference studies [113], Alzheimer's disease effect evaluation [111], or depression phenotypic traits detection [146], it has also met skepticism, relatively to the implied assumptions statistical validity [25]. Sha et al. [116] considered only the regions with significant SNP heritability ($FDR > 0.05$), as calculated using restricted maximum likelihood estimation [143], while at the same time controlling for age, ethnicity, data collection and other covariates. Subsequently, they performed a multivariate GWAS using a meta-analysis tool, called MetaPhat [80], on the filtered areas. MetaPhat joins the univariate GWAS per AI results, through the application of CCA, extracting the first canonical correlation coefficient r_i, along with a p-value of the hypothesis of all CCA coefficients being 0. The steps of snp-to-gene and functional annotation were performed using FUMA toolbox [139], that makes use of a large variety of databases and other tools, including the SNP annotation tool ANNOVAR [137] and the gene-set enrichment tool, MAGMA [30]. The lead SNPs filtering and functional characterization was specifically done with brain-related

provided eQTL and chromatin datasets [116]. Additional temporal developmental analysis was made feasible, once again through FUMA, and the authors identified high genetic enrichment during the early weeks of brain development, albeit with limitations in the dataset size and time resolution [67]. From the described pipeline, Sha et al. [116] were able to identify 21 genetic loci, implicated in microtubule organization and prenatal brain development. Furthermore, SNP-level significant similarities were identified with regards to schizophrenia, autism and educational attainment[116]. Because of the fact that multivariate GWAS were applied, the authors argue that no LD score analysis, discussed in subsection 2.4.5, could be performed. However, based on the work of Naqvi et al. [93], published the same year, a connection with multivariate studies can be, as aforementioned, successfully established. The dataset used to perform GWAS was the self-proclaimed white European cohort of UK Biobank, amounting to 32,256 individuals [81].

In the work of Kong et al. [73], the performed phenotypic measurements remained similar, enhanced with cortical thickness asymmetry statistics. However, the target goal did not implicate genetic factors at all and was conversely related to the first part of the current study, that is the statistical understanding of asymmetry across partition segments. The authors identified high AI dependency on age, sex and intercranial volume, raising awareness about the need of such covariates control. They managed to validate, through their analysis, global patterns of asymmetry in the inferior frontal gyrus, transverse temporal gyrus, parahippocampal gyrus, and entorhinal cortex [73], sub-regions of the inferior temporal and prefrontal cortices. Kong et al. [74] extended their research in performing GWAS on a global proxy of asymmetry, the Yakovlevian torque effect as measured using the skewness coefficients extracted from the 3D affine transformation matrix required to align a real asymmetric brain shape onto a template symmetric cortex, by merging a variety of different datasets, including UK Biobank. The authors were able to identify significant associations, without considering underlying genetics, between handedness and horizontal or vertical brain skewness, strengthening the position that functional lateralization follows structural asymmetry. Skewness also showed significant correlations with cognitive ability, behavior, language skills and mental health. The place and the country of birth appeared highly correlated with the observed asymmetry, without them being controlled prior to the correlation analysis. BMI was also found to be considerably associated. On a genetic basis, Kong et al. [74] identified high polygenicity, low heritability and genetic overlaps with autism spectrum disorder (ASD), though without detecting any significant genetic correlation with skewness.

By far the most complete reported study on cortical asymmetry was performed by Zhao et al. [144]. They focused on local geometrical features and the alteration of brain torque during development, by considering contiguous brain slices along with skewness coefficients and average landmark asymmetry differences on regions defined by automatic parcellation Destrieux atlas [34], collecting 348 features per individual. To handle the temporal factor, specialized datasets for studying brain development during adolescence, ABCD [135], PING [66] and PNC [112], came into play [144]. By using meta-analysis applied on univariate GWAS, they were able to identify only

2 significant lead SNPs, in chromosomes 1 and 10, from the originally discovered 86 lead SNPs, after adjusting p-values for the number of traits under a strict Bonferroni threshold. Without considering this adjustment, they identified significant genetic correlations; under moderately high positive r_g (>0.2), with Alzheimer's disease (AD), ADHD, and under negative r_g (<-0.2) with bipolar disorder, educational attainment, intelligence and schizophrenia [144].

1.3.2 Present work contributions

The current work aims to provide a data-driven approach of studying cortical asymmetry, under a coarse-to-fine segmentation strategy. The scientific goal is to fundamentally identify the underlying factors that give rise to this phenotypic trait throughout the different regions of the cortex, both in the general, statistical context, as well as in the more specific genetic landscape. To this end, two stages of analysis take place.

The first stage consists of a non-parametric statistical analysis, that demarcates the degree different effects have on asymmetry, using a 2-way permutation ANOVA, on landmark-defined brain regions. It closely follows the analysis performed by Claes et al. [22] on the facial asymmetry, as firstly introduced by Klingenberg and Mcintyre [70]. The outcomes pinpoint the degree under which different regions in the average brain are asymmetric, referring to the average genetic effect, and the degree under which being a different individual affects the shape of the average hemisphere, which is related to the individuals' upbringing, as well as their specific genetic background. Lastly, the interaction between the effect of sides and individuals is assessed, to offer an understanding of whether being a different individual affects the underlying asymmetry. To our knowledge, no similar analysis has been applied in the past, and an anatomic analysis of the outcomes is being provided by Vanbiervliet et al. [131].

The second stage dissects the genetic background of cortical asymmetry, by identifying the variants on healthy individuals that are most likely to affect different brain regions, under the regimen of a 4-level partitioning, performed in a bifurcating manner. The phenotype analyzed is the principal components (PCs) of the difference of coordinates of contra-lateral landmarks of the normalized hemispherical shapes. Hence, no prior assumptions on anatomical regions, global morphology and geometry are made, in contrast to relevant studies [75, 144]. For each identified partition, the following analysis is performed. CCA linearly relates the multivariate phenotypic features contribution with covariates adjusted multi-allelic SNPs, making the minimal assumptions that the distributions of the studied variables are multivariate normal and that the underlying phenotypic differences are linearly correlated. The GWAS design power of generalization is evaluated by considering two different datasets and comparing, qualitatively and quantitatively, the extracted results. The results from these analyses are subsequently combined, to increase the statistical power and to identify originally underrepresented associations. The heritability of the analyzed phenotypic traits is consecutively measured, following the work of Bulik-Sullivan et al. [15]. Next, functional, developmental, tissue and cell-type specific analysis is applied on each partition, to identify the genetic correlation of cortical asymmetry with other

traits, diseases, biological pathways and age, in a similar format as Sha et al. [116], enhanced with the methods suggested by McLean et al. [87] and Finucane et al. [47]. Last but not least, the genes that are found to correspond to identified lead SNPs have their functional relations distinguished and clustered, their conservation among hominids closely related to humans are assessed, and significantly enriched publications are retrieved, using the work of Szklarczyk et al. [122]. The results of this study are compared to the ones in the literature. No such detailed, and at the same time generic analysis is known to have been performed on the cortical asymmetry phenotypic trait, like it has successfully been applied on other traits, such as brain shape [93] and face [23].

Chapter 2

Materials and Methods

In Figure 2.1 a brief overview of the processes applied in this work is displayed. In the coming sections, each compartment will be separately analyzed.

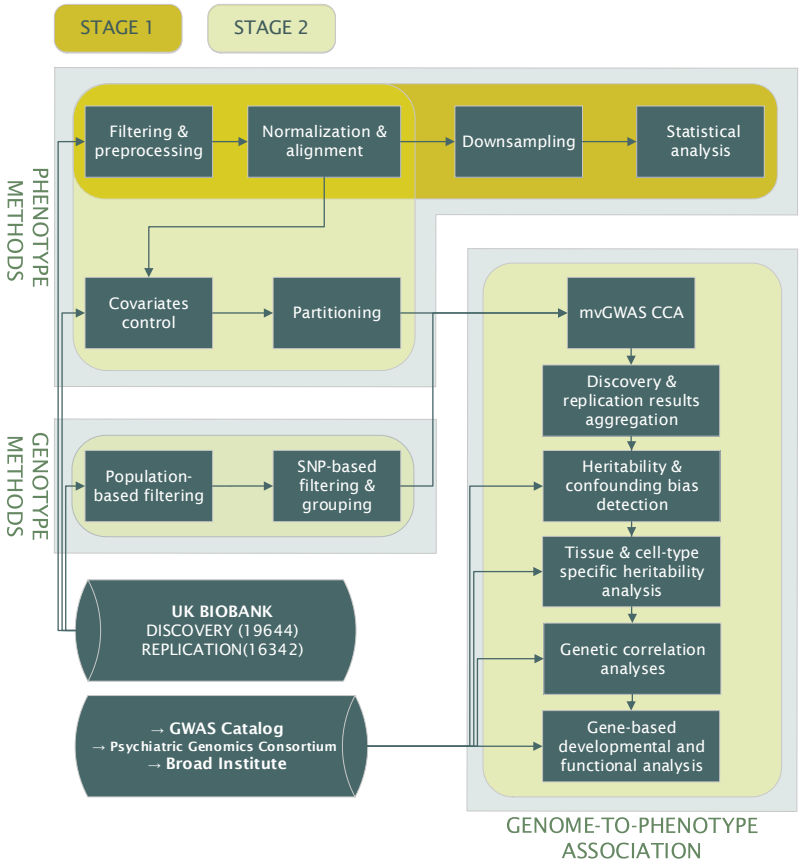


FIGURE 2.1: Visual overview of the applied methods and used materials. Stage 1 corresponds to the statistical analysis of brain shape asymmetry. Overlapping stage 2 refers to the steps performed in the genetic and functional studies.

2.1 Data description

2.1.1 Primary Data Source

With the advent of technology capable to collect and process genomes from different individuals in relatively high speed, vast databases targeting human physiology have been constructed. One of the main players in the data collection has been UK Biobank; a large-scale database from a randomized consortium of 500,000 individuals, whose genome has been collected, from whom 48,000 subjects had also participated in brain MRI collection process, as of December 2020 [81]. Participants are male and female, with the age range spanning 40 to 69. Apart from MRI scans and genetic information, of interest are general individual biomarkers that were collected, such as age, height, weight, BMI and blood pressure, as well as information about the imaging process, like date of acquisition, seat height, head (x,y,z) coordinates inside the scanner, and diagnostic center. Those are considered as covariates in the genetic analysis and their effect is assessed and removed. The present study focuses on healthy self-reported white individuals of European ancestry, filtering and preprocessing them based on the work of Naqvi et al. [93]. Specifically, the discovery dataset, amounting to 19,644 individuals and 9,705,931 SNPs, of the present work is identical to the one used in their study. In addition to that dataset, a smaller one, coming from a different, collected at a later time but under the same protocol, batch of 16,342 individuals is used as a replication dataset during GWAS.

2.1.2 Other sources

A 20 samples test-retest MRI dataset from Human Connectome Project (HCP) was collected to simulate replications during symmetry analysis. Various other sources are directly used during the meta-analysis, mainly to collect external GWAS scores. Those were selected from GWAS Catalog [16] or the Psychiatric Genomics Consortium and are summarized in the following tables:

	# Cases	# Controls	Ancestry
ADHD[32]	20,183	35,191	Undefined
AD[65]	24,087 late-onset 47,793 with family history	383,378	European
ASD[54]	18,381	27,969	Danish
BD[92]	41,917	371,459	European
Handedness[29]	31,856	299,181	British
MDD[57]	7,264	49,373	European
OCD [5]	2,688	7,037	European
Red Hair [91]	15,731	328,153	European
Schizophrenia [108]	36,989	113,075	European & Asian

Table 2.1: Qualitative traits GWAS sources.

	# Individuals	Ancestry
Cortical surface asymmetry[116]	32,256	European
Cortical surface shape[93]	19,644	European
Educational attainment[100]	405,072	European
Intelligence[119]	78,308	European
LFC[88]	32,186	European
Neuroticism[83]	329,821	European

Table 2.2: Quantitative traits GWAS sources.

Maybe something surprising presented in Table 2.1 is the red hair trait, which was assessed with data from the UK Biobank [91]. The reason of its inclusion is that, as the current work's approach is data-driven, significant signal is observed in the gene-based analysis, regarding this trait, raising questions about subpopulation stratification and leading to a deeper investigation of this association. As a global reference genome for identifying the LD structure and comparing GWAS from different sources, 1000G (Phase 3) data is used [8]. For epigenetic studies, the chromatin data from Roadmap Epigenomics [109] and ENCODE [37] projects are utilized, proposed and preprocessed by Finucane et al. [47] and offered by Broad Institute. Lastly, the out-of-the-box tools FUMA [139] and STRING[122] use their own abundant sets of resources.

2.2 Methods applied on Phenotype

2.2.1 Initial filtering and preprocessing

T1-weighted MRI scans are analyzed. The analysis is performed by initially converting the raw DICOM MRI volumetric images to well-defined 3D surface triangular meshes through the pipeline applied by FreeSurfer 'recon-all' [106] and Ciftify 'ciftify-recon-all' commands [35], on a space of 32 thousand vertices, with the average edge length being 2mm. Subsequently, the mid-cortical surface is arbitrarily selected, purportedly because of its plausible smoothness, allowing the distinction of sulci and gyri without over-representing their geometry [93]. After quality control, the vertices from the sub-cortical part of the surface, referring to corpus callosum connection points, are removed based on a mask derived from the Conte69 atlas [50], getting reduced to 29,759.

2.2.2 MRI Shapes normalization and alignment

The current work applies principles from general symmetry studies to model cortical asymmetry. For any of these analyses to occur, the aberrations of 3D shapes produced from MRI scans need to be considered. MRI output is affected by the subject positioning and technical error [141]. Volumetric differences also increase the level of discrepancies among MRI samples. To prevent positioning and volume deviations from gravely affecting shape comparisons, a normalization is required[69].

The samples of the derived 3D triangular mesh are represented as a set of vertices $\mathcal{V}_S$ of predefined dimensionality P, with a single landmark coded in the format of (x,y,z) coordinates. Those are joined together with a predefined faces matrix $\mathcal{E}_S$, with each of its elements containing three indices referring to $\mathcal{V}_S$, with the additional constraint that S is a multiple-connected structure, namely a graph in which there is at least one path joining any two vertices. Shapes normalization is performed through the application of generalized Procrustes analysis (GPA). GPA is an algorithm that iteratively performs translation, scaling and rotation on a given set of structures S, given initially a reference S_0, aiming to minimize the euclidean distance of corresponding points and the average shape. The translation is performed in such a way that the centroid C, defined by $\frac{\sum_{\forall i} \mathcal{E}_{Si}}{P}$, becomes the system origin. The scaling is such that the centroid size of the normalized S structure, defined by $\sqrt{\sum_{\forall i} ||\mathcal{E}_{Si} - C||_2}$ becomes equal to 1. The transformed samples then belong to what it has been coined as Kendall Space [69]. Under the framework of cortical surface analysis, a single hemisphere is considered to be one of the S structures. To apply any symmetry analysis, therefore, one of the individual hemispheres needs to be mirrored on the other side of the midsagittal plane, and then GPA is applied to align all hemispheres at once. The mirroring is performed by normalization of right and left hemispheres sets separately, and, then, x coordinate sign inversion of the right hemisphere landmarks. Aligning, finally, the entire dataset marks the end of the shapes normalization for the two tasks, statistical asymmetry analysis and GWAS, resulting into left (H_L) and mirrored right (H_R) shapes. In the case of GWAS, the difference between the left and right landmarks of the aligned shapes $D_A = \mathcal{E}_{H_L} - \mathcal{E}_{H_R}$ is computed, a proxy of directional asymmetry (DA) defined in subsection 2.2.4, before being reshaped to merge the dimensions of L landmarks and coordinates, resulting an array with size $N \times 3L$. This structure plays the role of the asymmetry phenotype in further analysis.

2.2.3 Downsampling

An intermediate step is followed when performing cortical symmetry statistical analysis, in order to reduce the computational burden of the process. A rather simple algorithm, in MATLAB context, has been derived, that computes the subset of indices of a given shape S, *approximately* with a given factor, that best describe the downsampled shape M provided from the proprietary function 'reducepatch' output [82]. With this method, the analyzed shapes are downsampled by a factor of 10, with the average shape retaining most morphological characteristics. The key idea is to find the one-to-many correspondence between faces from the two meshes. Let $\mathbf{Cn_T}$ be the centroids of each face of a shape T. The faces correspondence is found by firstly identifying for each face $x \in \mathcal{E}_M$ a part S_x of S, with $\mathcal{E}_{Sx} \subset \mathcal{E}_S$, joined to the closest, to x, face of S with index $y_{min}(x) = \arg\min_{\mathbf{Cn_S}} ||\mathbf{Cn_S} - \mathbf{Cn_M}||_2$, through a path of utmost 10 edges, that is the desired reduction rate. The faces of S_x are having non-zero entries in the 10th power of the S's adjacency matrix, at the y_{min}-th row. Then, the optimal vertex correspondence is found by taking all the vertices

corresponding to the faces subset $\mathbf{V_{Sx}}$ and identifying which of them is the closest to each of the vertices of x. The resulting downsampled shape R then has the faces of M, but projected on the vertices of S. This mapping allows for instant, although naive and of reduced quality, downsampling of 29,759 to 3,098 landmarks (Figure 2.2).

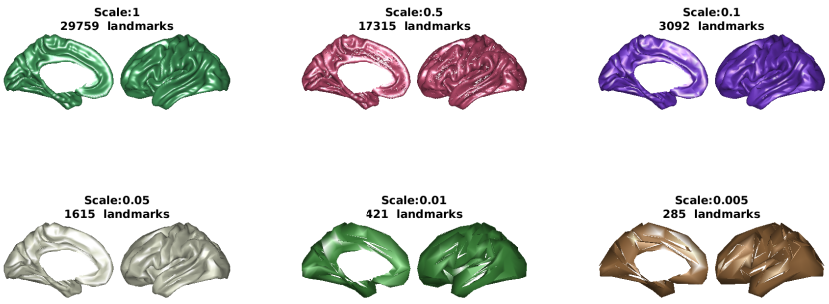


FIGURE 2.2: Downsampling indices of original average template using the novel algorithm and specific reduction scales (1/r). As no inter-face edge connectivity criterion is being considered, artifacts occur in the approximated shape, in the form of scars.

2.2.4 Symmetry Statistical analysis

Bilateral asymmetry is mainly described using three components in literature [71][132]. Directional asymmetry (DA), the main focus of this study, corresponds to the hemispheric side effect, namely how the intrinsic (i.e. genetic) properties of the studied population are manifesting across individuals. Antisymmetry, which is related to the effect where sidedness is random in a population (i.e. left-right pattern is mirrored to a right-left pattern), is not observed in the human cerebral cortex, in contrast to other internal organs positions, or organisms [94]. The third component, fluctuating asymmetry (FA), encompasses any random developmental and environmental effects, that cannot be explained with the existing knowledge. The observed deviations can be statistically linearly modeled as products of two effects, the hemisphere side studied and the individual specimen analyzed, as well as their interaction [71]. Given that the analysis is performed on a pair of symmetric objects, and not on a single symmetric object, this configuration is named **matching asymmetry analysis**. Formally, based on [129] assuming the presence of replications for each observation per individual, to account for technical error, a mixed linear model representing the aforementioned dependencies is defined as:

$$Y_{ijk} = \mu + \beta + I_i + S_{ij} + E_{ijk} \tag{2.1}$$

where Y_{ijk} is the phenotype of the i-th individual, from the j-th side, under the k-th replication, μ and β are the fixed intercept and fixed side effect respectively, $I_i \sim \mathcal{N}(0, \sigma^2_{ind})$ is the random individual effect, $S_{ij} \sim \mathcal{N}(0, \sigma^2_{FA})$ is the random

side and individual specific effect, matched to FA, and $E_{ijk} \sim \mathcal{N}(0, \sigma^2_{ME})$ is the measurement error. Given this definition, a way to measure the statistical significance is performed through an F-test applied on a 2-way nonparametric permutation-based ANOVA, to relate the RSS ratios of effects to observable error terms, and of fluctuating effect to the measurement error. Extra care needs to be given on the determination of the DOF of each term, given the preprocessing applied to bring the hemispheres surfaces into Kendall shape space [71]. Specifically, the constructed F-ratios, for each pair of contralateral landmarks coordinates separately, on N individuals and R replications, are:

$$F_I = \frac{RSS_I}{RSS_S},\ F_{DA} = (N-1)\frac{RSS_D}{RSS_S},\ F_{FA} = \frac{2(R-1)N}{N-1}\frac{RSS_S}{RSS_E}$$

RSS_I, RSS_D, RSS_S and RSS_E are the rows, columns, interaction and error RSSs respectively, as computed by MATLAB 'anova2' function on the $NR \times 2$ array that contains in each group of R rows information about each individual. Replications are necessary in such analysis, in order to distinguish the FA effect from measurement error, and manage to detect F_{FA}. To this end, an MRI test-retest subset of 20 individuals from HCP is retrieved [130], and the preprocessing mentioned in subsection 2.2.1 is performed. For each landmark and coordinate, and for each hemisphere separately, the mean observed replication variance across individuals is computed. Subsequently, assuming that the technical measurement error is normally distributed, an augmented dataset is produced for the MRI samples in the UK Biobank dataset. Three ($R = 3$) replications per individual are generated by sampling from the identified distributions.

While simple ANOVA bases the F-score significance on the assumption of normality, permutation-based ANOVA makes no assumptions on the underlying distribution [4]. Instead it bases significance of a F-statistic on the number of permutations which resulted in F-scores equal or higher than the F-statistic measured in the simple ANOVA scenario on the original data, divided by the total number of permutations [70]. By definition, the observable p-value resolution is the reciprocal of the permutations number. This is introduced in the fraction described above by adding 1 to the nominator and denominator, namely considering the non-permuted case as well. Let N be the number of individuals. A reshaping operation is performed, after which the set of size L landmarks becomes a set of size 3L coordinates per individual per side, in other words a 3D dataset. The permutations are generated considering each time the dimension being investigated, pursuing biologically feasible result when possible; for assessing the individual effect and computing RSS_I, the hemispheres are randomly shuffled across individuals (N^{3RL} possible orderings); for the side effect test and RSS_D measurement, landmarks of each individual are reassigned a random side (2^{NRL} configurations); for the fluctuating effect, quantified by RSS_{FA}, the whole dataset is randomly permuted ($(6NRL)!$ orderings). As it can be foreseen, a handicap of the method is the largely unequal size of the possible permutations among the components analyses. This fact renders the last test more sensitive to assign low p-values to each landmark, as the configuration that could possibly produce a better f-score is exponentially less likely to be selected. However,

it is worth noting that in all tests the possible cases number is prohibitively large, and that analyses in Monte Carlo simulations, suggest the size of 1000 replications as good enough [86].

The consecutive analysis has also been demonstrated in the work of Vanbiervliet et al. [131]. 1000 replications are selected to test the significance of each asymmetry component, which means that the analysis is computationally intensive, but facilitated by the downsampling described in subsection 2.2.3. Five random subsets of 50 samples are collected from the discovery dataset. The number of samples is chosen experimentally, as it was observed that the size of 1000 replications is actually not enough for larger datasets and the method is generally sensitive in assigning high significance (p-value$<$0.05) to each landmark, the larger the set size assessed. A number of different random subsets is selected, so that to reduce the effect of cherry-picking. The final counts are computed to be the average of the experimental iterations. Although the last step is not theoretically correct, as independent pools of combinations are assessed, it achieves a smoothing operation CITATION NEEDED.

2.2.5 Covariates control

In the case of GWAS, variance caused because of non-genetic factors needs to be excised from the underlying data. To this end, covariates adjustment is performed, by retrieving the residue of a partial least squares regression (PLSR) [56] describing D_A relatively to the factors mentioned in subsection 2.1.1, along with the 20 genetic PCs, to account for population stratification and reduce confounding biases (Table 2.3).

age(1)	age squared(1)
height(1)	weight (pre-imaging)(1)
diastolic blood pressure(1)	systolic blood pressure(1)
date of measurement(1)	genetic PCs (20))
volumetric scaling from T1 head image to standard space(1)	
X-position of center-of-gravity of brain mask in scanner coordinates(1)	
Y-position of back of brain mask in scanner coordinates(1)	
Z-position of center-of-gravity of brain mask in scanner coordinates(1)	
Z-position of table/coil in scanner coordinates(1)	
one-hot encoded assessment location (21)	
left & right hemisphere centroid sizes prior scaling (2)	

TABLE 2.3: Covariates used to control phenotype, totaling 57. Numbers in parenthesis show the dimensionality of each covariate.

2.2.6 Shapes Partitioning

The present work evaluates the brain asymmetry genetic landscape in a coarse-to-fine segmentation, through hierarchical spectral clustering (HSC). The technique has been used in a number of different related phenotypic studies [23][93], yielding results that are in accordance with the underlying anatomic features. The main reason

behind this partitioning is the intrinsic complexity of the studied phenotype, eliciting expected differences in the genomic profiles of each cerebral cortex region. This type of distance-based clustering is governed by the least quantity of assumptions, regarding the shape or form of the cluster [136]. The partitions' genetic juxtaposition is valuable for identifying which regions share similar significant genetic loci, highlighting the corresponding genes contribution, or showcasing the specialization of certain regions that share little to no similarities with their neighbors.

HSC is an unsupervised method of iterative partitioning, that makes use of the distance matrix eigenvectors [95]. The distance matrix between pairs of rows of D_A, across individuals, is constructed using RV coefficients [110], a generalization of Pearson correlation in N-dimensional space. The resulting matrix gets enhanced pairwise similarities, by becoming sharper through the application of the Laplacian transformation dictated by the Shi-Malik method [118]. The eigenvectors of the matrix are computed and Kmeans++ clustering [6] is applied with 2 clusters. The process is repeated on each cluster for a desirable amount of levels, resulting into a binary tree structure (i.e. each parent shape is partitioned into two children). In the current study, a level-4 partitioning is performed, resulting into a tree of 31 partitions, with the root partition, the entire hemisphere, included. Subsequently, the rows of D_A corresponding to each of those partitions are transformed by principal component analysis (PCA), keeping maximally 500 PCs that explain at most 80% of the variance, resulting into a structure P_A that contains 31 arrays, $P_{Ai}, i = 1..31$, one for each partition, with varying dimensionality. The last step is not only performed for reasons of dimensionality reduction and computational efficiency, but also to ensure that the resulting phenotypic traits are orthogonal with each other, and therefore compatible for LD score correlation (LDSC) analyses, discussed in subsection 2.4.5. The partitioning is derived by the discovery dataset only, so that the GWAS results between discovery and replication datasets correspond to the same partitions and are directly comparable. The computed clustering is subsequently compared to the DK atlas parcellation, through a symmetric measure of similarity called normalized mutual information (NMI). Let $U := \{\mathbf{v_i}\}, i = 1..n_{DK}$ and $V := \{\mathbf{v_j}\}, j = 1..31$ be the supersets of sets of indices in $\mathcal{V}_S$ associated to partitions by DK atlas and HSC respectively. Then, the NMI score is:

$$\mathrm{NMI}(U, V) := \frac{1}{S} \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

where in this study the normalizing factor S, generally variable [120], is defined as the square root of the product of the sets entropies, namely

$$S := \sqrt{\sum_{i=1}^{|U|} (|U_i| log |U_i|) \sum_{j=1}^{|V|} (|V_j| log |V_j|)}$$

This score is by definition independent on the ordering and the number of the labels. However, a differential number of labels which is expected between the measured

labels affects the maximally possible result. Therefore, the score is further normalized by scaling it by the approximate (based on integer division) theoretical maximum score that can observed, if each lesser partitioning does not 'share', in correspondence, any labels of the greater partitioning, apart from a single placeholder label, to account for integer division.

2.3 Methods applied on genotype

2.3.1 Population-based filtering

PCA, using as reference data the 1000G (Phase 3), is applied to select European individuals only. First, SNPs in LD, as computed by PLINK 1.9 with parameters 50 variant window-size, 5 variant step size and 0.2 r^2, are excluded from the reference dataset. KMeans algorithm is fitted on 25 PCs of the reference dataset. Then, only individuals from discovery and replication datasets that are included in the clusters with a EURO label are considered. The individuals identifiers of the genotype are consecutively matched to those in the phenotype.

2.3.2 SNP-based filtering and grouping

In addition, SNPs referring to indels, with low genotyping rate (<50%) , low MAF < 1% (i.e. corresponding to rare alleles), or not in Hardy-Weinberg equilibrium (P<10.6) are excluded from the analysis. Subsequently, the LD filtering applied on the reference dataset is also done for the remaining SNPs of the discovery and replication data. The filtered discovery and replication datasets contain 9,705,931 and 8,305,363 variants respectively. The multi-allelic SNPs rows are grouped together, so that single association test is applied.

2.4 Genome-to-phenotype association

2.4.1 mvGWAS CCA

For each i-th partition, CCA is applied between each SNP and P_{Ai}. The operation is repeated for both discovery and replication datasets. The produced χ^2 scores are transformed into the quantity dictated by multivariate LDSC analysis. The resulting p values are transformed into -log10P values and the GWAS results are compared qualitatively, as well using LDSC genetic correlation analysis.

2.4.2 Discovery and replication results aggregation

Once the comparisons between the results from discovery and replication datasets are made, they are aggregated into a single output, with the p-values combined using Stouffer's method. This method is applied by first projecting the p-values corresponding to the i-th sample into z-scores, through the calculation of the complementary

inverse error function erfc^{-1} of $p_i = \{p_{i1}, p_{i2}\}$. The combined p-value p_{ci} then is:

$$p_{ci} = \text{erfc}\left(\frac{\sum_{k=1}^{2} \text{erfc}^{-1}(p_{ik})}{\sqrt{2}}\right)$$

Apart from this, the χ^2 scores are summed, under the theoretical basis that the sum of two independent χ^2 values with d_1 and d_2 DOFs respectively follows a χ^2 distribution with $d_1 + d_2$ DOFs.

2.4.3 Heritability and confounding bias detection

The LD between two alleles A and B from different loci is generally quantified using one of the following values:

- the coefficient of linkage disequilibrium $\mathcal{D}$:

  $$\mathcal{D} := p_{AB} - p_A p_B$$

  with p_{AB} referring to the haplotype AB frequency and p_i to the frequency of allele i. This coefficient is scaled by theoretical maximum $\mathcal{D}$, $\mathcal{D}_{max}$, to render it independent of the per-pair frequencies magnitudes, producing $\mathcal{D}'$.

- the genetic correlation r^2, a proxy of the Pearson coefficient, defined by:

  $$r^2 := \frac{\mathcal{D}}{p_A(1-p_A)p_B(1-p_B)}$$

Based on simulations, it has been shown that $\mathcal{D}'$ is inflated when the sample size is small or the minor allele is rare [124], thus genetic correlation is generally preferred. In a seminal research work from Bulik-Sullivan et al. [15], it was found that there is a closed mathematical expression that connects the j-th allele χ^2 expected value with the average heritability explained per SNP h and its LD score, defined by $\sum_k r_{jk}^2$, r_{jk}^2 being the r^2 of the j-th with the k-th allele:

$$E[\chi^2|l_j] = \frac{Nh^2 l_j}{M} + N\alpha + 1$$

α is the contribution of population-related effects, such as population stratification, that are not being controlled, known as confounding biases. The gains from this regression are dual; a measurement of heritability can be obtained by estimating the slope, and the confounding bias effect can be measured by the intercept. This formula, which originally referred to a univariate phenotype, was extended in [93] to incorporate D-dimensional multivariate traits:

$$E\left[\frac{\chi_j^2}{D\left(1+\frac{\chi_j^2}{N}\right)}\right] = \frac{N-1}{P}\left(\frac{\sum_{d=1}^{D} h_d^2}{D}\right) l_j + 1 + O\left(\frac{1}{N}\right)$$

where $O(1/N)$ term is corresponding to the confounding biases effect. Therefore, this tool is used to estimate heritability and confounding biases, per partition, from the combined GWAS results. The basic underlying assumptions, or limitations, of such a model are:

- The SNP heritability follows a uniform distribution, i.e. it is on average the same genome-wide. Extensions have been made to relax this, generally wrong [126], assumption, by considering partitions of SNPs separately and doing what is known as stratified LD score regression (LDSR) [46, 47].
- Each SNP effect is assessed independently from the rest, therefore no between-SNPs interactions can be included in the computation.
- The covariance matrix of the phenotype equals the identity matrix multiplied by N, that is the studied traits are orthogonal to each other.

Another limitation of LDSR is that the heritability is under-estimated when the effective sample size is small[77].

2.4.4 Tissue and cell-type specific heritability analysis

As mentioned in subsection 2.4.3, extensions have been devised to account for different heritability profiles of disparate biological and functional regions. Finucane et al. [47] incorporate the notion of chromatin regulation and gene expression profiles in LDSR. Let m be the number different gene expression profiles, each on a separate tissue or a cell type, an annotation class. For each of those a set of genes, corresponding to a set of SNPs $K_j, (j = 1..m)$, has been found to be significantly enriched, relatively to the rest of the comparison group. The discussed extension is of the form:

$$E\left[\chi_i^2\right] = N\sum_j \tau_j l(i,j) + N\alpha + 1$$

with the LD score of the i-th SNP for the j-th genes cluster $l(i,j)$ being defined as $\sum_{k\in K_j} r_{ik}^2$. τ_j is the estimated coefficient per annotation class and explains the signed effect of each class on the heritability of the observed phenotype. In other words, the coefficient τ_j relates the cumulative LD effect of the j-th set of SNPs with the observed capacity of the i-th SNP to affect the phenotypic trait.

Under this framework, in the present study, a significant association is sought between chromatin regulation studies related to gene expression, to identify the degree under which the identified significant SNPs effects are also likely to be regulated by epigenetic modifications, using the same dataset used by Finucane et al. [47], as mentioned in subsection 2.1.2.

2.4.5 Genetic correlation analyses

Bulik-Sullivan et al. [14] also invented a way to utilize GWAS scores produced for two different traits as a proxy to relate the genetic correlation of these traits, namely

the extent over which the two characteristics are being regulated by similar genetic drivers:

$$E[z_{1j}z_{2j}|l_j] = \frac{\sqrt{N_1N_2}\rho_g}{M}l_j + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

The conversion between χ^2 with 1 DOF and z score values is straightforward, as, by definition, the square of a standard normal distribution follows the χ^2 one with 1 DOF. In other words, the equation above retrogresses to the LDSR one, if traits 1 and 2 are considered the same. With LDSC, seemingly independent phenotypes can be compared, testing for pleiotropic SNP effects and discovering novel biological pathways [14]. As immediate result by comparison, the correlation is retrieved to be:

This method is used to relate the GWAS scores presented in subsection 2.1.2. One of those comparisons is done versus the work from Sha et al. [116], with the aim to quantitatively relate the results presented there to the current study. In addition, as an internal measure of similarity, LDSC is applied to measure the degree of concordance between the results presented by the discovery and the replication dataset. LDSC fails to be applied when the heritability of the partition of the studied dataset (discovery, replication or combined) is assessed to be prohibitively low to make an assessment of the genetic correlation of that partition with a trait.

2.4.6 Gene-based developmental and functional analysis

Last but crucial step, analyses at the gene level are performed. The focus is mainly shifted on 5 partitions; the entire hemisphere, and the 4 ones on the second level, as greater discrepancies in GWAS are identified among partitions at that level relatively to others. The subsequent analysis refers to the process applied on each partition separately.

Initially, underlying gene sets are retrieved, in the exact same way as it was done by Sha et al. [116]. More specifically, FUMA toolbox SNP2GENE utility is used [139], by accumulating the outputs of positional, eQTL and chromatin interaction mapping, with default parameters only taking into consideration brain-related samples when dimmed necessary [116, 139].

Lead SNPs are also retrieved by applying a GRM r^2 upper cutoff of 0.1 on the list of significant SNPs, selected imposing a 5×10^{-8} threshold on the produced p-values $\mathbf{p_c}$. A further extension of the gene sets is achieved by using GREAT tool [87], supplying it with the identified lead SNPs. This tool models possible gene regulatory domains based on empirical evidence and assigns SNPs in such intronic regions to the corresponding genes [87].

Leveraging the power of another statistical tool through FUMA, MAGMA, a time-dependent analysis is performed, identifying the degrees under which identified genes are enriched in genetic expression profiles from brain tissues from different developmental stages [30]. MAGMA gene-set analysis uses the full distribution of SNP p-values, hence it is fundamentally different from a GSEA kind of test. However, long-range relationships, assessed and introduced by the methodologies of FUMA and GREAT, defined above, are not considered. Instead, MAGMA process examines

the joint association signals of all SNPs within a given gene, in a 100kb region, while considering the LD between those SNPs [30, 116].

The resulting gene set is supplied to the FUMA GENE2FUNC process, where a variety of GSEAs takes place [139], extracting functional relationships with biological pathways. Concurrently, the identified gene set interactions are graphically represented, significantly enriched publications are identified and gene conservation profiles are inspected using the STRING suite [122].

Chapter 3

Results

3.1 Statistical brain shape analysis

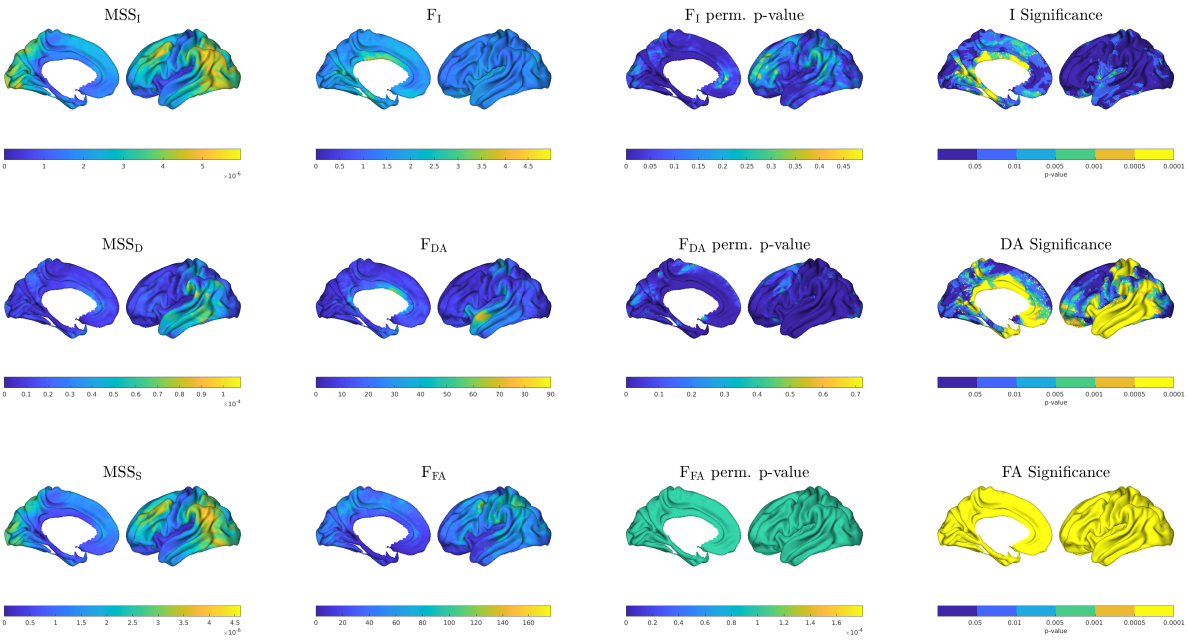


FIGURE 3.1: Asymmetry components significance analysis. Information about each landmark has been used to fill the 3D average shape of the left hemisphere, displayed medially and laterally. The mean sums of squares are produced by dividing RSSs with the appropriate DOFs. Both the first and the second column refer to information retrieved from the non-permuted data. The last column is derived by applying various thresholds (shown in the colorbar) on the p-value response presented on the third column.

The main results from the 2-way permutation ANOVA are summarized in Figure 3.1 and are also being more formally medically followed by Vanbiervliet et al. [131]. The individual effect (first row), which corresponds to the average shape variability across individuals, exhibits greater variation in the medial surface, particularly in the middle-anterior and middle-posterior parts of the cingulate gyrus and sulcus (⊆BA28, processing emotions and behavior regulation), the rostral part of the cuneus (⊆BA17, processing of visual information), the parahippocampal gyrus (⊆BA27, memory encoding and retrieval), and the fusiform gyrus (⊆BA37, recognition of faces).

DA (second row), the focus of this study, relevant to the general aptitude of individuals to exhibit certain asymmetric traits, is found to be highly significant in almost half of the studied surface. In line with the general identified asymmetry patterns presented in subsection 1.1.6, it is greatly localized around the peri-sylvian fissure and the temporal lobe (see peri-sylvian asymmetry). Also, it occurs in the medial surface, and the occipital lobe, implying relationship with the yakovlevian torque, although lower significance of the effect is demonstrated on the prefrontal lobe.

FA effect (last row), which, as a reminder, is related to environmentally and developmentally induced variations, has been found generally significant across the cortical surface. This finding can be partly justified to the large combinatorial space from where permutations are collected (see subsection 2.2.4), as well as the overall exhibited plasticity of the human cortex (see subsection 1.1.3), as raised by Vanbiervliet et al. [131]. However, a comparison across regions is possible by inspecting F_{FA} instead, where, in the caudal part of the middle frontal gyrus (⊆ BA40, phonological processing and emotional responses), the superior part of the precentral gyrus (⊆ BA07, space localization), and the caudal part of the superior frontal sulcus and gyrus (⊆ BA08, planning complex movements) greater effect of FA is exhibited.

3.2 Covariates control

In Figure 3.2, the average covariates explained variance on each segment of DK atlas is observed as retrieved from PLSR. The largest part of the frontal lobe, with greater impact on the inferior frontal, the inferior parietal gyrus and the parahippocampal gyrus appear to be more correlated with the collected metadata, shown in Table 2.3. The results point to less observed explained variation around the area of the sylvian fissure and the temporal lobe, and also raise a degree of uncertainty on the significant response observed during the statistical shape asymmetry analysis, regarding the medial surface and the inferior part of the frontal lobe.

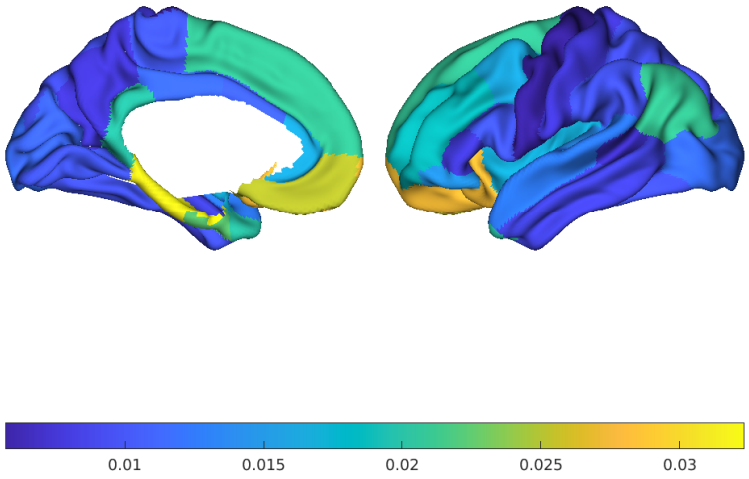
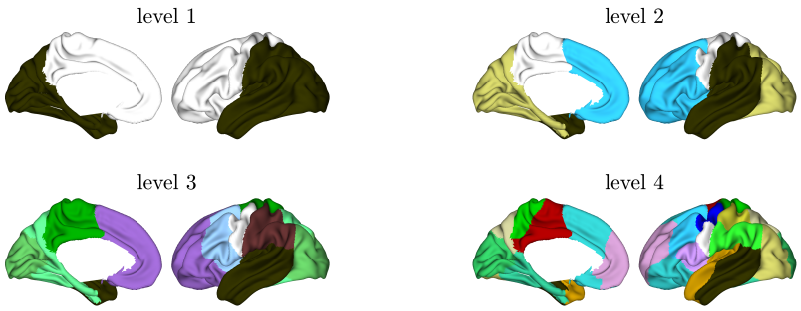


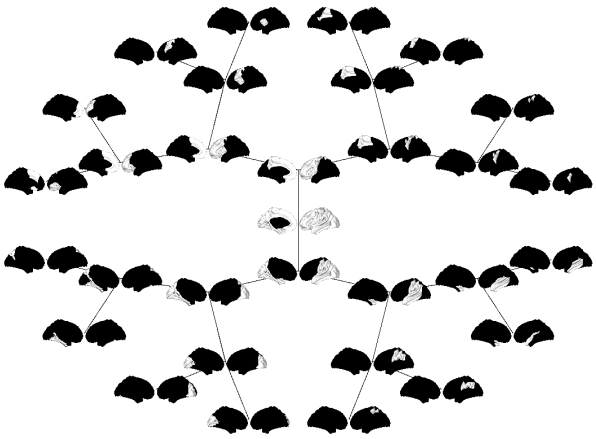
FIGURE 3.2: Explained variance from the covariates on each DK atlas segment, as retrieved from PLSR, mapped on the average left hemisphere medial and lateral side.

3.3 Partitioning and PCA

In Figure 3.3, the partitioning produced from the application of HSC is displayed. On the first level, although the cross-section accurately follows the sylvian fissure on the lateral part and partitions the frontal lobe from the rest of the hemisphere, on the medial surface it appears to split the precuneus in half ($\subseteq$ BA07). On the second level, the occipital lobe is separated from the temporal lobe, while the central gyrus appears to be dissected from the frontal lobe, while inspected the lateral surface. On the medial surface, the paracentral gyrus ($\subseteq$ BA04)is approximately separated from the superior-frontal, whereas another cross section appears to share its boundaries with the temporal pole ($\subseteq$ BA38). In general, the unsupervised clustering follows the functional partitioning, validating the close relationship between function and morphology of each cortical region. The calculated NMI score for the partitioning, compared to the DK atlas is displayed in Table 3.1 for each level. Although the finer the partitioning, the further away from the theoretical maximum value, quantitatively the two clusterings highly agree.



(A)



(B)

FIGURE 3.3: 4-level brain shape partitioning based on cortical asymmetry, using HSC. Shown in 2 different versions, on a level representation (top), and as a polar dendrogram plot, annotated with white color against black background (bottom). Those representations are used across the coarse-to-fine analysis in this study.

	NMI	NMI_{max}	ratio
Lvl 1	0.37	0.48	0.78
Lvl 2	0.49	0.66	0.74
Lvl 3	0.55	0.79	0.71
Lvl 4	0.62	0.90	0.69

TABLE 3.1: NMI scores across HSC partitioning levels, comparing DK atlas with computed partitioning levels. NMI_{max} is an approximate maximal value, given the different number of partitions in each clustering, and ratio is the scaled NMI using that value.

The required number of PCA features per partition given the constraints, as

computed by assessing the discovery dataset, is displayed in Figure 3.4. A significant dimensionality reduction was achieved, given that the total number of landmark coordinates per individual is 89367.

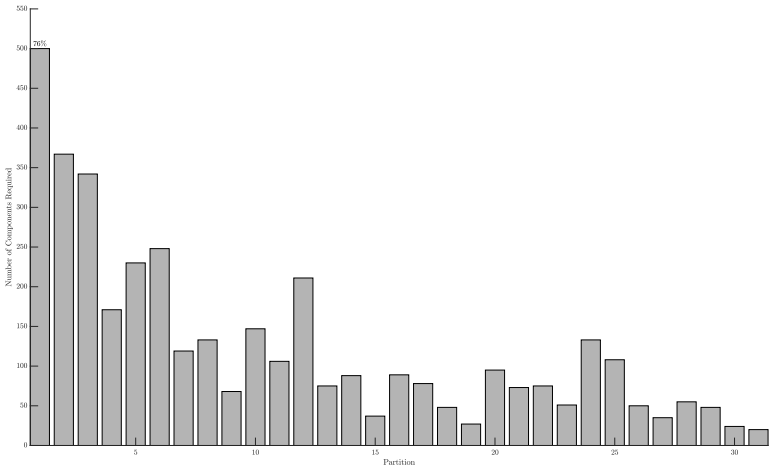


FIGURE 3.4: Number of PCs for each HSC cortical surface partition, required to explain 80% of its variance, relatively to the discovery dataset. For the first partition, the upper limit of 500 components is reached and only 74% of its variance is explained.

3.4 GWAS

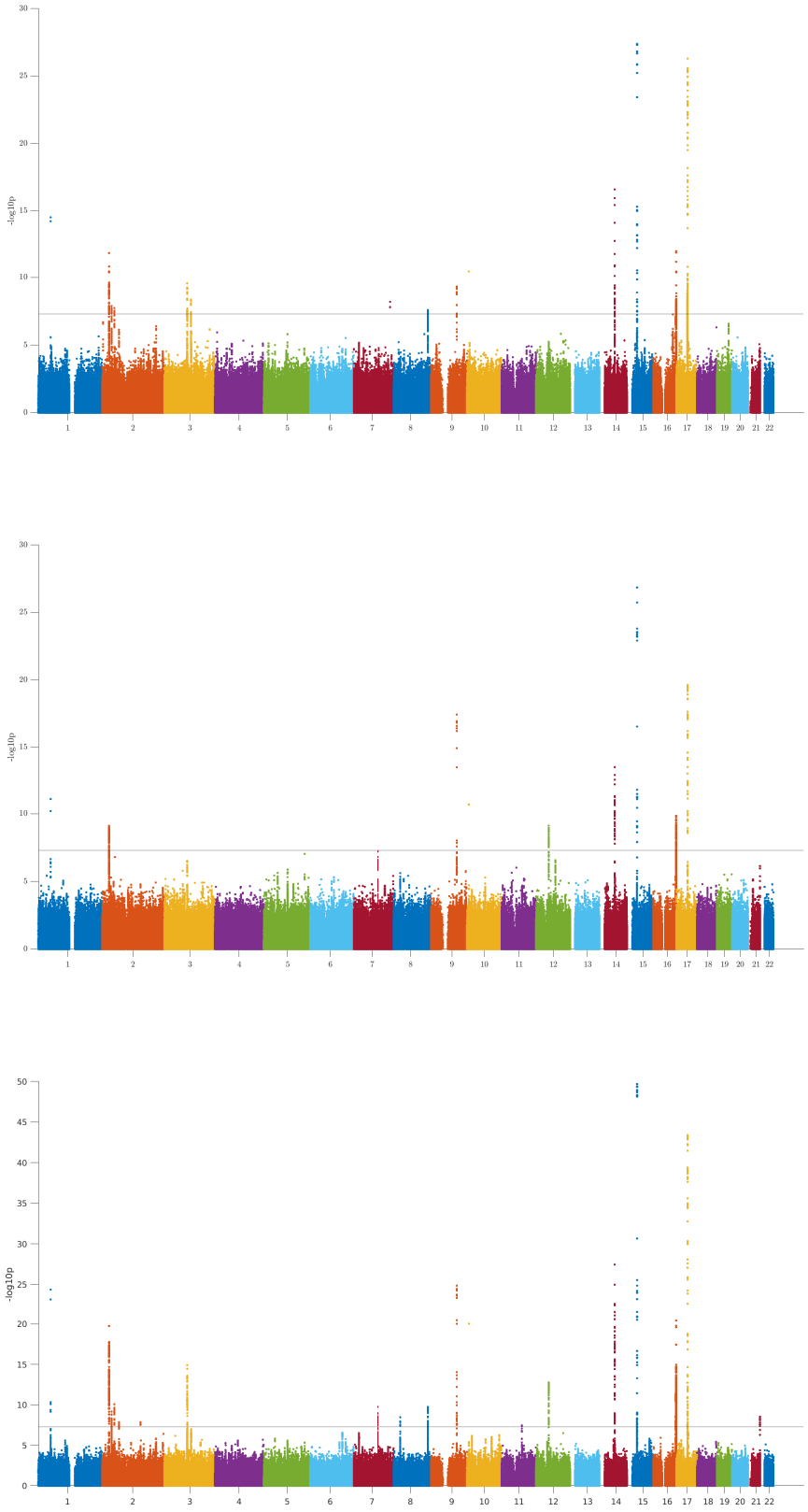


FIGURE 3.5: GWAS of the entire hemisphere shape asymmetry computed on the discovery (top) and the replication (middle) dataset, along with the meta-analysis union based on Stouffer's method (bottom).

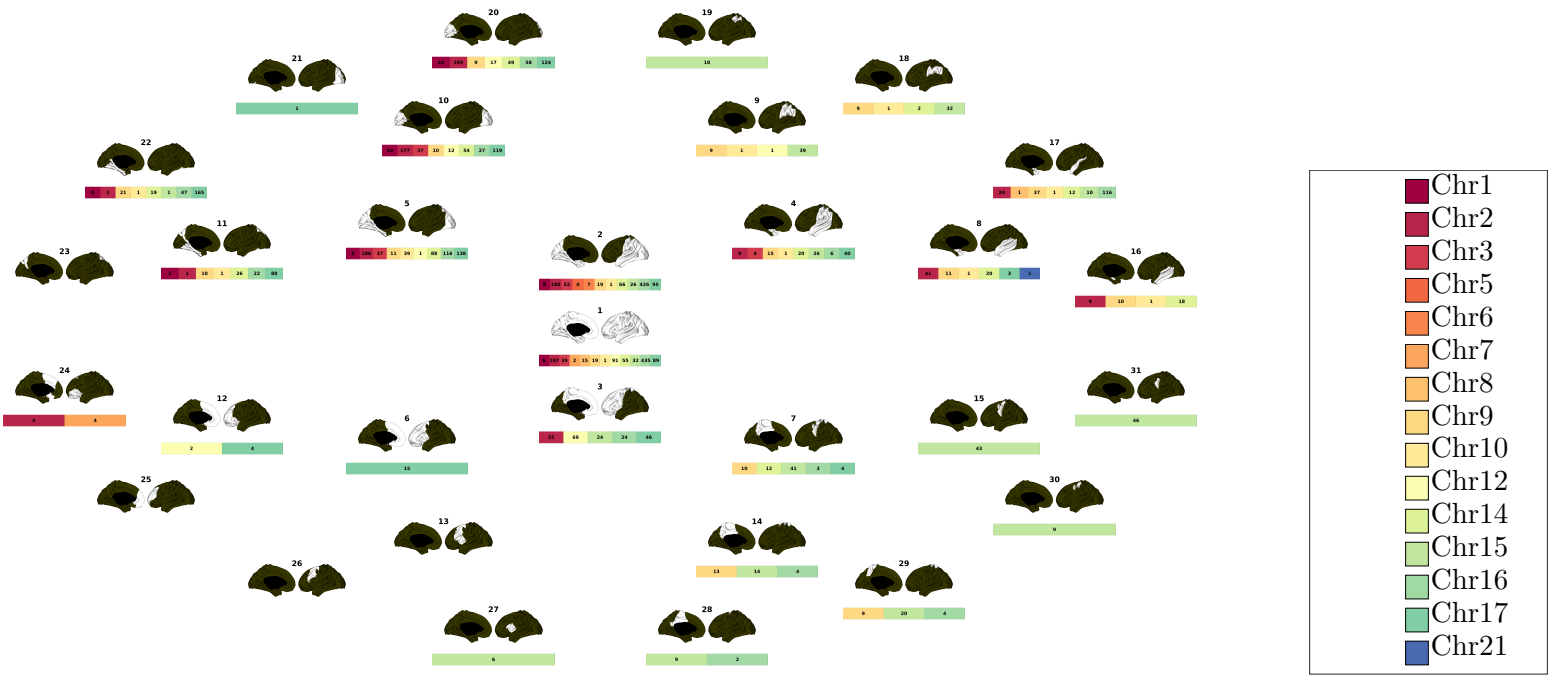


FIGURE 3.6: Number of significant SNPs per chromosome after Bonferroni correction across partitions. The number per chromosome is shown inside boxes of a certain color.

3.5 LDSR

3.6 LDSC

3.7 LDSC-SEG

3.8 Developmental analysis

3.9 Functional association

3.10 Evolutionary studies

Chapter 4

Discussion

Bibliography

[1] E. Aaku-Saraste, B. Oback, A. Hellwig, and W. B. Huttner. Neuroepithelial cells downregulate their plasma membrane polarity prior to neural tube closure and neurogenesis. *Mechanisms of Development*, 69(1-2):71–81, dec 1997. ISSN 0925-4773. doi: 10.1016/S0925-4773(97)00156-1.

[2] R. S. Abu-Rustum, M. F. Ziade, and S. E. Abu-Rustum. Reference Values for the Right and Left Fetal Choroid Plexus at 11 to 13 Weeks. *Journal of Ultrasound in Medicine*, 32(9):1623–1629, sep 2013. ISSN 1550-9613. doi: 10.7863/ULTRA.32.9.1623.

[3] D. Alejandro Gonzalez-Chica, J. Luiz Bastos, R. Pereira Duquia, R. Rangel Bonamigo, and J. Martínez-Mesa. EPIDEMIOLOGY AND BIOSTATISTICS APPLIED TO DERMATOLOGY. *An Bras Dermatol*, 90(4):523–531, 2015. doi: 10.1590/abd1806-4841.20154289.

[4] M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, feb 2001. ISSN 1442-9993. doi: 10.1111/J.1442-9993.2001.01070.PP.X.

[5] P. D. Arnold et al. Revealing the complex genetic architecture of obsessive–compulsive disorder using meta-analysis. *Molecular Psychiatry 2018 23:5*, 23(5):1181–1188, aug 2017. ISSN 1476-5578. doi: 10.1038/mp.2017.154.

[6] D. Arthur, D. Arthur, and S. Vassilvitskii. K-means++: the advantages of careful seeding. *IN PROCEEDINGS OF THE 18TH ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS*, 2007.

[7] M. Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, may 2000. ISSN 10614036. doi: 10.1038/75556.

[8] A. Auton et al. A global reference for human genetic variation. *Nature 2015 526:7571*, 526(7571):68–74, sep 2015. ISSN 1476-4687. doi: 10.1038/nature15393.

[9] S. Aw and M. Levin. Is left-right asymmetry a form of planar cell polarity? *Development*, 136(3):355–366, feb 2009. ISSN 0950-1991. doi: 10.1242/DEV.015974.

[10] S. Aw, D. S. Adams, D. Qiu, and M. Levin. H,K-ATPase protein localization and Kir4.1 function reveal concordance of three axes during early determination of left–right asymmetry. *Mechanisms of Development*, 125(3-4):353–372, mar 2008. ISSN 0925-4773. doi: 10.1016/J.MOD.2007.10.011.

[11] A. Balzeau, E. Gilissen, and D. Grimaud-Hervé. Shared Pattern of Endocranial Shape Asymmetries among Great Apes, Anatomically Modern Humans, and Fossil Hominins. *PLoS ONE*, 7(1):e29581, jan 2012. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0029581.

[12] J. W. Belmont et al. The International HapMap Project. *Nature 2004 426:6968*, 426(6968):789–796, dec 2003. ISSN 1476-4687. doi: 10.1038/nature02168.

[13] K. Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Johann Ambrosius Barth, Leipzig, 1909.

[14] B. Bulik-Sullivan et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics 2015 47:11*, 47(11):1236–1241, sep 2015. ISSN 1546-1718. doi: 10.1038/ng.3406.

[15] B. Bulik-Sullivan et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics 2015 47:3*, 47(3):291–295, feb 2015. ISSN 1546-1718. doi: 10.1038/ng.3211.

[16] A. Buniello et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, jan 2019. ISSN 1362-4962. doi: 10.1093/NAR/GKY1120.

[17] E. Cano-Gamez and G. Trynka. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, 11:424, may 2020. ISSN 16648021. doi: 10.3389/FGENE.2020.00424/BIBTEX.

[18] M. L. Cara, I. Streata, A. M. Buga, and D. G. Iliescu. Developmental Brain Asymmetry. The Good and the Bad Sides. *Symmetry 2022, Vol. 14, Page 128*, 14(1):128, jan 2022. ISSN 2073-8994. doi: 10.3390/SYM14010128.

[19] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics*, 14, apr 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-128.

[20] M. Chini and I. L. Hanganu-Opatz. Prefrontal Cortex Development in Health and Disease: Lessons from Rodents and Humans. *Trends in Neurosciences*, 44 (3):227–240, mar 2021. ISSN 0166-2236. doi: 10.1016/J.TINS.2020.10.017.

[21] A. Cichonska et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*, 32(13):1981–1989, jul 2016. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTW052.

[22] P. Claes, M. Walters, M. D. Shriver, D. Puts, G. Gibson, J. Clement, G. Baynam, G. Verbeke, D. Vandermeulen, and P. Suetens. Sexual dimorphism in multiple aspects of 3D facial symmetry and asymmetry defined by spatially dense geometric morphometrics. *Journal of Anatomy*, 221(2):97–114, aug 2012. ISSN 1469-7580. doi: 10.1111/J.1469-7580.2012.01528.X.

[23] P. Claes et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics*, 50(3):414–423, 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0057-4.

[24] M. Claussnitzer et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans . *New England Journal of Medicine*, 373(10):895–907, sep 2015. ISSN 0028-4793. doi: 10.1056/NEJMOA1502214/SUPPL_FILE/NEJMOA1502214_DISCLOSURES.PDF.

[25] P. Coles. Statistical errors and asymmetry indices. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45), nov 2008. ISSN 00278424. doi: 10.1073/PNAS.0806646105.

[26] M. L. Concha, I. H. Bianco, and S. W. Wilson. Encoding asymmetry within neural circuits. *Nature Reviews Neuroscience 2012 13:12*, 13(12):832–843, nov 2012. ISSN 1471-0048. doi: 10.1038/nrn3371.

[27] M. C. Corballis. The evolution and genetics of cerebral asymmetry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1519):867, 2009. ISSN 14712970. doi: 10.1098/RSTB.2008.0232.

[28] M. R. Corces et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nature genetics*, 52(11):1158, nov 2020. ISSN 15461718. doi: 10.1038/S41588-020-00721-X.

[29] C. G. de Kovel and C. Francks. The molecular genetics of hand preference revisited. *Scientific reports*, 9(1), dec 2019. ISSN 2045-2322. doi: 10.1038/S41598-019-42515-0.

[30] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, 11(4): e1004219, apr 2015. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1004219.

[31] Ö. de Manzano and F. Ullén. Same Genes, Different Brains: Neuroanatomical Differences Between Monozygotic Twins Discordant for Musical Training. *Cerebral Cortex*, 28(1):387–394, jan 2018. ISSN 1047-3211. doi: 10.1093/CERCOR/BHX299.

[32] D. Demontis et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, 51(1):63–75, jan 2019. ISSN 1546-1718. doi: 10.1038/S41588-018-0269-7.

[33] R. S. Desikan et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, jul 2006. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2006.01.021.

[34] C. Destrieux, B. Fischl, A. Dale, and E. Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15, oct 2010. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2010.06.010.

[35] E. W. Dickie, A. Anticevic, D. E. Smith, T. S. Coalson, M. Manogaran, N. Calarco, J. D. Viviano, M. F. Glasser, D. C. Van Essen, and A. N. Voineskos. ciftify: A framework for surface-based analysis of legacy MR acquisitions. *NeuroImage*, 197:818, aug 2019. ISSN 10959572. doi: 10.1016/J.NEUROIMAGE.2019.04.078.

[36] C. Do, A. Shearer, M. Suzuki, M. B. Terry, J. Gelernter, J. M. Greally, and B. Tycko. Genetic–epigenetic interactions in cis: a major focus in the post-GWAS era. *Genome Biology 2017 18:1*, 18(1):1–22, jun 2017. ISSN 1474-760X. doi: 10.1186/S13059-017-1250-Y.

[37] I. Dunham et al. An integrated encyclopedia of DNA elements in the human genome. *Nature 2012 489:7414*, 489(7414):57–74, sep 2012. ISSN 1476-4687. doi: 10.1038/nature11247.

[38] L. I. Dyck and E. M. Morrow. Genetic control of postnatal human brain growth. *Current opinion in neurology*, 30(1):114, 2017. ISSN 14736551. doi: 10.1097/WCO.0000000000000405.

[39] A. Ernst and J. Frisén. Adult Neurogenesis in Humans- Common and Unique Traits in Mammals. *PLoS Biology*, 13(1), 2015. ISSN 15457885. doi: 10.1371/JOURNAL.PBIO.1002045.

[40] M. F. Bear, B. W. Connors, and M. A. Paradiso. Appendix: An illustrated guide to human neuroanatomy. In *Exploring the brain*, chapter 7, pages 205–262. Wolters Kluwer, 4th edition, 2016. ISBN 0781778174.

[41] M. F. Bear, B. W. Connors, and M. A. Paradiso. Formation of the Neural Tube-Three Primary Brain Vesicles. In *Exploring the brain*, chapter 7, pages 193–196. Wolters Kluwer, 4th edition, 2016. ISBN 0781778174.

[42] M. F. Bear, B. W. Connors, and M. A. Paradiso. Differentiation of the Forebrain. In *Exploring the brain*, chapter 7, pages 196–199. Wolters Kluwer, 4th edition, 2016. ISBN 0781778174.

[43] S. B. Fernandes, K. S. Zhang, T. M. Jamann, and A. E. Lipka. How Well Can Multivariate and Univariate GWAS Distinguish Between True and Spurious Pleiotropy? *Frontiers in Genetics*, 11:1747, jan 2021. ISSN 16648021. doi: 10.3389/FGENE.2020.602526/BIBTEX.

[44] A. Ferng and D. Mytilinaios. Brodmann areas: Anatomy and functions | Kenhub, 2022.

[45] J. R. Finnerty. The origins of axial patterning in the metazoa: how old is bilateral symmetry? *International Journal of Developmental Biology*, 47(7-8): 523–529, dec 2003. ISSN 0214-6282. doi: 10.1387/IJDB.14756328.

[46] H. K. Finucane et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics 2015 47:11*, 47 (11):1228–1235, sep 2015. ISSN 1546-1718. doi: 10.1038/ng.3404.

[47] H. K. Finucane et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics 2018 50:4*, 50 (4):621–629, apr 2018. ISSN 1546-1718. doi: 10.1038/s41588-018-0081-4.

[48] T. E. Galesloot, K. Van Steen, L. A. Kiemeney, L. L. Janss, and S. H. Vermeulen. A comparison of multivariate genome-wide association methods. *PLoS ONE*, 9 (4), apr 2014. ISSN 19326203. doi: 10.1371/JOURNAL.PONE.0095923.

[49] M. D. Gallagher and A. S. Chen-Plotkin. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, 102(5):717–730, may 2018. ISSN 0002-9297. doi: 10.1016/J.AJHG.2018.04.002.

[50] M. F. Glasser and D. C. van Essen. Mapping Human Cortical Areas In Vivo Based on Myelin Content as Revealed by T1- and T2-Weighted MRI. *The Journal of Neuroscience*, 31(32):11597, aug 2011. ISSN 02706474. doi: 10.1523/JNEUROSCI.2180-11.2011.

[51] M. Götz and W. B. Huttner. The cell biology of Neurogenesis. *Nat Rev Mol Cell Biol*, 6:777–788, 2005. doi: 10.1038/nrm1739.

[52] J. H. Graham, N. Banura, A. Ohki, and S. Saito. Asymmetry Index Evaluation of Cerebral Volume and Cerebral Blood Flow in Neonatal Hypoxic and Ischemic Encephalopathy. *Symmetry 2022, Vol. 14, Page 596*, 14(3):596, mar 2022. ISSN 2073-8994. doi: 10.3390/SYM14030596.

[53] D. T. Grimes and R. D. Burdine. Left-right patterning: breaking symmetry to asymmetric morphogenesis. *Trends in genetics : TIG*, 33(9):616, sep 2017. ISSN 13624555. doi: 10.1016/J.TIG.2017.06.004.

[54] J. Grove et al. Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51(3):431–444, mar 2019. ISSN 1546-1718. doi: 10.1038/S41588-019-0344-8.

[55] T. Guadalupe et al. Asymmetry within and around the human planum temporale is sexually dimorphic and influenced by genes involved in steroid hormone receptor activity. *Cortex; a journal devoted to the study of the nervous system and behavior*, 62:41–55, jan 2015. ISSN 1973-8102. doi: 10.1016/J.CORTEX.2014.07.015.

[56] D. V. Guebel and N. V. Torres. Partial Least-Squares Regression (PLSR). *Encyclopedia of Systems Biology*, pages 1646–1648, 2013. doi: 10.1007/978-1-4419-9863-7_1274.

[57] M. Guindo-Martínez et al. The impact of non-additive genetic associations on age-related complex diseases. *Nature communications*, 12(1), dec 2021. ISSN 2041-1723. doi: 10.1038/S41467-021-21952-4.

[58] P. Heger, W. Zheng, A. Rottmann, K. A. Pan Lio, and T. Wiehe. The genetic factors of bilaterian evolution. *eLife*, 9:1–45, jul 2020. ISSN 2050084X. doi: 10.7554/ELIFE.45530.

[59] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–589, may 2010. ISSN 1097-4164. doi: 10.1016/J.MOLCEL.2010.05.004.

[60] A. Hejnol and K. Pang. Xenacoelomorpha's significance for understanding bilaterian evolution. *Current Opinion in Genetics And Development*, 39:48–54, aug 2016. ISSN 0959-437X. doi: 10.1016/J.GDE.2016.05.019.

[61] M. R. Herbert et al. Brain asymmetries in autism and developmental language disorder: a nested whole-brain analysis. *Brain*, 128(1):213–226, jan 2005. ISSN 0006-8950. doi: 10.1093/brain/awh330.

[62] Y. Huo, S. Li, J. Liu, X. Li, and X. J. Luo. Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nature Communications 2019 10:1*, 10(1):1–19, feb 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08666-4.

[63] D. Jackson, R. Riley, and I. R. White. Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, 30(20):2481, sep 2011. ISSN 02776715. doi: 10.1002/SIM.4172.

[64] R. Janky et al. iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLOS Computational Biology*, 10(7): e1003731, 2014. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1003731.

[65] I. E. Jansen et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature genetics*, 51(3):404–413, mar 2019. ISSN 1546-1718. doi: 10.1038/S41588-018-0311-9.

[66] T. L. Jernigan et al. The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. *NeuroImage*, 124:1149–1154, jan 2016. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2015.04.057.

[67] H. J. Kang et al. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, oct 2011. ISSN 1476-4687. doi: 10.1038/NATURE10523.

[68] R. J. Klein et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, apr 2005. ISSN 00368075. doi: 10.1126/SCIENCE.1109557/SUPPL_FILE/KLEIN_SOM.PDF.

[69] C. P. Klingenberg. Walking on Kendall's Shape Space: Understanding Shape Spaces and Their Coordinate Systems. *Evolutionary Biology*, 47(4):334–352, dec 2020. ISSN 19342845. doi: 10.1007/S11692-020-09513-X/FIGURES/9.

[70] C. P. Klingenberg and G. S. Mcintyre. Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with procrustes methods. *Evolution*, 52(5):1363–1375, oct 1998. ISSN 1558-5646. doi: 10.1111/J.1558-5646.1998.TB02018.X.

[71] C. P. Klingenberg, M. Barluenga, and A. Meyer. Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. *Evolution*, 56(10):1909–1920, 2002. ISSN 00143820, 15585646.

[72] E. M. Koch and S. R. Sunyaev. Maintenance of Complex Trait Variation: Classic Theory and Modern Data. *Frontiers in Genetics*, 12:2198, nov 2021. ISSN 16648021. doi: 10.3389/FGENE.2021.763363/BIBTEX.

[73] X. Z. Kong et al. Mapping cortical brain asymmetry in 17,141 healthy individuals worldwide via the ENIGMA consortium. *Proceedings of the National Academy of Sciences of the United States of America*, 115(22):E5154–E5163, may 2018. ISSN 10916490. doi: 10.1073/PNAS.1718418115/SUPPL_FILE/PNAS.1718418115.SD07.XLSX.

[74] X. Z. Kong, M. Postema, D. Schijven, A. C. Castillo, A. Pepe, F. Crivello, M. Joliot, B. Mazoyer, S. E. Fisher, and C. Francks. Large-Scale Phenomic and Genomic Analysis of Brain Asymmetrical Skew. *Cerebral Cortex (New York, NY)*, 31(9):4151, sep 2021. ISSN 14602199. doi: 10.1093/CERCOR/BHAB075.

[75] X.-Z. Kong et al. Mapping brain asymmetry in health and disease through the ENIGMA consortium. *Human Brain Mapping*, 43(1):167–181, jan 2022. ISSN 1065-9471. doi: https://doi.org/10.1002/hbm.25033.

[76] F. Kuo and T. F. Massoud. Structural asymmetries in normal brain anatomy: A brief overview. *Annals of Anatomy - Anatomischer Anzeiger*, 241:151894, apr 2022. ISSN 0940-9602. doi: 10.1016/J.AANAT.2022.151894.

[77] J. J. Lee, M. McGue, W. G. Iacono, and C. C. Chow. The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genetic epidemiology*, 42(8):783, dec 2018. ISSN 10982272. doi: 10.1002/GEPI.22161.

[78] S. Li, S. Wang, X. Li, Q. Li, and X. Li. Abnormal surface morphology of the central sulcus in children with attention-deficit/hyperactivity disorder. *Frontiers in Neuroanatomy*, 9(AUGUST):114, aug 2015. ISSN 16625129. doi: 10.3389/FNANA.2015.00114/BIBTEX.

[79] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47 (W1):W199–W205, jul 2019. ISSN 0305-1048. doi: 10.1093/NAR/GKZ401.

[80] J. Lin, R. Tabassum, S. Ripatti, and M. Pirinen. MetaPhat: Detecting and Decomposing Multivariate Associations From Univariate Genome-Wide Association Statistics. *Frontiers in Genetics*, 11:431, may 2020. ISSN 16648021. doi: 10.3389/FGENE.2020.00431/BIBTEX.

[81] T. J. Littlejohns et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications 2020 11:1*, 11(1):1–12, may 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15948-9.

[82] C. P. López. Display Volumes and Specialized Graphics. *MATLAB Graphical Programming*, pages 73–98, 2014. doi: 10.1007/978-1-4842-0316-3_4.

[83] M. Luciano et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature genetics*, 50(1):6–11, jan 2018. ISSN 1546-1718. doi: 10.1038/S41588-017-0013-8.

[84] S. Malik, G. Vinukonda, L. R. Vose, D. Diamond, B. B. Bhimavarapu, F. Hu, M. T. Zia, R. Hevner, N. Zecevic, and P. Ballabh. Neurogenesis continues in the third trimester of pregnancy and is suppressed by premature birth. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(2): 411–423, jan 2013. ISSN 1529-2401. doi: 10.1523/JNEUROSCI.4445-12.2013.

[85] J. J. Maller, R. Anderson, R. H. Thomson, J. V. Rosenfeld, Z. J. Daskalakis, and P. B. Fitzgerald. Occipital bending (Yakovlevian torque) in bipolar depression. *Psychiatry Research: Neuroimaging*, 231(1):8–14, jan 2015. ISSN 0925-4927. doi: 10.1016/J.PSCYCHRESNS.2014.11.008.

[86] M. Marozzi. Some remarks about the number of permutations one should consider to perform a permutation test. *Statistica*, 64(1):193–201, 2004. ISSN 1973-2201. doi: 10.6092/ISSN.1973-2201/32.

[87] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. GREAT improves functional interpretation

of cis-regulatory regions. *Nature Biotechnology 2010 28:5*, 28(5):495–501, may 2010. ISSN 1546-1696. doi: 10.1038/nbt.1630.

[88] Y. Mekki, V. Guillemot, H. Lemaître, A. Carrión-Castillo, S. Forkel, V. Frouin, and C. Philippe. The genetic architecture of language functional connectivity. *NeuroImage*, 249, apr 2022. ISSN 1095-9572. doi: 10.1016/J.NEUROIMAGE.2021.118795.

[89] Z. Molnár et al. New insights into the development of the human cerebral cortex. *Journal of Anatomy*, 235(3):432, 2019. ISSN 14697580. doi: 10.1111/JOA.13055.

[90] V. K. Mootha et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 2003 34:3*, 34(3):267–273, jun 2003. ISSN 1546-1718. doi: 10.1038/ng1180.

[91] M. D. Morgan, E. Pairo-Castineira, K. Rawlik, O. Canela-Xandri, J. Rees, D. Sims, A. Tenesa, and I. J. Jackson. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nature Communications 2018 9:1*, 9(1):1–10, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07691-z.

[92] N. Mullins et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics 2021 53:6*, 53(6):817–829, may 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00857-4.

[93] S. Naqvi et al. Shared heritability of human face and brain shape. *Nature Genetics*, 53(6):830–839, 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00827-w.

[94] S. Neubauer, P. Gunz, N. A. Scott, J. J. Hublin, and P. Mitteroecker. Evolution of brain lateralization: A shared hominid pattern of endocranial asymmetry is much more variable in humans than in great apes. *Science Advances*, 6(7):9935–9949, 2020. ISSN 23752548. doi: 10.1126/SCIADV.AAX9935/SUPPL_FILE/AAX9935_SM.PDF.

[95] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, 2002.

[96] H. L. Nicholls, C. R. John, D. S. Watson, P. B. Munroe, M. R. Barnes, and C. P. Cabrera. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. *Frontiers in Genetics*, 11:350, apr 2020. ISSN 16648021. doi: 10.3389/FGENE.2020.00350/BIBTEX.

[97] S. S. Nishizaki, N. Ng, S. Dong, R. S. Porter, C. Morterud, C. Williams, C. Asman, J. A. Switzenberg, and A. P. Boyle. Predicting the effects of SNPs on transcription factor binding affinity. *Bioinformatics*, 36(2):364–372, jan 2020. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTZ612.

[98] T. J. Nowakowski, A. A. Pollen, C. Sandoval-Espinosa, and A. R. Kriegstein. Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development. *Neuron*, 91(6):1219–1227, sep 2016. ISSN 10974199. doi: 10.1016/J.NEURON.2016.09.005/ATTACHMENT/5FB07326-F8FF-4C17-9374-553AF90F6D57/MMC2.XLSX.

[99] Y. Okada, S. Takeda, Y. Tanaka, J. C. I. Belmonte, and N. Hirokawa. Mechanism of Nodal Flow: A Conserved Symmetry Breaking Event in Left-Right Axis Determination. *Cell*, 121(4):633–644, may 2005. ISSN 0092-8674. doi: 10.1016/J.CELL.2005.04.008.

[100] A. Okbay et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539–542, may 2016. ISSN 1476-4687. doi: 10.1038/NATURE17671.

[101] W. Y. Ong, C. S. Stohler, and D. R. Herr. Role of the Prefrontal Cortex in Pain Processing. *Molecular Neurobiology*, 56(2):1137, feb 2019. ISSN 15591182. doi: 10.1007/S12035-018-1130-9.

[102] M. A. Purnell, P. J. Donoghue, S. E. Gabbott, M. E. McNamara, D. J. Murdock, and R. S. Sansom. Experimental analysis of soft-tissue fossilization: opening the black box. *Palaeontology*, 61(3):317–323, may 2018. ISSN 1475-4983. doi: 10.1111/PALA.12360.

[103] P. Rakic. Radial glial cells: Brain functions. In *Encyclopedia of Neuroscience*, volume R, pages 15–21. Elsevier Ltd, 2009. ISBN 9780080450469. doi: 10.1016/B978-008045046-9.01021-4.

[104] V. Ramensky, P. Bork, and S. Sunyaev. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30(17):3894–3900, sep 2002. ISSN 0305-1048. doi: 10.1093/NAR/GKF493.

[105] S. Rao, Y. Yao, and D. E. Bauer. Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Medicine 2021 13:1*, 13(1):1–20, mar 2021. ISSN 1756-994X. doi: 10.1186/S13073-021-00857-3.

[106] M. Reuter, N. J. Schmansky, H. D. Rosas, and B. Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61 (4):1402, jul 2012. ISSN 10538119. doi: 10.1016/J.NEUROIMAGE.2012.02.084.

[107] M. Ribolsi, Z. J. Daskalakis, A. Siracusano, and G. Koch. Abnormal Asymmetry of Brain Connectivity in Schizophrenia , 2014.

[108] S. Ripke et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature 2014 511:7510*, 511(7510):421–427, jul 2014. ISSN 1476-4687. doi: 10.1038/nature13595.

[109] Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, feb 2015. ISSN 14764687. doi: 10.1038/NATURE14248.

[110] P. Robert and Y. Escoufier. A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Applied Statistics*, 25(3):257, 1976. ISSN 00359254. doi: 10.2307/2347233.

[111] A. Sarica, R. Vasta, F. Novellino, M. G. Vaccaro, A. Cerasa, and A. Quattrone. MRI asymmetry index of hippocampal subfields increases through the continuum from the mild cognitive impairment to the alzheimer's disease. *Frontiers in Neuroscience*, 12(AUG):576, aug 2018. ISSN 1662453X. doi: 10.3389/FNINS.2018.00576/BIBTEX.

[112] T. D. Satterthwaite et al. The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage*, 124:1115–1119, jan 2016. ISSN 1053-8119. doi: 10.1016/J.NEUROIMAGE.2015.03.056.

[113] I. Savic and P. Lindström. PET and MRI show differences in cerebral asymmetry and functional connectivity between homo- and heterosexual subjects. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27):9403–9408, jul 2008. ISSN 1091-6490. doi: 10.1073/PNAS.0801566105.

[114] D. J. Schaid, W. Chen, and N. B. Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics 2018 19:8*, 19(8):491–504, may 2018. ISSN 1471-0064. doi: 10.1038/s41576-018-0016-z.

[115] J. Schmitz, O. Güntürkün, and S. Ocklenburg. Building an asymmetrical brain: The molecular perspective. *Frontiers in Psychology*, 10(APR):982, 2019. ISSN 16641078. doi: 10.3389/FPSYG.2019.00982/BIBTEX.

[116] Z. Sha, D. Schijven, A. Carrion-Castillo, M. Joliot, B. Mazoyer, S. E. Fisher, F. Crivello, and C. Francks. The genetic architecture of structural left–right asymmetry of the human brain. *Nature Human Behaviour 2021 5:9*, 5(9):1226–1239, mar 2021. ISSN 2397-3374. doi: 10.1038/s41562-021-01069-w.

[117] J. Sheng, L. Wang, H. Cheng, Q. Zhang, R. Zhou, and Y. Shi. Strategies for multivariate analyses of imaging genetics study in Alzheimer's disease. *Neuroscience Letters*, 762:136147, sep 2021. ISSN 0304-3940. doi: 10.1016/J.NEULET.2021.136147.

[118] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. ISSN 01628828. doi: 10.1109/34.868688.

[119] S. Sniekers et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature genetics*, 49(7):1107–1112, jul 2017. ISSN 1546-1718. doi: 10.1038/NG.3869.

[120] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3(3):583–617, mar 2003. ISSN 15324435. doi: 10.1162/153244303321897735.

[121] A. Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, oct 2005. ISSN 00278424. doi: 10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.

[122] D. Szklarczyk et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, jan 2021. ISSN 1362-4962. doi: 10.1093/NAR/GKAA1074.

[123] V. Tam, N. Patel, M. Turcotte, Y. Bossé, G. Paré, and D. Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics 2019 20:8*, 20(8):467–484, may 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0127-1.

[124] M. D. Teare, A. M. Dunning, F. Durocher, G. Rennart, and D. F. Easton. Sampling distribution of summary linkage disequilibrium measures. *Annals of human genetics*, 66(Pt 3):223–233, 2002. ISSN 0003-4800. doi: 10.1017/S0003480002001082.

[125] L. N. Telano and S. Baker. Physiology, Cerebral Spinal Fluid. *StatPearls*, jul 2021.

[126] G. Trynka, C. Sandor, B. Han, H. Xu, B. E. Stranger, X. S. Liu, and S. Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, 45(2):124–130, feb 2013. ISSN 1546-1718. doi: 10.1038/NG.2504.

[127] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1:59, 2021. ISSN 2662-8449. doi: 10.1038/s43586-021-00056-9.

[128] T. Usui, M. R. Macleod, S. K. McCann, A. M. Senior, and S. Nakagawa. Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. *PLOS Biology*, 19(5):e3001009, may 2021. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.3001009.

[129] S. Van Dongen, G. Molenberghs, and E. Matthysen. The statistical analysis of fluctuating asymmetry: REML estimation of a mixed regression model. *Journal of Evolutionary Biology*, 12(1):94–102, jan 1999. ISSN 1010061X. doi: 10.1046/J.1420-9101.1999.00012.X.

[130] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80:62–79, oct 2013. ISSN 1095-9572. doi: 10.1016/J.NEUROIMAGE.2013.05.041.

[131] A. Vanbiervliet, V. Lemonidis, and P. Claes. Mapping shape asymmetry of the human cerebrum. Master's thesis, KU Leuven, 2022.

[132] G. Vingerhoets, R. Gerrits, and H. Verhelst. Atypical Brain Asymmetry in Human Situs Inversus: Gut Feeling or Real Evidence? *Symmetry 2021, Vol. 13, Page 695*, 13(4):695, apr 2021. ISSN 2073-8994. doi: 10.3390/SYM13040695.

[133] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 90(1):7–24, jan 2012. ISSN 0002-9297. doi: 10.1016/J.AJHG.2011.11.029.

[134] G. Vogt. Disentangling the environmentally induced and stochastic developmental components of phenotypic variation. *Phenotypic Switching: Implications in Biology and Medicine*, pages 207–251, jan 2020. doi: 10.1016/B978-0-12-817996-3.00010-4.

[135] N. D. Volkow et al. The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, 32:4–7, aug 2018. ISSN 1878-9293. doi: 10.1016/J.DCN.2017.10.002.

[136] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing 2007 17:4*, 17(4):395–416, aug 2007. ISSN 1573-1375. doi: 10.1007/S11222-007-9033-Z.

[137] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, sep 2010. ISSN 0305-1048. doi: 10.1093/NAR/GKQ603.

[138] L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1):1–8, jul 2011. ISSN 0888-7543. doi: 10.1016/J.YGENO.2011.04.006.

[139] K. Watanabe, E. Taskesen, A. Van Bochoven, and D. Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications 2017 8:1*, 8(1):1–11, nov 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01261-5.

[140] T. White, N. C. Andreasen, and P. Nopoulos. Brain Volumes and Surface Morphology in Monozygotic Twins. *Cerebral Cortex*, 12(5):486–493, may 2002. ISSN 1047-3211. doi: 10.1093/CERCOR/12.5.486.

[141] M. M. J. Wittens et al. Inter- and Intra-Scanner Variability of Automated Brain Volumetry on Three Magnetic Resonance Imaging Systems in Alzheimer's Disease and Controls. *Frontiers in Aging Neuroscience*, 13:641, oct 2021. ISSN 16634365. doi: 10.3389/FNAGI.2021.746982/BIBTEX.

[142] Z. Yan et al. Integrating RNA-Seq with GWAS reveals novel insights into the molecular mechanism underpinning ketosis in cattle. *BMC Genomics*, 21(1): 1–12, jul 2020. ISSN 14712164. doi: 10.1186/S12864-020-06909-Z/FIGURES/5.

[143] J. Yang et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics 2010 42:7*, 42(7):565–569, jun 2010. ISSN 1546-1718. doi: 10.1038/ng.608.

[144] L. Zhao, W. Matloff, Y. Shi, R. P. Cabeen, and A. W. Toga. Mapping Complex Brain Torque Components and Their Genetic Architecture and Phenomic Associations in 24,112 Individuals. *Biological Psychiatry*, 91(8):753–768, apr 2022. ISSN 0006-3223. doi: 10.1016/J.BIOPSYCH.2021.11.002.

[145] Y. Zhou et al. Integrating RNA-Seq With GWAS Reveals a Novel SNP in Immune-Related HLA-DQB1 Gene Associated With Occupational Pulmonary Fibrosis Risk: A Multi-Stage Study. *Frontiers in Immunology*, 12:5822, jan 2022. ISSN 16643224. doi: 10.3389/FIMMU.2021.796932/BIBTEX.

[146] Z. Zuo, S. Ran, Y. Wang, C. Li, Q. Han, Q. Tang, W. Qu, and H. Li. Asymmetry in cortical thickness and subcortical volume in treatment-naïve major depressive disorder. *NeuroImage: Clinical*, 21:101614, jan 2019. ISSN 2213-1582. doi: 10.1016/J.NICL.2018.101614.