UNIVERSITY OF CALIFORNIA, SAN DIEGO

Genotype-Phenotype Mapping of the Human Brain in the Era of Large-Scale Genomics Databases

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive Science

by

Chun Chieh Fan

Committee in charge:

Professor Terry L Jernigan, Chair
Professor Anders M. Dale, Co-Chair
Professor Eran A. Mukamel
Professor Nicholas J. Schork
Professor Armin Schwartzman
Professor Zhuowen Tu

2017

The dissertation of Chun Chieh Fan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2017

DEDICATION

To Chi-Yu, dearest of my heart, for her steadyfast support. And to Alfie, for his incurable appetite for treats and plays that enriched my years of study.

EPIGRAPH

I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists.

— G. H. Hardy

For such a model there is no need to ask the question "Is the model true?".
If "truth" is to be the "whole truth" the answer must be "No".
The only question of interest is "Is the model illuminating and useful?".

— G. E. P. Box

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in *Plos Medicine* 2017. Rahul S. Desikan*, Chun Chieh Fan*, Yunpeng Wang, Andrew J. Schork, Howard J. Cabral, L. Adrienne Cupples, Wesley K. Thompson, Lilah Besser, Walter A. Kukull, Dominic Holland, Chi-Hua Chen, James B. Brewer, David S. Karow, Karolina Kauppi, Aree Witoelar, Celeste M. Karch, Luke W. Bonham, Jennifer S. Yokoyama, Howard J. Rosen, Bruce L. Miller, William P. Dillon, David M. Wilson, Christopher P. Hess, Margaret Pericak-Vance, Jonathan L. Haines, Lindsay A. Farrer, Richard Mayeux, John Hardy, Alison M. Goate, Bradley T. Hyman, Gerard D. Schellenberg, Linda K. McEvoy, Ole A. Andreassen, Anders M. Dale. PLoS, 2017. The dissertation author was the primary investigator and co-first author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in *Current Biology* 2015. Chun Chieh Fan, Hauke Bartsch, Andrew Schork, Chi-Hua Chen, Yunpeng Wang, Min-Tzu Lo, Timothy T. Brown, Joshua M. Kuperman, Donald J. Hagler Jr., Nicholas Schork, Terry L. Jernigan, Anders M. Dale. Cell Press, 2015. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in *NeuroImage:Clinical* 2017. Chun Chieh Fan, Timothy T. Brown, Hauke Bartsch, Joshua M. Kuperman, Donald J. Hagler Jr., Andrew Schork, Yvonne Searcy, Ursula Bellugi, Eric Halgren, Anders M. Dale. Elsevier, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is being prepared for submission for publication. Chun Chieh Fan, Andrew J. Schork, Timothy T. Brown, Barbara E. Spencer, Natacha Akshoomoff, Chi-Hua Chen, Joshua M. Kuperman, Donald J. Hagler Jr., Asta Kristine Hberg, Thomas Espeseth, Ole A. Andreassen, Anders M. Dale, Terry L. Jernigan, Eric Halgren. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is being prepared for submission for publication. Chun Chieh Fan, Andrew J. Schork, Westly K. Thompson, Asta Kristine Hberg, Thomas Espeseth, Ole A. Andreassen, Anders M. Dale, Terry L. Jernigan, Nicholas J. Schork. The dissertation author was the primary investigator and author of this paper.

VITA

2017	Doctor of Philosophy in Cognitive Science University of California, San Diego
2013-2017	Graduate Student Researcher Teaching Asistant Cognitive Science Department University of California, San Diego
2009-2013	Attending Physician in Psychiatry Ju-Shan Hospital, Taiwan
2009-2011	Master of Science in Epidemiology National Taiwan University, Taiwan
2005-2009	Residency in Psychiatry Taipei City Hospital, Taiwan
1997-2004	Medical Doctor National Yang-Ming University, Taiwan

PUBLICATIONS

Chun Chieh Fan, Timothy T Brown, Hauke Bartsch, Joshua M Kuperman, Donald J Hagler, Andrew Schork, Yvonne Searcy, Ursula Bellugi, Eric Halgren, Anders M Dale. "Williams syndrome-specific neuroanatomical profile and its associations with behavioral features", *NeuroImage:Clinical*, 15, 2017

Olav B Smeland, Oleksandr Frei, Karolina Kauppi, W David Hill, Wen Li, Yunpeng Wang, Florian Krull, Francesco Bettella, Jon A Eriksen, Aree Witoelar, Gail Davies, **Chun Chieh Fan**, Wesley K Thompson, Max Lam, Todd Lencz, Chi-Hua Chen, Torill Ueland, Erik G Jnsson, Srdjan Djurovic, Ian J Deary, Anders M Dale, Ole A Andreassen. "Identification of Genetic Loci Jointly Influencing Schizophrenia Risk and the Cognitive Traits of Verbal-Numerical Reasoning, Reaction Time, and General Cognitive Function", *JAMA Psychiatry*, 2017

A Devor, OA Andreassen, Y Wang, T Mki-Marttunen, OB Smeland, **Chun Chieh Fan**, AJ Schork, D Holland, WK Thompson, A Witoelar, CH Chen, RS Desikan, LK McEvoy, S Djurovic, P Greengard, P Svenningsson, GT Einevoll, AM Dale. "Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia", *Molecular Psychiatry*, 22, 6, 2017

Tan, Chin Hong; Sugrue, Leo; Broce, Iris; Tong, Elizabeth; Tan, Jacinth; Hess, Christopher ; Dillon, William; Bonham, Luke; Yokoyama, Jennifer; Rabinovici, Gil Dan; Rosen, Howard; Miller, Bruce; Hyman, Bradley T; Schellenberg, Gerard; Besser, Lilah; Kukull, Walter; Karch, Celeste; Brewer, James; Kauppi, Karolina; McEvoy, Linda; Andreassen, Ole; Dale, Anders; **Fan, Chun Chieh***; Desikan, Rahul*. "Polygenic hazard scores in preclinical Alzheimers disease", *Annals of Neurology*, 2017. *co-senior authors

Jennifer S Yokoyama, Celeste M Karch, **Chun Chieh Fan**, Luke W Bonham, Naomi Kouri, Owen A Ross, Rosa Rademakers, Jungsu Kim, Yunpeng Wang, Gnter U Hglinger, Ulrich Mller, Raffaele Ferrari, John Hardy, Parastoo Momeni, Leo P Sugrue, Christopher P Hess, A James Barkovich, Adam L Boxer, William W Seeley, Gil D Rabinovici, Howard J Rosen, Bruce L Miller, Nicholas J Schmansky, Bruce Fischl, Bradley T Hyman, Dennis W Dickson, Gerard D Schellenberg, Ole A Andreassen, Anders M Dale, Rahul S Desikan, International FTD-Genomics Consortium. "Shared genetic risk between corticobasal degeneration, progressive supranuclear palsy, and frontotemporal dementia", *Acta neuropathologica*, 133, 2017

Rahul S Desikan*, **Chun Chieh Fan***, Yunpeng Wang, Andrew J Schork, Howard J Cabral, L Adrienne Cupples, Wesley K Thompson, Lilah Besser, Walter A Kukull, Dominic Holland, Chi-Hua Chen, James B Brewer, David S Karow, Karolina Kauppi, Aree Witoelar, Celeste M Karch, Luke W Bonham, Jennifer S Yokoyama, Howard J Rosen, Bruce L Miller, William P Dillon, David M Wilson, Christopher P Hess, Margaret Pericak-Vance, Jonathan L Haines, Lindsay A Farrer, Richard Mayeux, John Hardy, Alison M Goate, Bradley T Hyman, Gerard D Schellenberg, Linda K McEvoy, Ole A Andreassen, Anders M Dale. "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score", *PLoS medicine*, 14, 2017. *equal contribution

Raffaele Ferrari, Yunpeng Wang, Jana Vandrovcova, Sebastian Guelfi, Aree Witeolar, Celeste M Karch, Andrew J Schork, **Chun Chieh Fan**, James B Brewer, Parastoo Momeni, Gerard D Schellenberg, William P Dillon, Leo P Sugrue, Christopher P Hess, Jennifer S Yokoyama, Luke W Bonham, Gil D Rabinovici, Bruce L Miller, Ole A Andreassen, Anders M Dale, John Hardy, Rahul S Desikan, International FTD-Genomics Consortium, International Parkinson's Disease Genomics Consortium. "Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases", *J Neurol Neurosurg Psychiatry*, 88, 2017.

Min-Tzu Lo, David A Hinds, Joyce Y Tung, Carol Franz, **Chun Chieh Fan**, Yunpeng Wang, Olav B Smeland, Andrew Schork, Dominic Holland, Karolina Kauppi, Nilotpal Sanyal, Valentina Escott-Price, Daniel J Smith, Michael O'Donovan, Hreinn Stefansson, Gyda Bjornsdottir, Thorgeir E Thorgeirsson, Kari Stefansson, Linda K McEvoy, Anders M Dale, Ole A Andreassen, Chi-Hua Chen. "Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders", *Nature Genetics*, 49, 2017.

Luke W Bonham, Ethan G Geier, **Chun Chieh Fan**, Josiah K Leong, Lilah Besser, Walter A Kukull, John Kornak, Ole A Andreassen, Gerard D Schellenberg, Howard J Rosen, William P Dillon, Christopher P Hess, Bruce L Miller, Anders M Dale, Rahul S Desikan, Jennifer S Yokoyama. "Agedependent effects of APOE 4 in preclinical Alzheimer's disease", *Annals of clinical and translational neurology*, 3, 2016.

J Yokoyama, **Chun Chieh Fan**, Y Wang, N Kouri, R Ferrari, O Andreassen, J Hardy, A Boxer, B Miller, G Schellenberg, D Dickson, A Dale, R Desikan. "Genetic overlap between 4-repeat tauopathies suggests a role for development in the pathobiology of corticobasal degeneration", *Journal of Neurochemistry*, 138, 2016.

YuJen Chen, YuChun Lo, YungChin Hsu, **ChunChieh Fan**, TzungJeng Hwang, ChihMin Liu, YiLing Chien, Ming H Hsieh, ChenChung Liu, HaiGwo Hwu, WenYih Isaac Tseng. "Automatic whole brain tractbased analysis using predefined tracts in a diffusion spectrum imaging template and an accurate registration strategy", *Human brain mapping*, 36, 2015.

Chun Chieh Fan, Hauke Bartsch, Andrew J Schork, Chi-Hua Chen, Yunpeng Wang, Min-Tzu Lo, Timothy T Brown, Joshua M Kuperman, Donald J Hagler, Nicholas J Schork, Terry L Jernigan, Anders M Dale. "Modeling the 3D geometry of the cortical surface with genetic ancestry", *Current Biology*, 25, 2015.

Chi-Hua Chen, Qian Peng, Andrew J Schork, Min-Tzu Lo, **Chun Chieh Fan**, Yunpeng Wang, Rahul S Desikan, Francesco Bettella, Donald J Hagler, Lars T Westlye, William S Kremen, Terry L Jernigan, Stephanie Le Hellard, Vidar M Steen, Thomas Espeseth, Matt Huentelman, Asta K Hberg, Ingrid Agartz, Srdjan Djurovic, Ole A Andreassen, Nicholas Schork, Anders M Dale. "Large-scale genomics unveil polygenic architecture of human cortical surface area", *Nature Communication*, 6, 2015.

Chien-Hsiun Chen, Chau-Shoun Lee, Ming-Ta Michael Lee, Wen-Chen Ouyang, Chiao-Chicy Chen, Mian-Yoon Chong, Jer-Yuarn Wu, Happy Kuy-Lok Tan, Yi-Ching Lee, Liang-Jen Chuo, Nan-Ying Chiu, Hin-Yeung Tsang, Ta-Jen Chang, For-Wey Lung, Chen-Huan Chiu, Cheng-Ho Chang, Ying-Sheue Chen, Yuh-Ming Hou, Cheng-Chung Chen, Te-Jen Lai, Chun-Liang Tung, Chung-Ying Chen, Hsien-Yuan Lane, Tung-Ping Su, Jung Feng, Jin-Jia Lin, Ching-Jui Chang, Po-Ren Teng, Chia-Yih Liu, Chih-Ken Chen, I-Chao Liu, Jiahn-Jyh Chen, Ti Lu, **Chun-Chieh Fan**, Ching-Kuan Wu, Chang-Fang Li, Kathy Hsiao-Tsz Wang, Lawrence Shih-Hsin Wu, Hsin-Ling Peng, Chun-Ping Chang, Liang-Suei Lu, Yuan-Tsong Chen, Andrew Tai-Ann Cheng. "Variant GADL1 and response to lithium therapy in bipolar I disorder", *New England Journal of Medicine*, 370, 2014.

ABSTRACT OF THE DISSERTATION

Genotype-Phenotype Mapping of the Human Brain in the Era of Large-Scale Genomics Databases

by

Chun Chieh Fan

Doctor of Philosophy in Cognitive Science

University of California, San Diego, 2017

Professor Terry L Jernigan, Chair
Professor Anders M. Dale, Co-Chair

Searching the genetic basis of human complex traits is an essential tool to illuminating biological processes and predicting risks of diseases. However, with the advance of genotyping technologies, e.g. genome sequencing, and sophistication of phenotypic measurements, e.g. magnetic resonance imaging, two prong challenges have imposed on the endeavor for mapping genotypes to phenotypes. First is the weak genetic signals due to the multifactorial contributions from common genetic variants. The effect sizes of those genetic variants become too small to be detected. The second is the inconsistency

of the phenotypic measurements, which the genetically fundamental units, or so called endophenotype, are not apparent. This polygenes-polymeasurements problem become even more prominent in recent surge of large-scale cross-traits genomic studies. This dissertation is a collection of my series studies to develop novel methods to tackle the polygenes-polymeasurement challenges. In particular, I focused on how to extract generalizable signals from diverse genomic databses. The extracted signals can be useful in either predicting disease risks or elucidating biological processes in human brain.

Chapter 1

Overview

The genetic basis of dynamic biological processes that shape human complex traits is the fundamental building block for each persons individuality. Complex traits, such as intelligence and personality, have been found to vary closely with genetics even before the discovery of double-strand DNA [1]. Efforts to find the associations between DNA and human traits, i.e. to map genotypes to phenotypes, have led to important discoveries about development, aging, and etiologies of diseases [2]. With the advance of technology, now the mapping has reached down to the single-base pair level in the human genome, in hopes of pinpointing which single-base can lead to variation in a human trait [3].

However, as the resolution of measurements of genetic variation has become finer, the challenges of genotype-phenotype mapping loom larger. The common traits in human populations, such as intelligence, personality, and mental disorders, have substantial amount of variations attributable to common genetic variants scattering across genome [3]. There can be thousands of genetic variants jointly shaped individuals traits through different biological mechanisms [4]. The contribution of each single-base pair, thus, would be tiny. How to identify the associations from those scattered weak signals has

become a steep hurdle for understanding human complex traits. Increasing the resolution of measurements of genetic variation only provides better chance to localize the causal genetic variants rather than better power to detect weak signals.

Moreover, the inconsistency of measurements of some human traits stacks additional difficulties on genotype-phenotype mapping. For instance, many psychometric tools are used to measure human intelligence, each targeting different cognitive domains, such as working memory, executive function, and verbal fluency. Although evidence suggests certain combinations of scores exhibit high heritability, i.e., have large amount of shared variations with genetics [5], the definition of those cognitive domains is solely based on the observable measurements across subjects. A complex trait with high heritability is not necessarily mean the measured variations are closer to the biological processes. The genetically fundamental unit of a complex phenotype, or so-called endophenotype, is an internal subunit of a complex trait that specifically shaped by a set of genes [6]. For understanding how genetic perturbation cascading into the complex phenotype, what we need are the endophenotype rather than the apparent phenotype. Yet it is unclear how to extract genetically relevant subunit from multiple measurements, all intending to measure one internal phenotype, such as a domain of intelligence. This makes researchers used a score based on observable variations from multiple measurements, which oftentimes makes the score inconsistent with the underlying true genetic component [7]. This problem is particularly evident in the brain-imaging field where measurements from magnetic resonance imaging can be millions of sampled points across brain with diverse metrics [8, 9, 10, 11, 12, 13, 14]. All different metrics across brain may actually measure the same process where there are a more parsimonious model can characterize it. Here I termed this phenomonen as ”poly-measurement” for human complex trait.

Current approaches to the challenges of polygenes and poly-measurements often

resort to brute force methods, by obtaining very large sample sizes for mapping [14]. The unintended consequence of large-scale studies, i.e. big data, is that the mapping becomes a fishing expedition, yielding non-replicable associations and drawing criticism for identifying weak unimportant genetic effects [4]. As George Box famously said, all models are wrong but some are useful; the essence of the challenge in modern genotype-phenotype mapping with large-scale data is to find generalizable signals that can either be used for predicting individuals phenotypic outcome or illuminate the biological processes.

Under this premise, this dissertation is a collection of my series of inquiries navigating the space between polygenes and poly-measurements for understanding the genetic basis of human brain phenotypes. The polygenic architecture of variation in human brains observed with high-dimensional measurements mandates novel analytic strategies to identify a generalizable signal that can either illuminate the underlying biological process or predict the eventual outcome of a given trait.

1.1 Polygenic signals for generalizable predictions

Chapter 2 of this dissertation is a project focusing on how to extract polygenic signals for predicting the outcome of a given trait. The polygenic signals in this context are the sum of small genetic perturbations that associate with a given trait, namely Alzheimers disease (AD). Previous approaches using polygenic signals treated the outcome as a static state, e.g. either you have the disease or not. Yet constant and stable conditions are the exception rather than a common rule among human complex traits. Since human development and aging are dynamic and time-dependent processes, diseases resulting from disrupting these processes would also have strong time-dependent characteristics. For example, Alzheimers disease (AD) has increasing incidence among older individuals, and APOE risk variants, genes with large effect on AD, shift age of onset

earlier [15]. Current approaches to defining a generalizable polygenic signal ignore the strong time-dependent properties of human traits, treating age as a nuisance parameter in the models [16, 17, 18]. The project described in Chapter 2 tackled this missing link by explicitly taking time into consideration when deriving a generalizable polygenic signal.

1.2 Defining a new phenotype based on poly-measurements

While chapter 2 focuses on the polygenic aspect of genotype-phenotype mapping, chapter 3 begins to tackle the challenges imposed by poly-measurements. Brain imaging measures structural or functional variation in the brain at millions of sampled points in magnetic resonance images (MRI) [8, 9]. Currently, most genotype-phenotype mapping with brain imaging uses the derived measures that group sets of sampled points into discretized regions, such as hippocampus, and then do the associations with these regions [13, 14]. As previous imaging studies based on co-inheritance among twins, the landmark-defined regions are not necessarily endophenotypes [19, 10]. For a set of genes involving in one particular molecular pathway can has its influence on several discontiguous brain regions, resulting a covariance structure among measurements. Pure data-driven machine learning approach is unlikely to discover this genetically driven covariance structure neither. The algorithms have to search through all possible combinations across MRI measurements, which are almost inexhaustible if we do not impose constraints [20, 21]. In my dissertation, I used our knowledge about genetics to redefine poly-measurement phenotypes that closely aligned with the genetic basis. In chapter 3, I used genetic background, i.e. genetic ancestry, to define brain morphological indices representing geometrical variation of the human cortical surface. Chapter 4 utilized a naturally occurring large genetic perturbation to derive a neuroanatomical score from

multiple imaging measures, representing the holistic morphological difference between individuals who had the large genetic perturbation and those who did not. Chapter 5 critically examined the generalizability of the neuroanatomical score. In particular, we investigated whether the combined poly-measurements score based on the large-genetic perturbation can enhance the signal for genotype-phenotype mapping among normally developing individuals, identifying genetic variants relevant to the neurodevelopment.

1.3 A joint map between polygenic signals and poly-measurement phenotypes

In the final chapter, I approached genotype-phenotype mapping from both ends to see if we can simultaneously find links between polygenes and poly-measurements. In this case, we inherit a double-curse of dimensionality. Computational resources and sample size required both have great impact on joint mapping attempts. Previous studies used co-heritability among twins as proxy measures for the genetic basis [19, 10, 12]. Although those studies provide great insight into the global genetic influence on the brain, they nevertheless provide no direct link to the genetic variations at the base-pair level in the human genome. High genetic correlations between two measurements from twin studies mean that those two measurements are, on average, genetically similar. It does not necessarily mean they both strongly associate with one molecular process driven by a small set of genes, making the contribution of each gene large and easier to be detected through associations. In chapter 6, we came up with a simple and intuitive approach to reveal the genetic topology of the human cortical surface directly from association signals with base-pair level variants of the human genome. This method is a preliminary demonstration that the joint polygenes-poly-measurements mapping can be achieved without the unquenchable need for large sample sizes.

Chapter 2

Genetic Assessment of Age-associated Alzheimers Disease Risk: Development and Validation of a Polygenic Hazard Score

2.1 Introduction

Late onset Alzheimers disease (AD), the most common form of dementia, places a large emotional and economic burden on patients and society. With increasing health care expenditures among cognitively impaired elderly [22], identifying individuals at risk for developing AD is of utmost importance for potential preventative and therapeutic strategies. Inheritance of the 4 allele of apolipoprotein E (APOE) on chromosome 19q13 is the most significant risk factor for developing late-onset AD [15]. APOE 4 has a dose dependent effect on age of onset, increases AD risk three-fold in heterozygotes and fifteen-fold in homozygotes, and is implicated in 20-25% of patients with AD [23].

In addition to APOE, recent genome-wide association studies (GWAS) have identified numerous AD associated single nucleotide polymorphisms (SNPs), most of which have a small effect on disease risk [24, 25]. Although no single polymorphism may be informative clinically, a combination of APOE and non-APOE SNPs may help identify older individuals at increased risk for AD. Despite the detection of novel AD associated genes, GWAS findings have not yet been incorporated into a genetic epidemiology framework for individualized risk prediction.

Building on a prior approach evaluating GWAS-detected genetic variants for disease prediction [26] and using a survival analysis framework, we tested the feasibility of combining AD associated SNPs and APOE status into a continuous measure polygenic hazard score (PHS) for predicting the age-specific risk for developing AD. We assessed replication of the PHS using several independent cohorts.

2.2 Methods

2.2.1 Participants

IGAP: To select AD associated SNPs, we evaluated publicly available AD GWAS summary statistic data (p-values and odds ratios) from the International Genomics of Alzheimers Disease Project . We used IGAP Stage 1 data, consisting of 17,008 AD cases and 37,154 controls, for selecting AD associated SNPs [24].

ADGC: To develop the survival model for the polygenic hazard scores (PHS), we first evaluated age of onset and raw genotype data from 6,409 patients with clinically diagnosed AD and 9,386 cognitively normal older individuals provided by the Alzheimers Disease Genetics Consortium (ADGC, Phase 1, a subset of the IGAP dataset), excluding individuals from the National Institute of Aging Alzheimers Disease Center (NIA ADC) samples and Alzheimers Disease Neuroimaging Initiative (ADNI). To evaluate replication

of PHS, we used an independent sample of 6,984 AD patients and 10,972 cognitively normal older individuals from the ADGC Phase 2 cohort. A detailed description of the genotype and phenotype data within the ADGC datasets has been described in detail elsewhere [26]. Briefly, the ADGC Phase 1 and 2 datasets consist of multi-center, case-control, prospective, and family-based sub-studies of Caucasian participants with AD occurrence after age 60. Participants with autosomal dominant (APP, PSEN1 and PSEN2) mutations were excluded. All participants were genotyped using commercially available high-density SNP microarrays from Illumina or Affymetrix. Clinical diagnosis of AD within the ADGC sub-studies was established using NINCDS/ADRDA criteria for definite, probable or possible AD [27]. For most participants, age of AD onset was obtained from medical records and defined as the age when AD symptoms manifested, as reported by the participant or an informant. For participants lacking age of onset, age at ascertainment was used. Patients with an age-at-onset or age-at-death less than 60 years, and Caucasians of European ancestry were excluded from the analyses. All ADGC Phase 1 and 2 control participants were defined within individual sub-studies as cognitively normal older adults at time of clinical assessment. The institutional review boards of all participating institutions approved the procedures for all ADGC sub-studies. Written informed consent was obtained from all participants or surrogates.

NIA ADC: To assess longitudinal prediction, we evaluated an ADGC-independent sample of 2,724 cognitively normal elderly individuals with at least 2 years of longitudinal clinical follow-up derived from the NIA funded ADCs (data collection coordinated by the National Alzheimers Coordinating Center) [28]. Specifically, we focused on older individuals defined at baseline as having an overall Clinical Dementia Rating (CDR) of 0.0. To assess the relationship between polygenic risk and neuropathology, we assessed 2,960 participants from the NIA ADC samples with genotype and neuropathological evaluations. For the neuropathological variables, we examined the Braak stage for neu-

rofibrillary tangles (NFTs) (0: none; I-II: entorhinal; III-IV: limbic, and V-VI: isocortical) [29] and the Consortium to Establish a Registry for Alzheimers Disease (CERAD) score for neuritic plaques (none/sparse, moderate, or frequent) [30]. Finally, as an additional independent replication sample, we evaluated all NACC AD cases with genetic data who were classified at autopsy as having a High level of AD neuropathologuc change (n = 361), based on the revised NIA-AA AD neuropathology criteria [31]. The institutional review boards of all participating institutions approved the procedures for all NIA ADC sub-studies. Written informed consent was obtained from all participants or surrogates.

ADNI: To assess the relationship between polygenic risk and in vivo biomarkers, we evaluated an ADGC-independent sample of 692 older controls, mild cognitive impairment and AD participants from the ADNI. On a subset of ADNI1 participants with available genotype data, we evaluated baseline CSF levels of A1-42 and total tau, as well as longitudinal clinical dementia rating-sum of box (CDR-SB) scores. In ADNI1 participants with available genotype and quality-assured baseline and follow-up MRI scans, we also assessed longitudinal sub-regional change in medial temporal lobe volume (atrophy) on 2471 serial T1-weighted MRI scans.

2.2.2 Statistical Analysis

We followed three steps to derive the polygenic hazard scores (PHS) for predicting AD age of onset: 1) we defined the set of associated SNPs, 2) we estimated hazard ratios for polygenic profiles, and 3) we calculated individualized absolute hazards.

Using the IGAP Stage 1 sample, we first identified a list of SNPs associated with increased risk for AD, using a significance threshold of p ¡ 10-5. Next, we evaluated all IGAP-detected, AD-associated SNPs within the ADGC Phase 1 case-control dataset. Using a stepwise procedure in survival analysis, we delineated the final list of SNPs for constructing the polygenic hazard score [32, 17]. Specifically, using Cox proportional

hazard models, we identified the top AD-associated SNPs within the ADGC Phase 1 cohort (excluding NIA ADC and ADNI samples), while controlling for the effects of gender, APOE variants, and top five genetic principal components (to control for the effects of population stratification). We utilized age of AD onset and age of last clinical visit to estimate age appropriate hazards [33] and derived a PHS for each participant. In each step of the stepwise procedure, the algorithm selected one SNP from the pool that most improved model prediction (i.e. minimizing the Martingale residuals); additional SNP inclusion that did not further minimize the residuals resulted in halting of the SNP selection process. To prevent over-fitting in this training step, we used 1000x bootstrapping for model averaging and estimating the hazard ratios for each selected SNPs. We assessed the proportional hazard assumption in the final model using graphical comparisons.

To assess for replication, we first examined whether the ADGC Phase 1 derived predicted PHSs could stratify individuals into different risk strata within the ADGC Phase 2 cohort. We next evaluated the relationship between predicted age of AD onset and the empirical/actual age of AD onset using cases from ADGC Phase 2. We binned risk strata into percentile bins and calculated the mean of actual age in that percentile as the empirical age of AD onset. In a similar fashion, we additionally tested replication within the NACC subset classified at autopsy as having a high level of AD neuropathologic change [31].

Because case-control samples cannot provide the proper baseline hazard [34], we used the previously reported annualized incidence rates by age, estimated from the general United States of America (US) population [35]. For each participant, by combining the overall population-derived incidence rates and genotype-derived PHS, we calculated an individuals instantaneous risk for developing AD, based on their genotype and age. To independently assess the predicted instantaneous risk, we evaluated longitudinal follow-

up data from 2,724 cognitively normal older individuals from the NIA ADC with at least 2 years of clinical follow-up. We assessed the number of cognitively normal individuals progressing to AD as a function of the predicted PHS risk strata and examined whether the predicted PHS-derived incidence rate reflects the empirical/actual progression rate using a Cochran-Armitage trend test.

We examined the association between our PHS and established in vivo and pathologic markers of AD neurodegeneration. Using linear models, we assessed whether the PHS associated with Braak stage for NFTs and CERAD score for neuritic plaques as well as CSF A1-42, and CSF total tau. Using linear mixed effects models, we also investigated whether the PHS was associated with longitudinal CDR-SB score and volume loss within the entorhinal cortex and hippocampus. In all analyses, we co-varied for the effects of age and sex.

2.3 Results

From the IGAP cohort, we found 1854 SNPs associated with increased risk for AD at a p ¡ 10^{-5}. Of these, using the Cox stepwise regression framework, we identified 31 SNPs, in addition to two APOE variants, within the ADGC cohort for constructing the polygenic model (Table 2.1). Figure 2.1 illustrates the relative risk for developing AD using the ADGC case/control Phase 1 cohort. The graphical comparisons among Kaplan-Meier estimations and Cox proportional hazard models indicate the proportional hazard assumption holds for the final model (Figure 2.1).

To quantify the additional prediction provided by polygenic information beyond APOE, we evaluated how PHS modulates age of AD onset in APOE 3/3 individuals. Among these individuals, we found that age of AD onset can vary by more than 10 years, depending on polygenic risk. For example, for an APOE 3/3 individual in the 10th decile

(top 10%) of PHS, at 50% risk for meeting clinical criteria for AD diagnosis, the expected age for developing AD is approximately 84 years (Figure 2.2); however, for an APOE 3/3 individual in the 1st decile (bottom 10%) of PHS, the expected age of developing AD is approximately 95 years (Figure 2.2). The hazard ratio of 10th decile to 1st decile is 3.34 (95% CI: 2.62 - 4.24, logrank test: p = 1 x 10^{-22}).

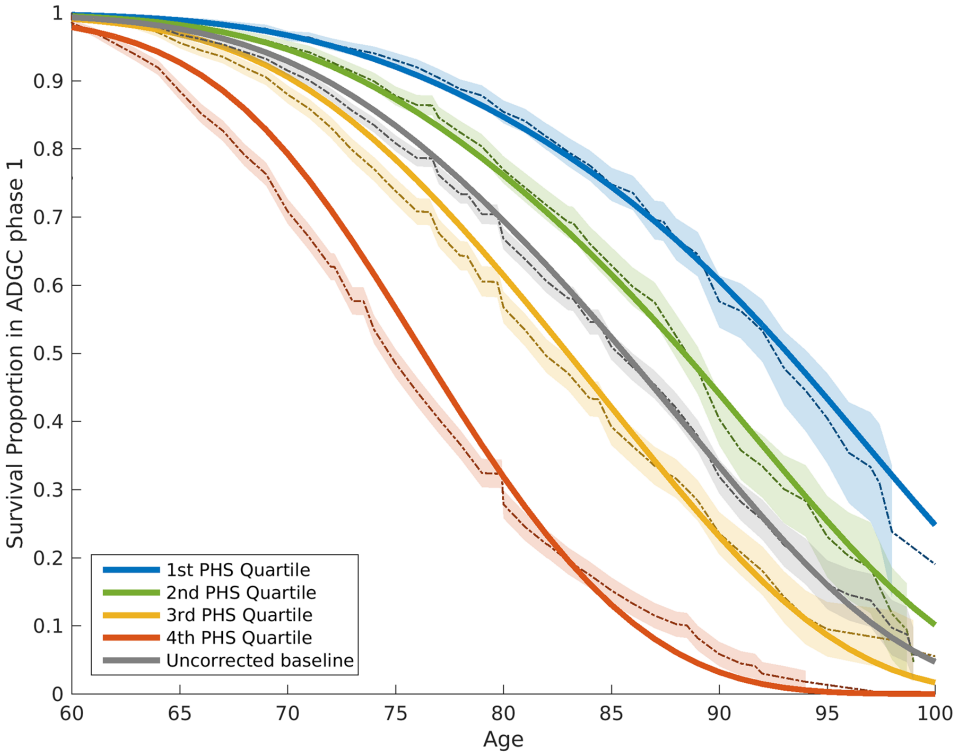

Figure 2.1: Survival analysis on ADGC phase 1 dataset. Kaplan-Meier estimates and Cox proportional model fits from the case-control ADGC phase 1 dataset, excluding NACC and ADNI samples. The proportional hazard assumptions were checked based on the graphical comparisons between Kaplan-Meier estimation (dashed line) and Cox proportional hazard models (solid line). 95% confidence intervals of Kaplan-Meier estimation are also demonstrated (shaded with corresponding colors). The baseline hazard (gray line) in this model is based on the mean of ADGC data.

To assess replication, we applied the ADGC Phase 1-trained model on independent samples from ADGC Phase 2. Using the empirical distributions, we found that the

Table 2.1: Information of selected SNP for polygenic hazard scores. Selected 31 SNPs, their closest genes, log hazard ratio estimates, and their conditional p values in the final joint model, after controlling for effects of gender and APOE variants.

	Chr	Position	Gene	log(HR)	Conditional p in -log10
ε 2 allele	19		APOE	-0.47	>15
ε 4 allele	19		APOE	1.03	>20
rs4266886	1	207685786	CR1	-0.09	2.7
rs61822977	1	207796065	CR1	-0.08	2.8
rs6733839	2	127892810	BIN1	-0.15	10.5
rs10202748	2	234003117	INPP5D	-0.06	2.1
rs115124923	6	32510482	HLA-DRB5	0.17	7.4
rs115675626	6	32669833	HLA-DQB1	-0.11	3.2
rs1109581	6	47678182	GPR115	-0.07	2.6
rs17265593	7	37619922	BC043356	-0.23	3.6
rs2597283	7	37690507	BC043356	0.28	4.7
rs1476679	7	100004446	ZCWPW1	0.11	4.9
rs78571833	7	143122924	AL833583	0.14	3.8
rs12679874	8	27230819	PTK2B	-0.09	4.2
rs2741342	8	27330096	CHRNA2	0.09	2.9
rs7831810	8	27430506	CLU	0.09	3
rs1532277	8	27466181	CLU	0.21	8.3
rs9331888	8	27468862	CLU	0.16	5.1
rs7920721	10	11720308	CR595071	-0.07	2.9
rs3740688	11	47380340	SPI1	0.07	2.8
rs7116190	11	59964992	MS4A6A	0.08	3.9
rs526904	11	85811364	PICALM	-0.2	2.3
rs543293	11	85820077	PICALM	0.3	4.2
rs11218343	11	121435587	SORL1	0.18	2.8
rs6572869	14	53353454	FERMT2	-0.11	3
rs12590273	14	92934120	SLC24A4	0.1	3.5
rs7145100	14	107160690	abParts	0.08	2
rs74615166	15	64725490	TRIP4	-0.23	3.1
rs2526378	17	56404349	BZRAP1	0.09	4.9
rs117481827	19	1021627	C19orf6	-0.09	2.5
rs7408475	19	1050130	ABCA7	0.18	4.3
rs3752246	19	1056492	ABCA7	-0.25	8.4
rs7274581	20	55018260	CASS4	0.1	2.1

PHS successfully stratified individuals from independent cohorts into different risk strata (Figure 2.3). Among AD cases in the ADGC Phase 2 cohort, we found that the predicted age of onset was strongly associated with the empirical (actual) age of onset (binned in

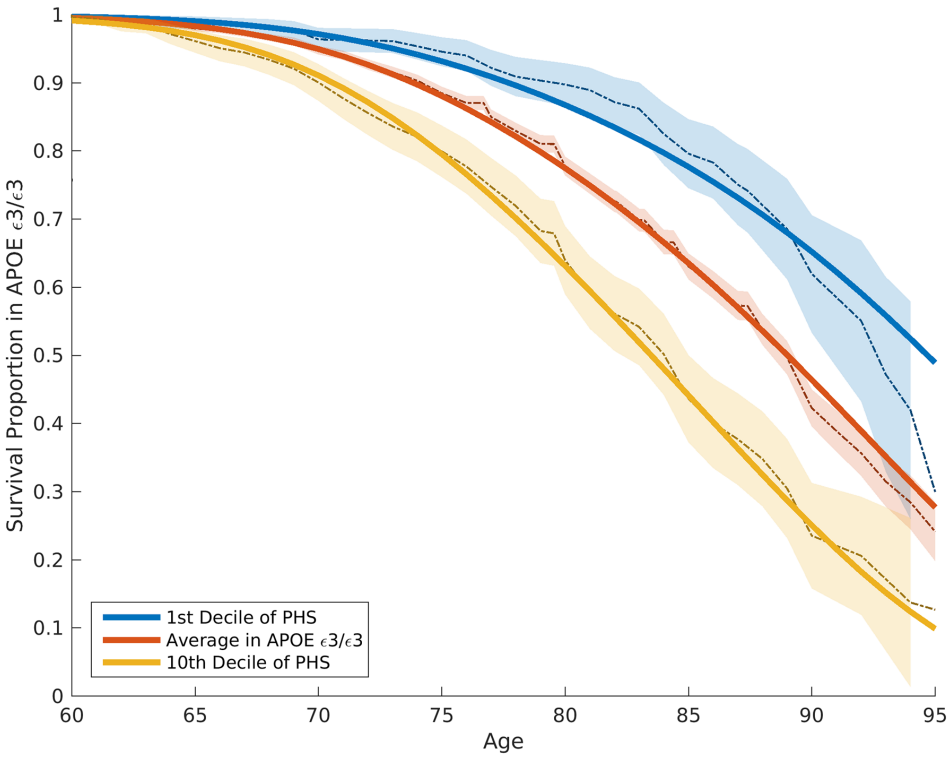

Figure 2.2: Survival models among APOE 3/3 individuals. Kaplan-Meier estimates and Cox proportional model fits among APOE 3/3 individuals in ADGC phase 1 dataset, excluding NACC and ADNI samples. The solid line represent the Cox fit whereas the dashed line and shaded regions represent Kaplan-Meier estimation with 95% confidence interval.

percentiles, r = 0.90, p = 1.1 x 10-26, Figure 2.3). Similarly within the NACC subset with a high level of AD neuropathologic change, we found that PHS strongly predicted time to progress to neuropathologically defined AD (Cox proportional hazard model, z = 11.8723, p = 2.82 x 10-32).

To evaluate risk for developing AD, combining the estimated hazard ratios from the ADGC cohort, allele frequencies for each of the AD-associated SNPs from the 1000 Genomes Project and the disease incidence in the general US population [35], we generated the population baseline-corrected survival curves given an individuals genetic profile and age. iven an individuals genetic profile and age, the corrected survival

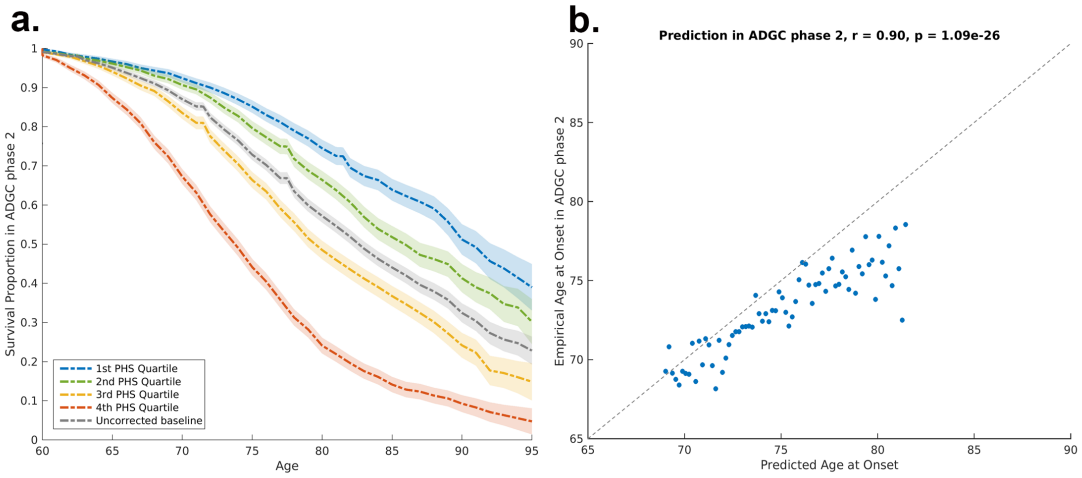

Figure 2.3: Model performance in replication sample. (a) Risk stratification in ADGC phase 2 cohort, using PHS derived from ADGC phase 1 dataset. (b) Predicted age of AD onset as a function of empirical age of AD onset among cases in ADGC phase 2 cohort. Prediction is based on the final survival model trained in the ADGC phase 1 dataset. The dashed line and shaded regions represent Kaplan-Meier estimation with 95% confidence interval.

proportion can be translated directly into incidence rates (Figure 2.4). As previously reported in a meta-analysis summarizing four studies from the US general population, the annualized incidence rate represents the proportion (in percent) of individuals in a given risk stratum and age, who have not yet developed AD but will develop AD in the following year; thus the annualized incidence rate represents the instantaneous risk for developing AD conditional on having survived up to that point in time. For example, for a cognitively normal 65 year-old individual in the 80th percentile PHS, the incidence rate would be: 0.29 at age 65, 1.22 at age 75, 5.03 at age 85, and 20.82 at age 95 (Figure 2.4); in contrast, for a cognitively normal 65 year old in the 20th percentile PHS, the incidence rate (per 100 person-years) would be 0.10 at age 65, 0.43 at age 75, 1.80 at age 85, and 7.43 at age 95. As independent validation, we examined whether the PHS predicted incidence rate reflects the empirical progression rate (from normal control to clinical AD). We found that the PHS predicted incidence was strongly associated with

empirical progression rates (Cochrane Armitage trend test, p = 1.54 x 10-10).

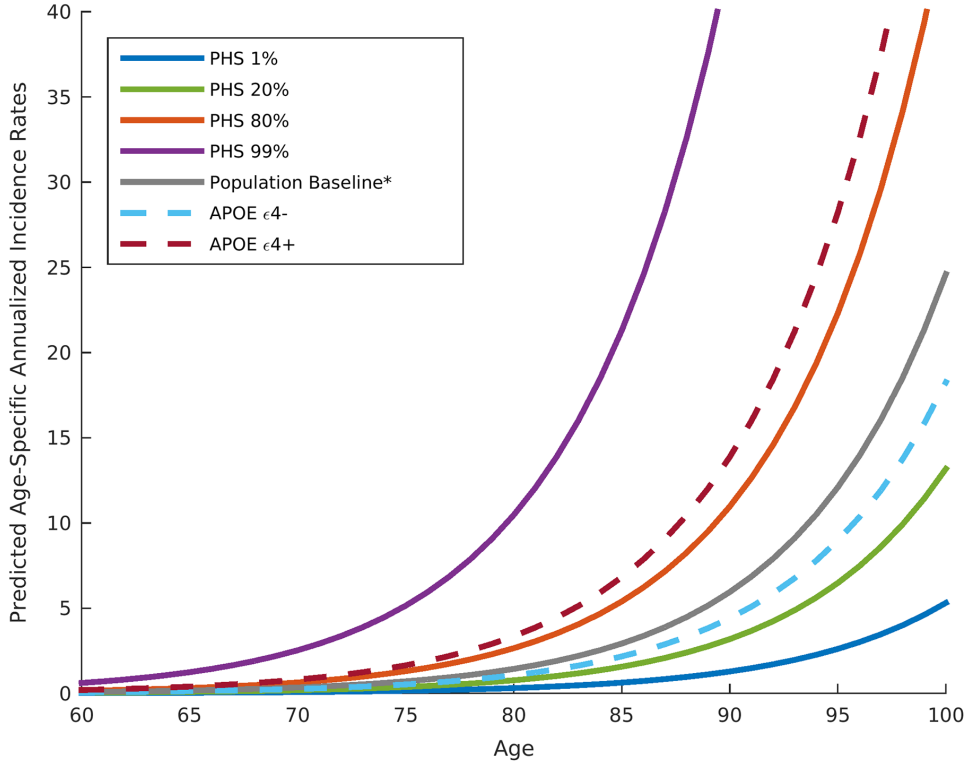

Figure 2.4: Predicted annualized incidence rate given PHS. Annualized incidence rates showing the instantaneous hazard as a function of PHS percentiles and age. The gray line represents the population baseline estimate.

We found that the PHS was significantly associated with Braak stage of NFTs (-coefficient = 0.115, standard error (SE) = 0.024, p-value = 3.9 x 10-6) and CERAD score for neuritic plaques (-coefficient = 0.105, SE = 0.023, p-value = 6.8 x 10-6). We additionally found that the PHS was associated with worsening CDR-Sum of Box score over time (-coefficient = 2.49, SE = 0.38, p-value = 1.1 x 10-10), decreased CSF A1-42 (reflecting increased intracranial A plaque load) (-coefficient = -0.07, SE = 0.01, p-value = 1.28 x 10-7), increased CSF total tau (-coefficient = 0.03, SE = 0.01, p-value = 0.05), and increased volume loss within the entorhinal cortex (-coefficient = -0.022, SE = 0.005, p-value = 6.30 x 10-6) and hippocampus (-coefficient = -0.021, SE = 0.0054, p-value =

7.86 x 10-5).

2.4 Discussion

In this study, by integrating AD-associated SNPs from recent GWAS and disease incidence estimates from the US population into a genetic epidemiology framework, we have developed a novel polygenic hazard score for quantifying individual differences in risk for developing AD, as a function of genotype and age. The PHS systematically modified age of AD onset, and was associated with known in vivo and pathologic markers of AD neurodegeneration. In independent cohorts (including a neuropathologically confirmed dataset), the PHS successfully predicted empirical (actual) age of onset and longitudinal progression from normal aging to AD. Even among individuals who do not carry the 4 allele of APOE (the majority of the US population), we found that polygenic information is useful for predicting age of AD onset.

Using a case/control design, prior work has combined GWAS-associated polymorphisms and disease prediction models to predict risk for AD [36, 37, 38, 39, 40, 41]/ Rather than representing a continuous process where non-demented individuals progress to AD over time, the case/control approach implicitly assumes that normal controls do not develop dementia and treats the disease process as a dichotomous variable where the goal is maximal discrimination between diseased cases and healthy controls. Given the striking age-dependence of AD, this approach is clinically suboptimal for estimating risk of AD. Building on prior genetic estimates from the general population [15, 42], we employed a survival analysis framework to integrate AD-associated common variants with established population-based incidence [35] to derive a continuous measure, polygenic hazard score (PHS). We note that the PHS can estimate individual differences in AD risk across a lifetime and can quantify the yearly incidence rate for developing AD.

These findings indicate that the lifetime risk of age of AD onset varies by polygenic profile. For example, the annualized incidence rates (risk for developing AD in a given year) are considerably lower for an 80-year old individual in the 20th percentile PHS relative to an 80-year old in the 99th percentile PHS (Figure 2.4). Across the lifespan, our results indicate that even individuals with low genetic risk (low PHS) develop AD, but at a later peak age of onset. Certain loci (including APOE 2) may protect against AD by delaying, rather than preventing, disease onset.

Our polygenic results provide important predictive information beyond APOE. Among APOE 3/3 individuals, who constitute 70-75% of all individuals diagnosed with late-onset AD, age of onset varies by more than 10 years, depending on polygenic risk profile (Figure 2.2). At 60% AD risk APOE 3/3 individuals in the 1st decile of PHS have an expected age of onset of 85 whereas for individuals in the 10th decile of PHS, the expected age of onset is greater than 95. These findings are directly relevant to the general population where APOE 4 only accounts for a fraction of AD risk 3 and are consistent with prior work 21 indicating that AD is a polygenic disease where non-APOE genetic variants contribute significantly to disease etiology.

We found that the PHS strongly predicted age of AD onset in within the ADGC phase 2 dataset and the NACC neuropathology confirmed subset demonstrating independent replication of our polygenic score. Within the NIA ADC sample, the PHS robustly predicted longitudinal progression from normal aging to AD illustrating that polygenic information can be used to identify cognitively normal older individuals at highest risk for developing AD (preclinical AD). We found a strong relationship between PHS and increased tau associated NFTs and amyloid plaques suggesting that elevated genetic risk may make individuals more susceptible to underlying Alzheimers pathology. Consistent with recent studies showing correlations between AD polygenic risk scores and markers of Alzheimers neurodegeneration [38, 39], our PHS also demonstrated robust associa-

tions with CSF A1-42 levels, longitudinal MRI measures of medial temporal lobe volume loss and longitudinal CDR-SB scores illustrating that increased genetic risk may increase likelihood of clinical progression and developing neurodegeneration measured in vivo.

From a clinical perspective, our genetic risk score may serve as a risk factor for accurately identifying older individuals at greatest risk for developing AD, at a given age. Conceptually similar to other polygenic risk scores (for a review of this topic see [18]) for assessing coronary artery disease risk [43] or breast cancer [44], our PHS may help in predicting which individuals may test positive for clinical, CSF or imaging markers of AD pathology. Importantly, a continuous, polygenic measure of AD genetic risk may provide an enrichment strategy for prevention and therapeutic trials and could also be useful for predicting which individuals may respond to therapy. From a disease management perspective, by providing an accurate, probabilistic assessment regarding the likelihood of Alzheimers neurodegeneration, determining a genomic profile of AD may help initiate a dialogue on future planning. Finally, a similar genetic epidemiology framework may be useful for quantifying the risk associated with numerous other common diseases.

There are several limitations to our study. We primarily focused on Caucasian individuals of European descent. Given that AD incidence [45], genetic risk [42, 46] and likely linkage disequilibrium in African-Americans and Latinos is different from Caucasians, additional work will be needed to develop a polygenic risk model in non-Caucasian (and non-US) populations. The majority of the participants evaluated in our study were predominantly recruited from specialized memory clinics or AD research centers and may not be representative of the general US population. In order to be clinically useful, we note that our PHS needs to be prospectively validated in large community based cohorts, preferably consisting of individuals from a range of ethnicities. The previously reported population annualized incidence rates were not separately provided for males and females [35]. Therefore, we could not report PHS annualized

incidence rates stratified by sex. Another limitation is that our PHS may not be able to distinguish pure AD from a mixed dementia presentation since cerebral small vessel ischemic/hypertensive pathology often presents concomitantly with Alzheimers neurodegeneration and additional work will be needed on cohorts with mixed dementia to determine the specificity of our polygenic score. Finally, we focused on APOE and GWAS-detected polymorphisms for disease prediction. Given the flexibility of our genetic epidemiology framework, it can be used to investigate whether a combination of common and rare genetic variants along with clinical, cognitive and imaging biomarkers may prove useful for refining the prediction of AD age of onset.

In conclusion, by integrating population based incidence proportion and genome-wide data into a genetic epidemiology framework, we have developed a polygenic hazard score for quantifying the age-associated risk for developing AD. Measures of polygenic variation may prove useful for stratifying AD risk and as an enrichment strategy in clinical trials.

2.5 Acknowledgement

Chapter 2, in full, is a reprint of the material as it appears in *Plos Medicine* 2017. Rahul S. Desikan*, Chun Chieh Fan*, Yunpeng Wang, Andrew J. Schork, Howard J. Cabral, L. Adrienne Cupples, Wesley K. Thompson, Lilah Besser, Walter A. Kukull, Dominic Holland, Chi-Hua Chen, James B. Brewer, David S. Karow, Karolina Kauppi, Aree Witoelar, Celeste M. Karch, Luke W. Bonham, Jennifer S. Yokoyama, Howard J. Rosen, Bruce L. Miller, William P. Dillon, David M. Wilson, Christopher P. Hess, Margaret Pericak-Vance, Jonathan L. Haines, Lindsay A. Farrer, Richard Mayeux, John Hardy, Alison M. Goate, Bradley T. Hyman, Gerard D. Schellenberg, Linda K. McEvoy, Ole A. Andreassen, Anders M. Dale. PLoS, 2017. The dissertation author was the

primary investigator and co-first author of this paper.

Chapter 3

Modeling the 3D Geometry of the Cortical Surface With Genetic Ancestry

3.1 Introduction

Knowing how the human brain is shaped by migration and admixture is a critical step in studying human evolution [47, 48], as well as preventing the bias of hidden population structure in brain research [49, 50]. Yet the neuroanatomical differences engendered by population history are still poorly understood. Most of the inference relies on craniometric measurements, because morphology of the brain is presumed to be the neurocraniums main shaping force before bones are fused and ossified [51]. Although studies have shown that the shape variations of cranial bones are consistent with population history [52, 53, 54], it is unknown how much human ancestry information is retained by the human cortical surface. In our groups previous study, we found that the area measures of cortical surface and total brain volumes of European descendants

in the United States correlate significantly with their ancestral geographic locations in Europe [55]. Here, we demonstrate that the 3-dimensional geometry of cortical surface is highly predictive of individuals genetic ancestry in West Africa, Europe, East Asia, and America, even though their genetic background has been shaped by multiple waves of migratory and admixture events. The geometry of the cortical surface contains richer information about ancestry than the areal variability of the cortical surface, independent of total brain volumes. Besides explaining more ancestry variance than other brain imaging measurements, the 3D geometry of the cortical surface further characterizes distinct regional patterns in the folding and gyrification of the human brain associated with each ancestral lineage.

3.2 Results

The participants were recruited as part of the Pediatric Imaging, Neurocognition, and Genetics (PING) study. A detailed overview of the study can be found in previous publications (e.g., [49, 50, 56]). Briefly, PING was a multi-site project recruiting children and adolescents from ages 3 to 21 at 10 sites in the United States. All participants were screened for history of major developmental, psychiatric, or neurological disorders; brain injury; or other medical conditions that affect development. Participants then received neurodevelopmental assessments, standardized multimodal neuroimaging, and genome-wide genotyping. The overall PING sample consisted of 1,493 participants; 1,152 individuals remained after quality control of the genotyping and neuroimaging data (for quality control processes and demographics of the participants, see Supplemental Information and Table S1). We focused our analyses on 562 individuals older than 12 years (289 males, mean age: 16.6 years, standard deviation: 2.6 years). Considering that the morphological features of cortical surface change little after age 12 [56], this

stratified approach further reduced the residual confounds of developmental effects.

The proportions of genetic ancestry were estimated using principal component (PC) analysis with whole-genome single nucleotide polymorphism (SNP) reference panels for ancestry [57, 58, 59]. Four continental populations were used as ancestral references: West Africa (YRI, as Yoruba in Ibadan), Europe (CEU, as Utah residents with northern and western European ancestry), East Asia (EA), and America (NA, as America natives). The metrics for summarizing genetic ancestry in each ancestral component were standardized as proportions, ranging from 0% to 100%. These proportions represent how similar an individual is to the reference population genetically [59].

3.2.1 Morphological Prediction for Genetic Ancestry

We first tested whether the surface geometry of the cerebral cortex predicted the proportion of genetic ancestry among participants. To characterize variation in the geometry, we reconstructed the cortical surfaces from all individuals T1-weighted scans, then represented the positions of the corresponding surface vertices using standard 3-D Cartesian coordinates. The reconstruction and registration processes ensure that each vertex on the reconstructed cortical surface is located in a homologous position with respect to the curvature patterns for individuals [15, 16]. Taking the coordinates of all vertices as a whole, we then have information about shape variation of the cortical surface, including aspect ratios, sulcal depth, and gyrification. The prediction models were fit with ridge regression while treating gender, age, age squared, total brain volumes, and the scanner on which the image data were acquired, as nuisance covariates. The model performance was evaluated using leave-one-out cross-validation (LOOCV).

As Figure 3.1 shows, the geometry of the cortical surface has good predictive value for each of the ancestry components. The variances explained by the models are 66% for ancestry in YRI, 55% for ancestry in CEU, 49% for ancestry in EA, and 47% for

ancestry in NA. To determine to what degree the geometric differences reflect variation in area expansion of cortical surface, comparable models were computed using vertex-wise surface area (Table 3.1). Also, to examine possible roles in the prediction of simpler morphological attributes, such as aspect ratios of the cerebrum and volumes of subcortical structures, we conducted comparable analyses predicting ancestry from these measures. None has as much information about ancestry as the geometry of cortical surface (Table 3.1).

Table 3.1: Percentage of Variance Explained in Different Predictive Models. Cortical surface geometry and cortical surface area are sampled in icosahedral level 4, which contains 642 vertices in each hemisphere. All models are fit with the same setting and evaluated with leave-on-out cross validation.

	Cortical Surface Geometry	Cortical Surface Area	Brain Aspect Ratio	Subcortical Volumes
YRI	66%	17%	10%	5%
CEU	55%	12%	2%	2%
EA	49%	9%	6%	6%
NA	47%	9%	9%	0%

3.2.2 Characterization of the Cortical Shape Morphs

We then reconstructed the 3D geometry of the cortical surface based on the linear relationship we observed between cortical surface geometry and the proportion of genetic ancestry. This allowed us to visualize how the geometry of the cortical surface changes as a function of increasing proportion of genetic ancestry in each ancestral component. The morphing of 3D cortical surfaces from neutral ancestry (25% of genetic ancestry in all four components) to 100% ancestry in each component is demonstrated in Figure 3.2. As Figure 3.2 illustrates, the textural contrasts between regions of the cortical surface indicate that the morphing process has complex, unique patterns for each ancestral component, while the intensity varies from region to region. For example, as

the proportion of the YRI component increases, the temporal surfaces move posteriorly and inward. The proportion of the CEU component is associated with protrusion of the occipital and frontal surfaces. Increases in the proportion of the EA component are accompanied by variations in temporal-parietal regions. The NA component is associated with flattening of the frontal and occipital surfaces.

Figure 3.3 summarizes the mean magnitudes and variations of the morphing in each cortical surface region defined by genetic correlations [19]. The mean magnitudes vary from cortical region to cortical region, corresponding to the description above. In addition, YRI, EA, and NA all have relatively high magnitude and variations of morphing in the posterolateral-temporal region.

3.3 Discussion

Our data indicate that the unique folding patterns of gyri and sulci are closely aligned with genetic ancestry. The geometry robustly predicts each individuals genetic background even though the population has been shaped by waves of migration and admixtures [57, 60]. Previously, only modeling of facial features has achieved 64% of explained variance in the YRI ancestry among African Americans [61]. Our 3D representation of cortical surface geometry performs similarly in predicting YRI ancestry and also performs well for the other three continental ancestries. As data in Table 3.1 show, the explanatory power is not due to the differences in total brain volumes, nor to the differences in areal expansion of the cortical surface. Instead, regional folding patterns characterize each ancestral lineage.

On the other hand, the global shapes of the reconstructed cortical surface geometry match W. W. Howells description on craniometry of 2,524 ancient human crania from 28 populations [62]. Crania of African ancestry tended to have a narrower cranial base, and

those of Northern European ancestry had elongated occipital and frontal regions. Crania of East Asian ancestry had a high cranial vault, and those of Native American ancestry had a flatter cranium. Regarding the morphing differences of YRI, EA, and NA, all had high magnitude and variations in the posterior-temporal regions (Figure 3.3).These findings are consistent with the notion that temporal bones contain more variations across ancestral groups [52].

At first glance, these results are surprising because our model is based on the contemporary United States population, which is the historical product of migrations, slave trades, and local admixture events [60, 63, 64]. Nevertheless, the coordinates of reference-inferred PC space reflect information about individuals ancestral origins [59, 63, 65]. Our groups previous study also showed that the individuals positions in the PC space are matched with their ancestral locations, rather than their current geographic locations [55]. Therefore, our 3D representation might, to a certain degree, reflect the neuroanatomical and/or neurocranial changes along the human migratory path in the dispersal from Africa [66]. More precise characterization of an individuals ancestral origins would require more complex estimates of ancestry based on global-scale reference panels [32]. Further understanding of neuroanatomical change along the Out-of-Africa scenario based on brain imaging data will require future studies using sampling methods similar to the Human Genome Diversity Project [67].

It is important to note that these ancestry-related geometric features of the cortical surface are not substantially attributable to variation in cortical surface area. Previous studies of ancient crania often interpreted the shape differences as evidence of relative size alterations of different cortical functional domains [68, 51]. Our results suggest that in the case of the contemporary population, the differences in cortical surface geometry might not reflect variation in the relative surface area of different functional cortical regions. In prior studies, regionalization of the cortex has been linked to cognitive differences in

humans [49, 50]. Any functional significance of the cortical surface geometry, per se, remains to be established. The effects reported here might be mediated by neutral drift of the phenotypic variations [69]. They can also result from a complex interaction between the brain and neurocranium, with the former expanding while the latter acts as physical resistance. Nevertheless, the causal relationships between the observed shapes and crania are beyond the scope of our current study.

An implication of our ancestry-related 3D models is that, unless properly controlled, hidden population structures could present a challenge in brain imaging studies of admixed populations [65]. The regional differences between ancestral groups include changing sulcus depths and folding angles. This issue becomes particularly relevant in large, multi-site U.S. and international brain imaging studies [70]. With the advent of inexpensive, high-throughput genotyping, it is now possible to control for spurious effects due to ancestry admixture using genetically derived admixture factors in the statistical analysis of data [49, 50]. It is also possible that the phenomena we observed are linked with specific ancestral haplotypes. It may therefore be possible to use the ancestral information to improve statistical power for gene discovery with methods such as admixture mapping [71].

3.4 Acknowledgement

Chapter 3, in full, is a reprint of the material as it appears in *Current Biology* 2015. Chun Chieh Fan, Hauke Bartsch, Andrew Schork, Chi-Hua Chen, Yunpeng Wang, Min-Tzu Lo, Timothy T. Brown, Joshua M. Kuperman, Donald J. Hagler Jr., Nicholas Schork, Terry L. Jernigan, Anders M. Dale. Cell Press, 2015. The dissertation author was the primary investigator and author of this paper.

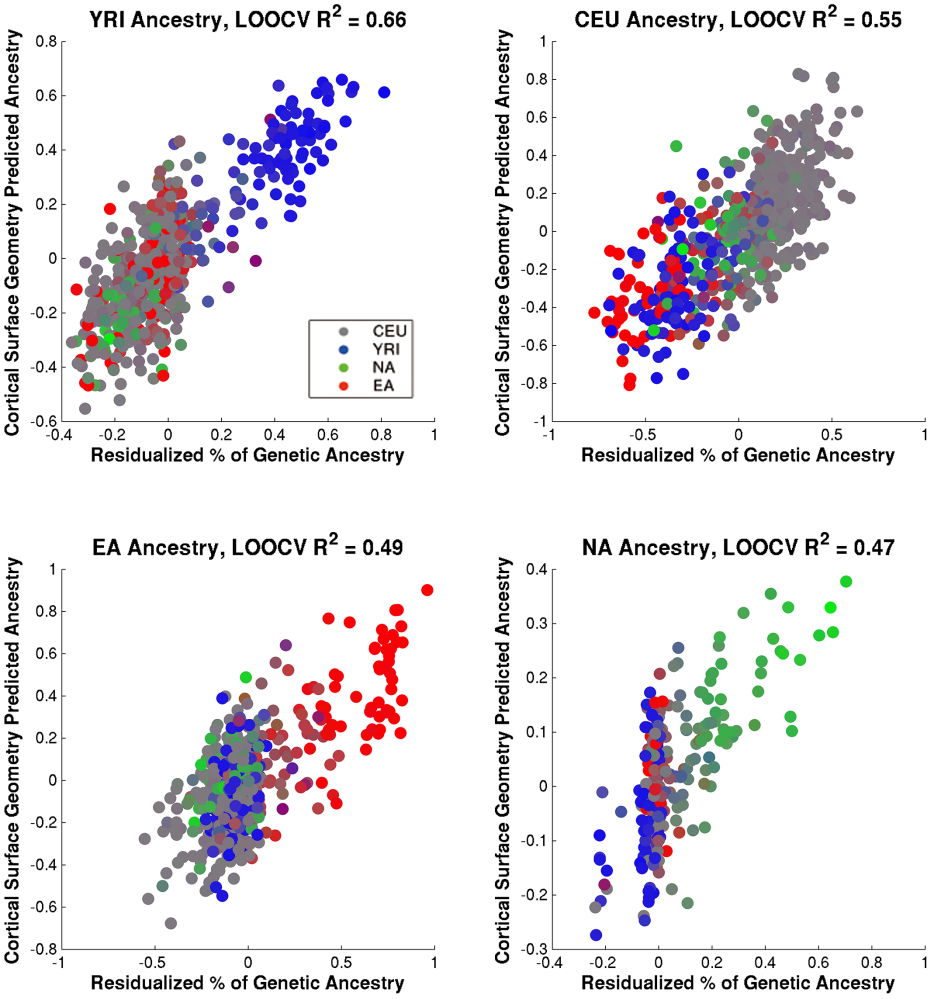

Figure 3.1: Predicting the proportion of genetic ancestry by cortical surface geometry. YRI: Yoruban, as the West Africa ancestry. CEU: Utah residents with northern and western European ancestry. EA: East Asia. NA: America natives. In all predictive models, the variables have been residualized with respect to the age, age squared, gender, total brain volumes, and scanner used. All models excluded individuals with a 0 proportion of genetic ancestry to that specific component. LOOCV: leave-one-out cross-validation. The colors of the data points are determined by the proportion of genetic ancestry as illustrated in the figure legend.

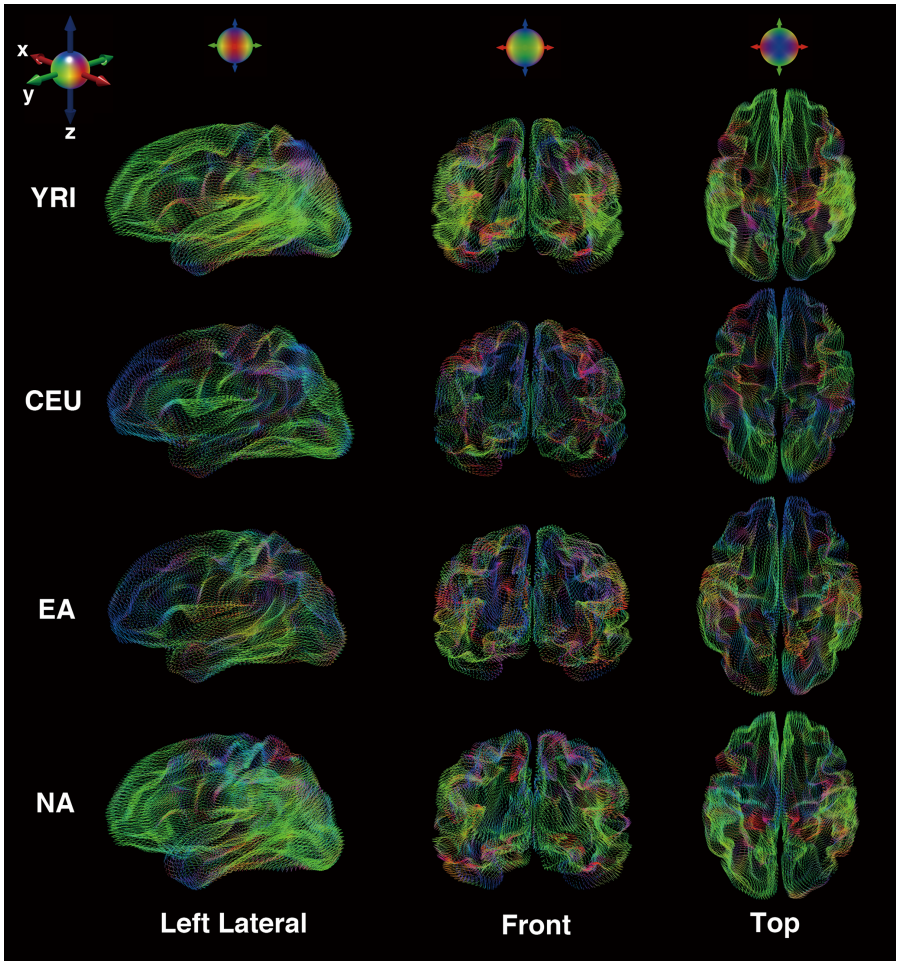

Figure 3.2: Color-coded morphing process of the 3D geometry of the cortical surface. The still image illustrates how each vertex on the cortical surface morphs from an ancestry-neutral 3D cortical surface (a 25% proportion of genetic ancestry in all ancestral components) to a 3D cortical surface with a 100% proportion of genetic ancestry in a specific ancestral component. The morphing coefficients were estimated from the PING sample. Here, the colors represent the direction of the morphing process. Moving along the medial-lateral axis is coded in red, along the anterior-posterior axis in green, along the dorsal-ventral axis in blue. The final color is the combination of these three, depending on which direction the vertices move. For each viewing perspective, the coloring frame of reference is rendered on the top of each column. The length of each morphing line is the actual distance between two 3D cortical surfaces.

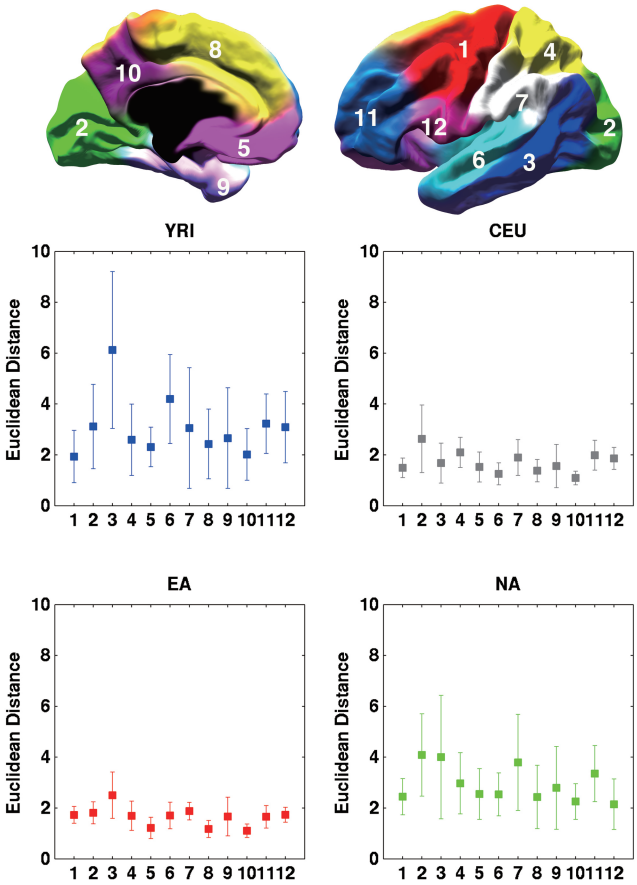

Figure 3.3: Mean magnitude and variations of morphing across 12 regions of cortical surface. Labeled in the topmost images, the following regions are defined in a previous publication [19]: 1. central region, 2. occipital cortex, 3. posterolateral-temporal region, 4. superiorparietal region, 5. orbitofrontal region, 6. superiotemporal region, 7. inferiorparietal region, 8. dorsomedialfrontal region, 9. anteromedial-temporal region, 10. precuneus, 11. dorsolateral-prefrontal cortex, 12. parsopercularis. The Euclidean distances between cortical surface of 100% ancestry and neutral ancestry were calculated for each vertex. Depending on the surface regions where the verticesare situated, the mean and standard deviations of the Euclidean distances are shown in the boxplots.

Chapter 4

Williams Syndrome-Specific Neuroanatomical Profile and Its Associations with Behavioral Features

4.1 Introduction

Williams Syndrome (WS) is a rare multi-system disorder caused by hemideletion of 26 genes on chromosome 7. Although the cognitive impact of WS is evident in general intelligence and visuospatial capabilities, the cardinal feature of WS cognition is overly social behavior [72]. WS individuals express heightened social approach behavior and social emotional behavior very early on, distinguishing them from others with disorders that include intellectual impairment [73]. This had led to extensive research using magnetic resonance imaging (MRI), in the hope of identifying the mediating neural processes from genetic deletions to social behavioral impact [74]. Previous MRI studies had found that what distinguishes WS from other genetic disorders with intellectual impairment e.g., Down syndrome is not the reduced total brain volume per se, but the

aberrant regionalization of the brain [75]. The most consistent findings are the gyral patterns in the superior parietal regions and orbital frontal cortex, which were found to be different between WS patients and healthy individuals [76, 77, 78, 79].

Yet the specificity of these findings to WS and relevance to its distinct behavioral features were left unanswered. Differences in regional cortical surface area, such as in lingual gyrus, post-central gyrus, and temporal poles, were also reported [80, 81]. Abnormalities in the Sylvian fissures (Eckert et al., 2006) and disproportional volumetric changes of subcortical structures were also reported, but not consistent [79, 82, 83, 78]. Furthermore, the diagnostic process for WS requires that clinicians identify individuals with WS features and use fluorescent in situ hybridization (FISH) to confirm. This precludes identification of individuals who have different deletions in the WS chromosome region (WSCR), resulting in slightly altered profiles of WS features. A recent analysis focused on cases of individuals with atypical deletions in the WSCR suggested that the varying size of the deletion would result in different behavioral profiles [84], which conceivably would make it difficult to identify those individuals in clinical settings. The rarity of both typical and atypical WS individuals makes the quantitative comparisons across MRI measures and groups impractical.

Here, we re-examined the WS-specific neuroanatomical profile using a novel analytic approach with the aim of developing a scoring system to quantify WS neuroanatomical variations. First, we extracted the WS-specific neuroanatomical profile from an adult WS cohort, using multiple measures derived from structural MRI of cerebrum, including subcortical volumes, cortical surface area [8, 9], sulcal depth [77], and cortical surface geometry [85]. To deal with the large number of MRI measures and limited sample size, we used an elastic-net model to achieve balance between the robust prediction and sparseness for easy interpretation. The resulting model provides the basis for calculating WS neuroanatomical scores that represent the similarity of an

individuals brain to the WS given his/her multimodal MRI features. The generalizability of the WS-specific neuroanatomical profile was then tested in an independent child WS cohort. After establishing the generalizability of the model, we examined whether the WS neuroanatomical scores could reflect the reduced size of genetic deletions in WSCR and whether the scores were associated with the behavioral features of WS.

4.2 Methods

All participants were recruited as part of a multi-project program, including two cohorts in current analyses, one as child cohort and the other as adult cohort. Except time of recruitment, age differences, additional diagnostic groups, and behavior measures, the protocols for inclusion and imaging acquisition were kept the same, which were described in separate publications [86, 87]. Participants were screened based on the following measures: normal or corrected vision/hearing, English native-language speaker, and no remarkable mental health history. Caregivers completed an interview and extensive demographic and family history questionnaires to assess whether participants met the screening criteria. Caregivers and child participants provided consent and assent, respective, for participation. Individuals with intellectual disabilities required a more simple, verbally delivered description for assent along with guardian informed consent. All procedures were explained in person, within the testing environment, with the caregiver present, to show the participants more concretely what to expect. They could choose at any time to withdraw from participation, even after beginning. Study protocols were approved by the Institutional Review Boards at the Salk Institute and at UCSD.

4.2.1 Adult WS Cohort

The adult cohort, on which the WS-specific neuroanatomical profile was trained, consisted of 22 individuals with typical WS deletions (approximately 26 genes in the WSCR 7q11.23 region) as well as 16 healthy controls (HC) (Table 4.1). Part of this cohort has been involved in a series of MRI studies for WS that were published elsewhere [86, 88]. The diagnosis of WS was based on clinical presentation (WS Diagnostic Score Sheet) and confirmation of meeting genetic criteria for WS using fluorescent in situ hybridization. HCs were screened for a history of neurological disorders, psychiatric illness, and substance abuse. Intellectual functioning was assessed with the age-appropriate version of the Wechsler tests to include the Wechsler Adult Intelligence Scale 3rd Edition, Wechsler Abbreviated Scale of Intelligence (WASI), and Wechsler Intelligence Scale for Children 3rd Edition WISC-III [89]. Sociability was assessed with the Salk Institute Sociability Questionnaire (SISQ) [73].

4.2.2 Child Cohort

The generalizability of the WS-specific neuroanatomical profile was tested with a cohort of 60 children (age range 6 to 13 years): seven individuals with WS, 23 typical developing children (TD), and 30 individuals with heterogeneous diagnoses to include high-functioning autism (HFA), specific language impairment (SLI), and focal lesions in the brain (FL). The demographic characteristics of each cohort are shown in Table 4.1. Children with WS were diagnosed using the same criteria as adults with WS. Subjects in the TD group were recruited from the community, had scores on a standardized test of intellectual functioning (WASI) in the normal range and no history of developmental or language delay. Individuals with HFA, SLI, and FL were recruited from populations at a local pediatric neurology clinic and a clinic for speech and language disorders

[87]. Detailed recruiting procedures and diagnostic criteria can be found in previously published studies [87].

4.2.3 Individuals with Atypical Deletions in WSCR

We further examined if the scores from the trained model for WS-specific neuroanatomical profile can identify whose brain phenotypes lie between WS and HC, such as individuals with reduced deletion size on WSCR. We tested our model on five individuals from one family with small deletions on chromosome 7q11.23, sparing regions coding for *FZD9*, *GTF2I*, and *GTF2IRD1* [84].

4.2.4 Imaging Acquisition and Extracting Multimodal MRI Features

All participants were scanned on a 1.5 Tesla MRI scanner (GE HDxt, echo time (TE) = 3.0 msec, repetition time (TR) = 8.7msec, inversion time = 270 msec, flip angle = 8o, field of view = 24 cm, voxel size = 1.25x1.25x1.2 mm). To reduce and prevent possible motion artifacts, real-time prospective motion tracking and correction (PROMO) was used for all participating subjects [90, 91]. Distortions caused by nonlinearity of the spatial encoding gradient fields were corrected with predefined nonlinear transformations [92] . Non-uniformity of signal intensity was reduced with the nonparametric nonuniform intensity normalization method [93]. After initial image data inspection and quality control, T1-weighted images underwent automated volumetric segmentation and cortical surface reconstruction using methods implemented in Freesurfer software [8, 9]. This automated processing corrects variations in image intensity due to RF coil sensitivity inhomogeneities, registers to a common reference, then segments volumes into cortical and subcortical structures. For each cohort, one staff research associate performed quality

control (QC) of the surfaces and segmentations for all MRI images at the same time, blind to age and group identification. Both the child cohort and the adult cohort went through the same QC processes. The segmentations and reconstructed surfaces were inspected for accuracy, manually edited using control points, and iteratively re-processed, blind to age or group labels, to ensure consistent quality across different cohorts.

Four different morphological measures of T1-weighted images were derived, including the volumes of subcortical structures [8], sulcal depths of the cortical surface [77], cortical surface area [9], and geometric deformations of the cortical surface [85]. Sulcal depth is the distance from each point on the cortical surface to the average mid-plane of the cortical surface, measuring gyrification of the brain. Cortical surface area expansion is the area surrounding a given cortical surface point relative to total cortical surface area. The geometric deformation is the 3D Cartesian coordinates of the cortical surface, characterizing the folding patterns of the brain. Subcortical structure volumes were divided by total brain volumes, and sulcal depths and geometric deformations were divided by the cubic root of each total brain volume to produce a uniform index, as well as to control for the global brain volume differences. Those imaging features were selected as a comprehensive representation of the neuroanatomical variations of the human brain possible with structural MRI without unnecessary a priori defined regions of interest.

4.2.5 Model Training

To characterize the WS-specific neuroanatomical profile from MRI measures, we fit an elastic-net logistic regression using data from the adult cohort and checked their performance with 10-fold cross validation. The index for model performance was area under curve (AUC) in the ROC analysis. The model included all four types of MRI measures; that is, cortical surface area (642 vertices-per-hemisphere), sulcal depths of cortical surface (642 vertices-per-hemisphere), cortical surface geometry (642 vertices-

per-hemisphere, each with 3D Cartesian coordinates), and subcortical volumes (thalamus, caudate, putamen, globus pallidum, hippocampus, amygdala, nucleus accumbens, and ventral caudate). To achieve the goal of balancing between predictive power and parsimonious solution, we used the ridge penalties to reduce the problem of rank deficiencies and additional lasso penalties for removing less relevant features [94]. The tuning parameters were optimized during the cross-validation.

4.2.6 Model Validation

After deriving the WS-specific neuroanatomical profile from the previous training step, the model was applied to the whole child cohort in predicting WS status out of a heterogeneous group. The model was also applied to individuals with atypical deletion size in the WSCR to examine if the scores were in-between the typical WS and HC. Afterward, the relationships between model-predicted scores and behavioral measures were explored using mediation analysis. Within-group variations were examined using Pearson correlations while Sobel tests were used to test whether the group differences were mediated by the neuroanatomical profile.

4.3 Results

In classifying WS status, the 10-fold cross-validation AUC of the WS-specific neuroanatomical profile achieved 100% in the adult WS cohort (two-tail test for AUC greater than 0.5, p ¡ 0.05). The model removes 98.4% of the input variables, leaving 412 features from four MRI measures. Among individuals with atypical deletions in WSCR (atypical WS), their predicted scores of the WS-specific neuroanatomical profile lay between typical WS and HC, which is significantly greater than HC (t_{19} = 9.4, p ¡ 10^{-7}), and less than patients with typical WS (t_{25}=-2.2, p = 0.038). To further test the

generalizability of the model, we applied the WS-specific neuroanatomical profile to the whole child cohort. In this independent cohort, the profiling scores have AUC with 1.0 in predicting WS status, achieving 100% sensitivity and 100% specificity with various decision cut-points (Figure 4.1).

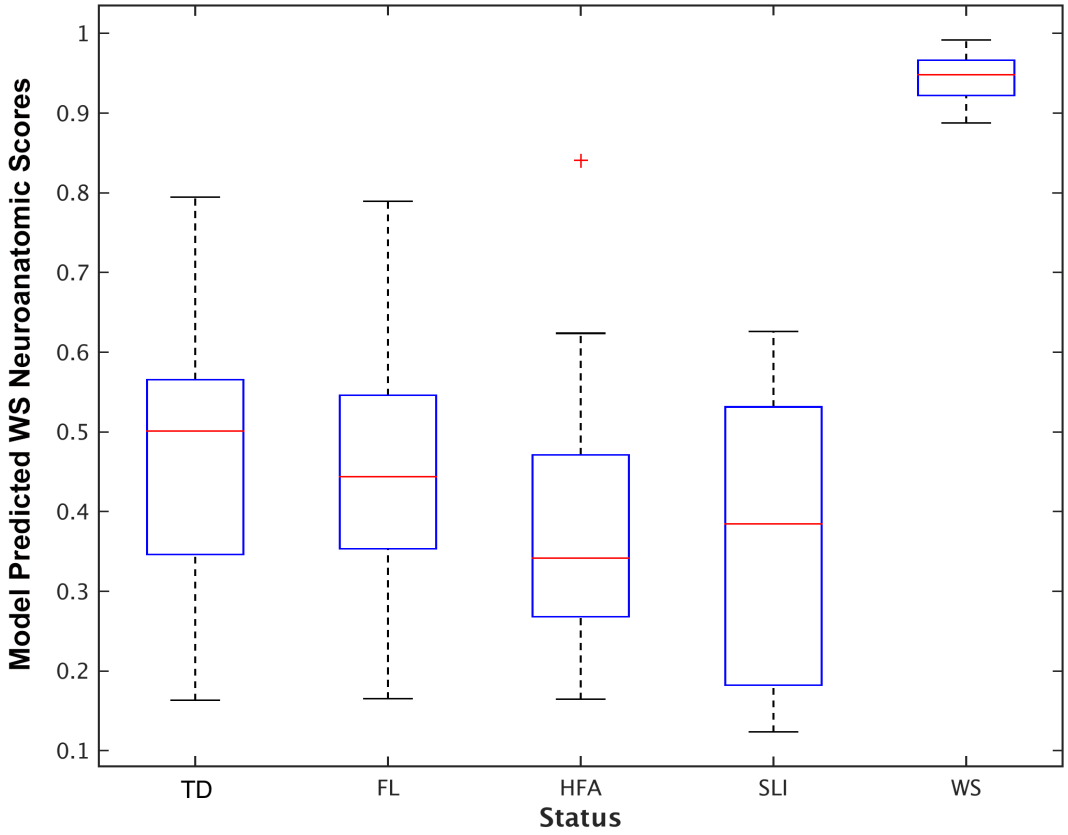

Figure 4.1: Boxplot of model predicted scores from trained WS-specific neuroanatomical profile across groups in the child cohort. The predicted scores of each group were demonstrated as median and inter-quartile range. The outliers were label as red-cross. Among them, children with WS have higher scores, none overlapping with any other group, that yield AUC of ROC analysis with one.

The cortical surface features extracted by the model are shown in Figure 4.2. The weights of selected features reflect the relative importance for predicting WS. Selected local features can be observed across different cortical surface regions, yet sparing the dorsal and medial part of the frontal cortex. The orbitofrontal cortex and superior

parietal cortex contain predictive features consistently across all three cortical surface measures (Figure 4.2). In addition, the cortical surface area contained predictive features in the Sylvian fissure and temporal poles. Two subcortical structures were also selected. Disproportionally decreasing sizes of left putamen (weights = -0.010) and left nucleus accumbens (weights = -0.014) were predictive for WS status.

The relationships among WS status, the WS-specific neuroanatomical profile, and behavioral function of WS are illustrated in Table 4.2. The Sobel tests for mediation indicate that the group differences in general intelligence, SISQ stranger score, and SISQ empathy score are largely explained by the mediating effect of the WS-specific neuroanatomical profile (all p values ¡ 10^{-3}, Bonferroni corrected). In the within-group analyses, the variations of the WS-specific neuroanatomical profile are significantly associated with SISQ empathy scores, with a trending p-value after applying Bonferroni correction for 9 independent tests (corrected p = 0.063).

4.4 Discussion

In this study, we sought to use a novel approach for characterizing the defining features of a WS-specific neuroanatomy and relating it to behavior. Features within the orbitofrontal cortex, superior parietal cortex, and Sylvian fissures were predictive for WS status across MRI measurements (Figure 4.1). Disproportional reductions in the putamen and nucleus accumbens are also important features for predicting WS status. The robust performance of our extracted WS-specific neuroanatomical profile are consistent in both adult and child cohorts (Figure 4.1). We also demonstrated that the scores for individuals with atypical deletions on WSCR lay between the WS and HC, and were associated with cardinal behavioral features of WS (Table 4.2).

MRI studies of WS have focused on localizing the neuroanatomical abnormalities [86, 76, 77, 95, 96, 88]. Although WS individuals have smaller brains in general, early studies have shown that the reductions are not uniformly distributed across brain regions [75]. Gyrification abnormalities in the orbitofrontal cortex, Sylvian fissures, and superior parietal regions have been reported [86, 76, 77, 88]. Some have found that amygdala volumes are disproportionally increased [82, 83] while others found no significant changes [79, 78]. The joint relationships across these neuroanatomical features were seldom examined in WS [74, 97]. One study had used tensor metrics of cortical surface to predict WS status in adult cohorts [97]. Different from what they attempted, our study aimed to evaluate all MRI measurements jointly, and the WS-specific neuroanatomical profile achieved 100% AUC in the independent child cohort (Figure 4.1).

Our sparse representation of WS profile matched with previously hypothesized causes of the behavioral profile of WS patients [79, 76, 88]. The selected features of cortical surface area located at the orbitofrontal, temporal parietal junction, and insula (Figure 4.2) are relevant to social functions [98, 96, 99]. The superior parietal region has been linked most strongly to the visuospatial processing deficits in WS [79]. Our mediation analyses using the Sobel test showed that the WS-specific neuroanatomical profile explained more variability in the behavioral measures than the WS status itself. This suggests that the WS-specific neuroanatomical profile may capture the underlying neuroanatomical factors that drive the related cognition and social behaviors. Since our behavioral analyses are limited in the WS adult cohort, we envision that longitudinal studies among children can be helpful to further establish the causal relationships between observed neuroanatomical profile and behavioral features of WS. Nevertheless, the robust performance of the WS-specific neuroanatomical profile in our child cohort suggests these features are already expressed during childhood.

Furthermore, case studies have indicated that atypical WS patients with smaller genetic deletions have lower social ratings than typical WS patients [73]. The telomere side of WS-related chromosomal regions, which tends to be spared in smaller deletions, contains genes such as *GTF2I* and *GTF2IRD1*, which have been associated with social behaviors in mouse models [100, 101]. Very recently, a study using induced pluripotent stem cells from WS suggested *FZD9* may be responsible for aberrant neurodevelopment [102]. Our data show that individuals with smaller deletions would have lower WS-specific anatomical scores than typical WS while those scores are positively correlated with hypersociability. These findings suggest that our extracted WS-specific profile of features might relate directly to the underlying genetic cause of hypersociability in WS.

Our study has several limitations. The training samples for the WS-specific neuroanatomical profile are relatively small compared with other machine-learning applications [94]. Small sample sizes are common in published studies of WS, considering that the prevalence of WS is rare [72]. Direct group comparisons across multiple MRI measures would suffer the burden of multiple hypotheses testing. Our approach for extracting WS-specific features circumvents this limitation of group comparisons. We kept a careful balance between interpretability and predictive power, achieving 100% AUC in both cross-validation of the adult cohort and the independent testing child cohort. Even though the robustness of the predictive performance is ensured, the feature selections are nevertheless constrained by the number of training samples [94]. This may explain why some previous reported neuroanatomical abnormalities, such as the amygdala [79, 82, 83], are not selected as predictive features. The neuroanatomical differences between WS patients and controls are not limited to regions we selected. The differences may be more similar to locally smoothed gradients. Meanwhile, it is also unclear how sensitive the WS-specific neuroanatomical profile is to the scanning protocols. Although our results indicate that our model can identify WS in different age

ranges from a very heterogeneous developmental cohort, the MRI images of training and testing samples were obtained and processed with the same protocol. Applying our WS-specific scores in other settings would be a further test of its clinical and research utilities.

Taken together, our novel multidimensional imaging approach captures the widespread differences observed within the neural architecture of individuals with WS. The model can have direct clinical applications, such as measuring the neuroanatomical phenotype of atypical WS with different sizes of deletions on WS chromosomal regions. Furthermore, a major benefit of our analytic strategy is that the extracted features can be readily applied to other imaging datasets. Applications of the extracted features on a large imaging genomic cohort would further inform research on the genetic influences of social behaviors.

4.5 Acknowledgement

Chapter 4, in full, is a reprint of the material as it appears in *NeuroImage:Clinical* 2017. Chun Chieh Fan, Timothy T. Brown, Hauke Bartsch, Joshua M. Kuperman, Donald J. Hagler Jr., Andrew Schork, Yvonne Searcy, Ursula Bellugi, Eric Halgren, Anders M. Dale. Elsevier, 2017. The dissertation author was the primary investigator and author of this paper.

Table 4.1: Demographics and global MRI measurements of participants in two cohorts.WS: Williams Syndrome. HC: Healthy controls. TD: Typical developed individuals. FL: Individuals with focal lesions in the MRI scans of brain. HFA: high function autism. SLI: specific language impairment.
Behavioral measures on Atypical WS were available for one male teenager. Hence, standard deviations were not shown.

Groups	n	Age - years		Gender -Male	Full IQ		SISQ - AS		SISQ - ES	
Adult WS cohort										
WS	22	31.6	(10.8)	59%	66.6	(5)	5.4	(1.4)	5.8	(0.8)
HC	16	25.9	(7)	37%	96.7	(14.6)	3.6	(1.2)	4.4	(1)
Atypical WS*	5	17.7	(2.5)	20%	74		4.2		5.8	
Child cohort										
WS	7	11.95	(1.75)	29%						
TD	23	9.48	(1.87)	52%						
FL	8	9.73	(1.26)	50%						
HFA	14	9.83	(1.45)	79%						
SLI	8	10.09	(1.48)	75%						

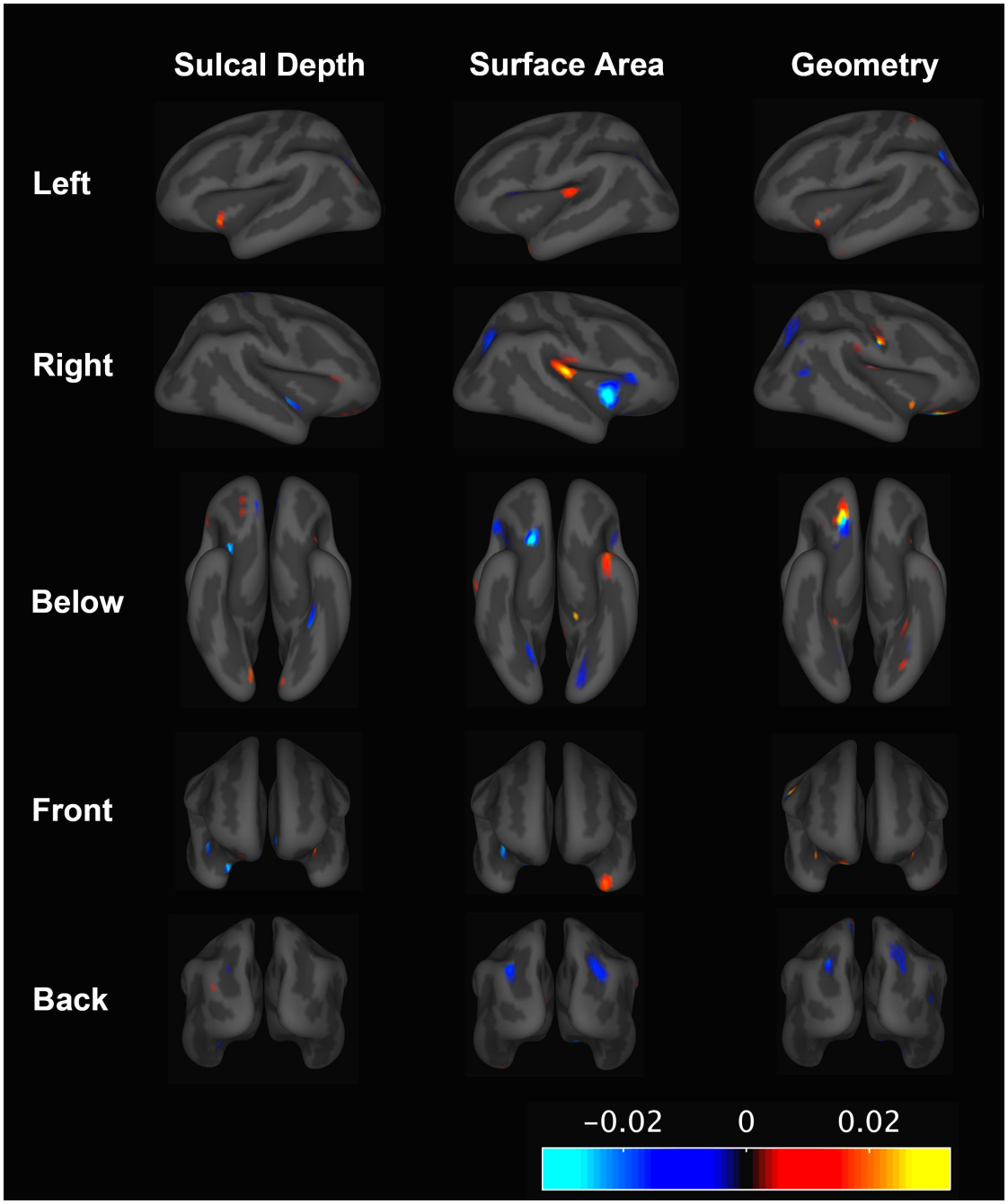

Figure 4.2: Elastic net model learnt features for predicting WS status. The blue or red indicates that the surface measures at that region were selected to be discriminative features. The red represents WS individuals with increased value of measures on that region ,whereas the blue represents the decreased value of measures among WS individuals. The magnitude of those colors indicates their relative importance for classifying WS and HC.

Table 4.2: Mediating effects and within-group correlations between model predicted WS neuroanatomic scores and behavioral measures. The mediating effect is checked with Sobel test for mediation, treating model predicted WS neuroanatomic scores as the mediator and each behavioral measure as dependent variable.

	Mediating Effect		Within HC		Within WS	
FIQ	z = -6.31	p = 1e-10	r = 0.29	p = 0.34	r = 0.18	p = 0.52
SISQ V Stranger	z = 3.73	p = 9e-5	r = -0.01	p = 0.96	r = 0.10	p = 0.74
SISQ - Empathy	z = 4.61	p = 2e-6	r = 0.09	p = 0.74	r = 0.70	p = 7e-3

Chapter 5

Williams Syndrome neuroanatomical score associates with GTF2IRD1 in large-scale magnetic resonance imaging cohorts: a proof of concept for multivariate endophenotypes

5.1 Introduction

The morphology of an adult brain represents a holistic snapshot of a unique neurodevelopmental history; its variations are an accumulation of dynamic processes working in concert with few constraints [103]. Different brain regions share the same original sets of proto-structures emerging from interactive molecular signaling programs during early embryonic stage. Post-natal brain growth, myelination and subsequent regressive processes leading to mature functional circuits provide further overlap in

the processes giving rise to adult brain morphology. These developmental processes, furthermore, are guided by distributed patterns of gene expression, interactions with the environment and operate under spatial constraints imposed by the cranium that may link the morphology of various parts of the adult brain [103, 104]. Consequently, the perturbation of a developmentally critical gene often results in diverse morphological abnormalities not limited to a single brain region [105, 100, 81]. Given this, it is reasonable to expect that variability interjected into neurodevelopment via a genetic variant may not only contribute to variability in the MRI derived morphology of a single delineated brain region, but also to covariance among multiple regions [104].

However, genetic studies of neuroanatomy using magnetic resonance imaging (MRI) continue to prioritize morphological measures on specific landmark-defined brain regions, such as the volumes of subcortical nuclei [13] or average thickness of cortical parcellations [106]. Although this approach captures some genetic effects of structural variations, it bypasses the fact that the morphological state of an adult brain is the sum of previous developmental processes across brain regions. These landmark-defined regions of interest (ROIs) therefore may have lost genetically relevant information by ignoring co-varied components, while concurrently introducing irrelevant variance by combining measures from genetically unrelated neighbors [107].

TThe limitations of this ROI approach are most evident in the context of studying effects on neurodevelopment, as the age-dependent processes have been shown to consist of a gradient spreading across the cortical surface without a discernable relationship to traditional anatomical landmarks [50]. Past efforts to redefine the imaging phenotypes beyond landmark-based ROIs include learning a sparse representation from patients with Alzheimers disease [107] or redrawing ROIs based on the genetic correlations from twin studies [19, 10]. These methods can be conceptualized as projecting the multidimensional measures of MRI onto a lower dimensional axis while filtering out components irrelevant

to the genetic signals. Such methods have seldom focused, however, on neurodevelopmental disorders, such as Williams Syndrome (WS), that have larger neuroanatomical impacts and more finite candidate genetic regions attributable to the neuroanatomical differences. Since statistical power is the most critical factor for identifying genes through associations [108], a redefined MRI measure that contains more relevant genetic signals and reduces the burden of multiple comparisons can greatly facilitate the discovery of neurodevelopmental genes.

WS is a multi-systemic disorder caused by hemi-deletion of roughly 27 genes on chromosome 7, resulting in cardiovascular morbidities, intellectual impairment, and hypersociability [95, 72]. Besides a decrease of about 11% in brain size, patients with WS have aberrant regionalization of cortical surfaces as assessed with brain MRI, particularly in superior parietal regions and the orbitofrontal cortex [76, 75, 77, 78, 96]. Animal models have suggested GTF2IRD1, a gene-encoded general transcription factor, as one of the most promising candidate genes for neuroanatomical differences in WS [100, 109, 101, 84]. Genetic perturbations on *GTF2IRD1* have recently been associated with dog friendliness toward humans [110].Despite such findings in animal models, associations of this gene with brain or behavioral phenotypes in the healthy human population are lacking. Without association studies on brain phenotypes in healthy human populations, it remains unclear whether common genetic variants on those genes have an impact on typical brain development.

Here, we describe a novel two-pronged approach to capturing genetic effects on neurodevelopment. First, using one single score to represent the global neuroanatomical variations, and a candidate genes approach by examining only the WS region, we limit the effect-size requirements imposed by Bonferroni correction. Second, and more important, we increase the sensitivity of the anatomical phenotype by using a single derived score calculated from multidimensional MRI measures. In our previous work,

we derived a single global measure that characterizes how WS brains are structurally different from controls, across multiple parameters in multiple locations [85]. In this study, we demonstrate that the WS neuroanatomical score can be regarded as an MRI endophenotype, enriched in genetic information pertaining to neurodevelopment. By applying the neuroanatomical scores to five imaging genetic cohorts with brain MRI and single nucleotide polymorphisms (SNP) data (n = 1863 healthy European descent), we demonstrate, for the first time, that a common variant in *GTF2IRD1* is associated with variation in brain structure (Bonferroni corrected p = 0.023). The genetic signals are more enriched than traditionally defined ROI and have significantly high SNP-heritability (h2 = 0.82, se = 0.25, p = 5e-4). Our results provide a proof of concept for the strategy of using multivariate structural measures as a derived intermediate phenotype for genetic association studies.

5.2 Methods

5.2.1 Participants

We selected 1,863 healthy imaging genetics subjects from five independent cohorts: 184 were from the Alzheimers Disease Neuroimaging Initiative (ADNI) [111], 653 were from the Nord-Trndelag Health Study (HUNT) [112], 325 were from the Norwegian Cognitive NeuroGenetics (NCNG) [113], 250 were from the Thematically Organized Psychosis study (TOP) [114], and 451 were from the Pediatric Imaging Neurocognition and Genetics Study (PING) [115]. From each study, only healthy, unrelated, European ancestry subjects were retained for analysis. Because the WS neuroanatomical scores were nevertheless trained on an adult WS cohort [116], the residual confounding of age effect might have an impact on the association. Given that the PING study contains the youngest individuals across all cohorts, we further stratified the PING sample into

two subcohorts, one for those ages 16 years and older, and the other for those younger than 16. Each study collected 3D T1 MRI images according to comparable acquisition protocols and was processed with the same FreeSurfer reconstruction protocols. Whole genome genotypes were imputed according to the same Mach/Minimac procedure using the 1000 Genomes Project as a reference. Estimated dosages of 110 SNPs falling within the WS hemi-deletion region (chromosome 7q11.23, 72Mb-74Mb, hg19) were imputed with good quality in all cohorts and selected for analysis.

5.2.2 Derivation of the Williams Syndrome Neuroanatomical Scores

We used a penalized regression model to calculate WS neuroanatomical scores given individuals MRI measures. Full details of the training and validation of the model have been published elsewhere [85]. Briefly, 3D T1 MRI images were obtained on 22 Williams Syndrome patients and 16 healthy controls. A multivariate regularized logistic regression was trained to discriminate WS patients from healthy controls on the basis of 30760 predictors, including estimated cortical surface area [9], cortical surface geometry [85], and sulcal depths [77] for each of 5124 reconstructed vertices and the volumes of 16 subcortical structures [8]. In order to capture the subtle morphological reorganizations of the WS brain, intra-cranial volumes (ICV) was used as a covariate to ensure overall brain size was not driving the classification. For each subject in our healthy imaging genetics cohort, we applied the resulting discriminative weights to the same neuroimaging feature space, summarizing this high dimensional data with a single, composite neuroanatomical score reflecting their morphological variations on the axis between healthy individuals and patients with WS. Figure 5.1 illustrates the flowchart of the analytic strategy and visualization of the weights for contributing neuroimaging measure to the final composite scores.

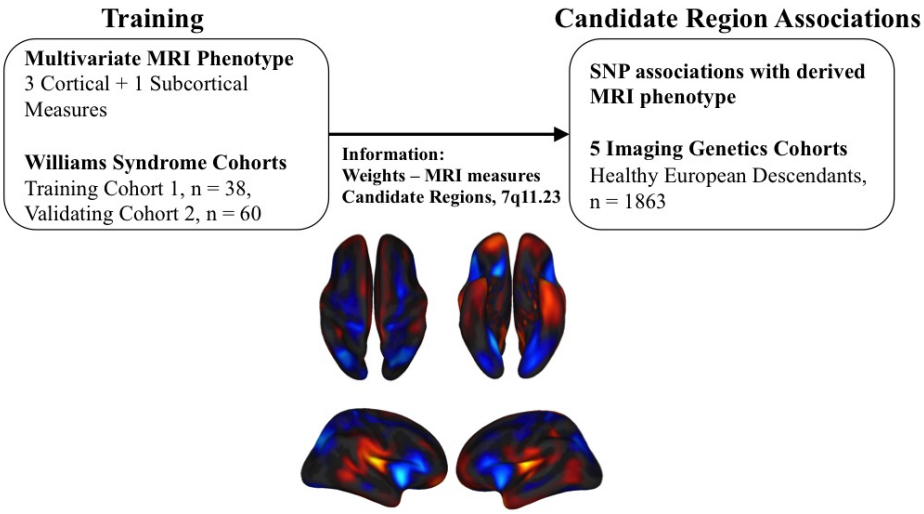

Figure 5.1: Flow chart of the study design. The first stage of the analysis (Training) was deriving neuroanatomical scores based on case-control data, which has been published elsewhere [116]. The second stage of the analysis (Candidate Region Associations) is the focus of this paper, wherein we directly apply the neuroanatomical scores from large scale imaging genetic cohorts without further calibration of the model parameters.

5.2.3 Candidate Region Association Analysis

Each imputed SNP dosage was regressed against the composite WS neuroanatomical score while controlling age, age squared, gender and the first seven principle components of genetic ancestry as potentially confounding covariates. For each SNP effect was estimated in each cohort separately and combined post-hoc according to an inverse variance weighted meta-analysis implemented in PLINK. To account for multiple comparisons, we used a Bonferroni adjustment for the 110 linked SNPs. Our significance threshold was set to p ¡ 0.05/110 = 0.00045, conservatively controlling for 110 correlated

tests. We then used CAVIAR to determine which SNP is the potential causal variant [117].

5.2.4 Local Enrichment and Global SNP Heritability

To demonstrate the enrichment of local genetic signals by using newly defined WS neuroanatomical scores, we performed the quantile-quantile plot comparing log10(p) between our SNP associations in the WS chromosomal regions and summary statistics of ROI approach from ENIGMA consortium (n = 12,596) 5. Despite of the scale of our imaging genetic cohorts, the sample size is considered as modest in the context of genome-wide association studies. Therefore, to prevent under powered genome wide analyses while quantifying the global genetic signals of WS neuroanatomical score, we used Genome-wide Complex Trait Analysis (GCTA) [118] to estimate the variance explained by all of the SNPs on the entire genome (i.e., the SNP-heritability).

5.3 Results

The training and validating of WS neuroanatomical scores have been published elsewhere [116]. In short, the derived neuroanatomical scores robustly distinguished WS from other groups in both the training set (leave-one-out cross-validation area under curve as 100%) and the validating set (area under curve as 100%). The composite WS score significantly mediates the cognitive differences between cases and controls, especially tests quantifying social behaviors 25. Having derived this multivariate measure which characterizes WS, we then applied the score to healthy imaging genomic cohorts. Each healthy individuals MRI measures were combined into one single score given the derived weights of WS neuroanatomical score. The score of cohort members is normally distributed and not correlated with genetic ancestry.

The associations between SNPs and neuroanatomical score in imaging genomic cohorts are shown in Figure 5.2 and Figure 5.3. One locus containing 3 SNPs located at *GTF2IRD1* showed statistical significance after Bonferroni correction (Figure 5.2, top SNP, rs2267824, corrected p = 0.023). Effect sizes of the associated SNP were consistent across cohorts (Figure 5.3) except for the cohort with individuals younger than 16 years old. After excluding individuals younger than 16 years old, the association of rs2267824 became stronger (reference allele: C, coefficient: 0.018, corrected p = 5.5e-3). CAVIAR confirmed that the region contains one single locus and rs2267824 was the potential causal variant. In addition, one SNP within 250kb of *FZF9* showed nominal significance (rs2237280, uncorrected p = 0.00627).

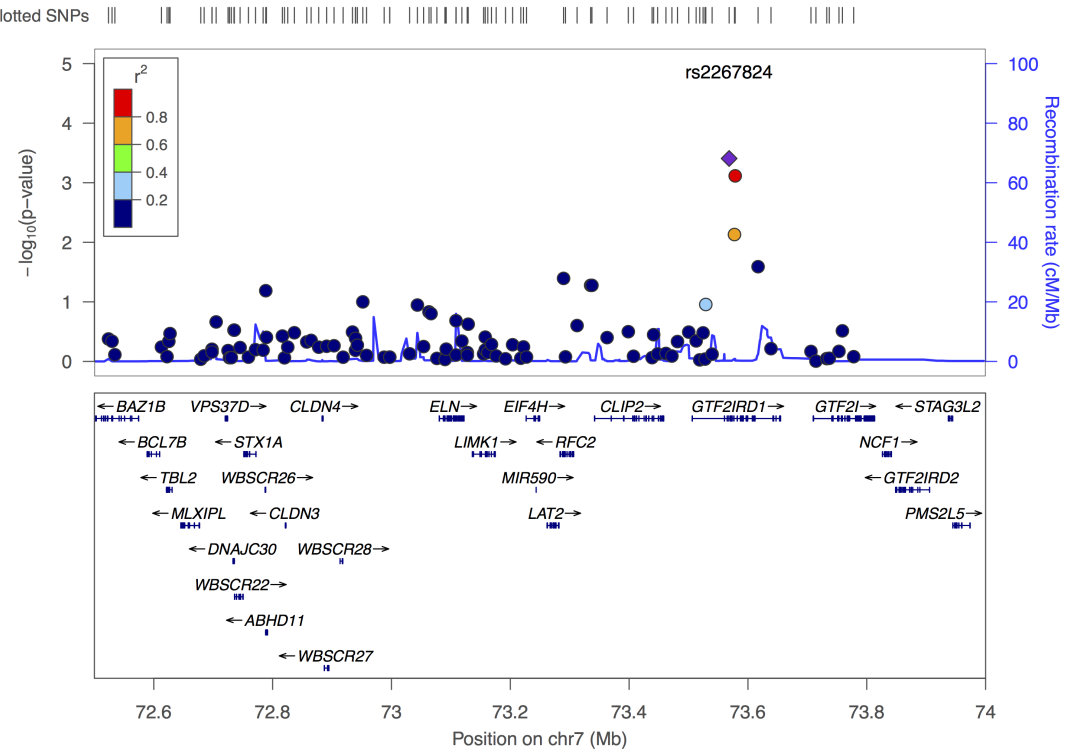

Figure 5.2: Regional plot of the associations between SNP dosage and WS neuroanatomical scores. The results of 110 SNP associations were plotted against gene annotations and physical positions. The coloring of each SNP represents the linkage disequilibrium with the top SNP, rs2267824.

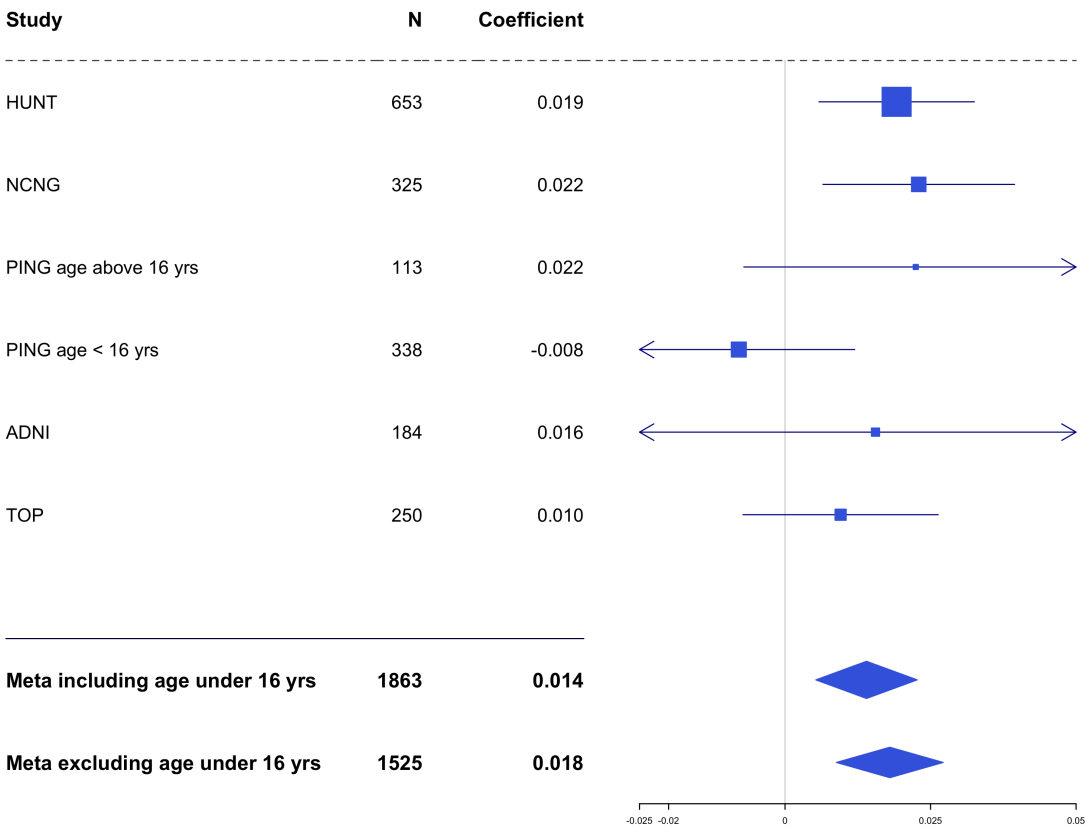

Figure 5.3: Meta-analysis and stratified analyses of the associations with rs2267824. The reference allele is set as C while the coefficients were unitless, as the WS neuroanatomical scores were similarity measures range from 0 to 1.

The quantile-quantile plots compared with associations from the ENIGMA study demonstrated significantly enriched genetic signals in the WS chromosomal regions when using the WS neuroanatomical score (Figure 5.4). In terms of global genetic signals, the WS neuroanatomical score has high heritability (h2 = 0.82, se = 0.25, p = 5e-4) despite the fact that less than one percent of phenotypic variation can be explained by the potential causal SNP, rs2267824.

5.4 Discussion

Here we demonstrate that the WS neuroanatomical score can be regarded as an MRI endophenotype, enriched in genetic information pertaining to neurodevelopment. By applying the neuroanatomical score to five imaging genetic cohorts, we show that a common variant in GTF2IRD1 is associated with variation in brain structure. The genetic signals were more enriched than traditionally defined ROI and have significantly high SNP-heritability. Our results provide a proof of concept for the strategy of using multivariate structural measures as a derived intermediate phenotype for genetic association studies. An optimized multivariate MRI procedure defines the intermediate phenotype that can accurately capture the continuous nature of the underlying brain variations, thus providing greater power for detecting genetic associations.

The associations between GTF2IRD1 and the WS neuroanatomical score support a critical role of this general transcription factor for normal brain development, and specifically for one of the characteristic personality traits of WS. WS has a unique neuroarchitecture compared to other developmental disorders with intellectual impairment, but few studies have tied anatomical changes to strikingly heightened social behavior [74, 76, 77, 78, 88]. Previous case studies of partial hemi-deletions in WS indicate that the region telomeric to 7q11.23, which includes *GTF2IRD1*, is crucial for the changes in social behaviors characteristic of WS [100, 84, 119]. Animal models also support the role of textitGTF2IRD1 in brain development [100, 109, 101]. In particular, a recent study on dog friendliness found the genetic variations on *GTF2IRD1* and textitGTF2I were positively selected for the tendency to socially engage with humans [110]. Our results provide converging evidence for the role of *GTF2IRD1* in human brain development and social cognition.

It is likely that other genes also affect the neuroanatomical profiles we defined

here, and they may act synergistically in producing the observed phenotype. For example, a study of neuron-like cells derived from stem cells in WS demonstrated reduced neuron proliferation and enhanced dendritic elaboration resulting from the perturbation on *FZD9* [102]. As our associations found a suggestive signal located at the *FZD9*, although much weaker than the main *GTF2IRD1* effects, it nevertheless jointly contributed to the variations in neuroanatomical profiles. This interpretation is supported by the effects of partial hemi-deletions which spare the *FZD9* gene [102, 84, 116]. We found that although WS neuroanatomical scores were increased among these subjects, it is much weaker than in those with a typical hemideletion [116]. Further evidence for synergistic effects were found in studies implicating both *GTF2IRD1* and *FZD9* in the Wnt pathway [105, 120, 121].

In addition, we found significantly high heritability of the observed variations in our defined neuroanatomical score, indicating polygenic contributions. Although the neuroanatomical scores were highly specific to WS status among patient groups [116], the variations in scores among healthy adults can represent the accumulation of multiple developmental processes with diverse genetic perturbations, each with small effects. This phenomenon is compatible with the theory of the modularized genetic networks in which canalized phenotypes, e.g. typically developed brains, can tolerate many small genetic perturbations unless genetic hubs are drastically disturbed [81, 4]. In this framework, the WS deletions would represent a large perturbation of a neurodevelopmental process which in typical developed individuals only shows small variations attributable to regulatory genes across the genome. Although our WS neuroanatomical scores were enriched for WS relevant genetic effects, it nevertheless characterized an underlying canalized developmental process. Using our analytic strategy with diverse genetic developmental disorders may provide further insight into this enduring question about phenotype-genotype mapping.

In sum, our results provide further support for the role of *GTF2IRD1* in the Williams Syndrome phenotype and a proof of concept for deriving multivariate MRI phenotypes for genotype-phenotype studies. This strategy may prove useful in other neurodevelopmental disorders that typically have restricted genetic deletions or alterations. In addition, more accurate measurement of the neuroanatomical phenotype should also provide greater power for genetic studies of diseases such as schizophrenia and autism spectrum disorders where the genetic basis is distributed across the genome, and should ultimately facilitate the discovery of other mediating paths from genes to disorders.

5.5 Acknowledgement

Chapter 5, in full, is being prepared for submission for publication. Chun Chieh Fan, Andrew J. Schork, Timothy T. Brown, Barbara E. Spencer, Natacha Akshoomoff, Chi-Hua Chen, Joshua M. Kuperman, Donald J. Hagler Jr., Asta Kristine Hberg, Thomas Espeseth, Ole A. Andreassen, Anders M. Dale, Terry L. Jernigan, Eric Halgren. The dissertation author was the primary investigator and author of this paper.

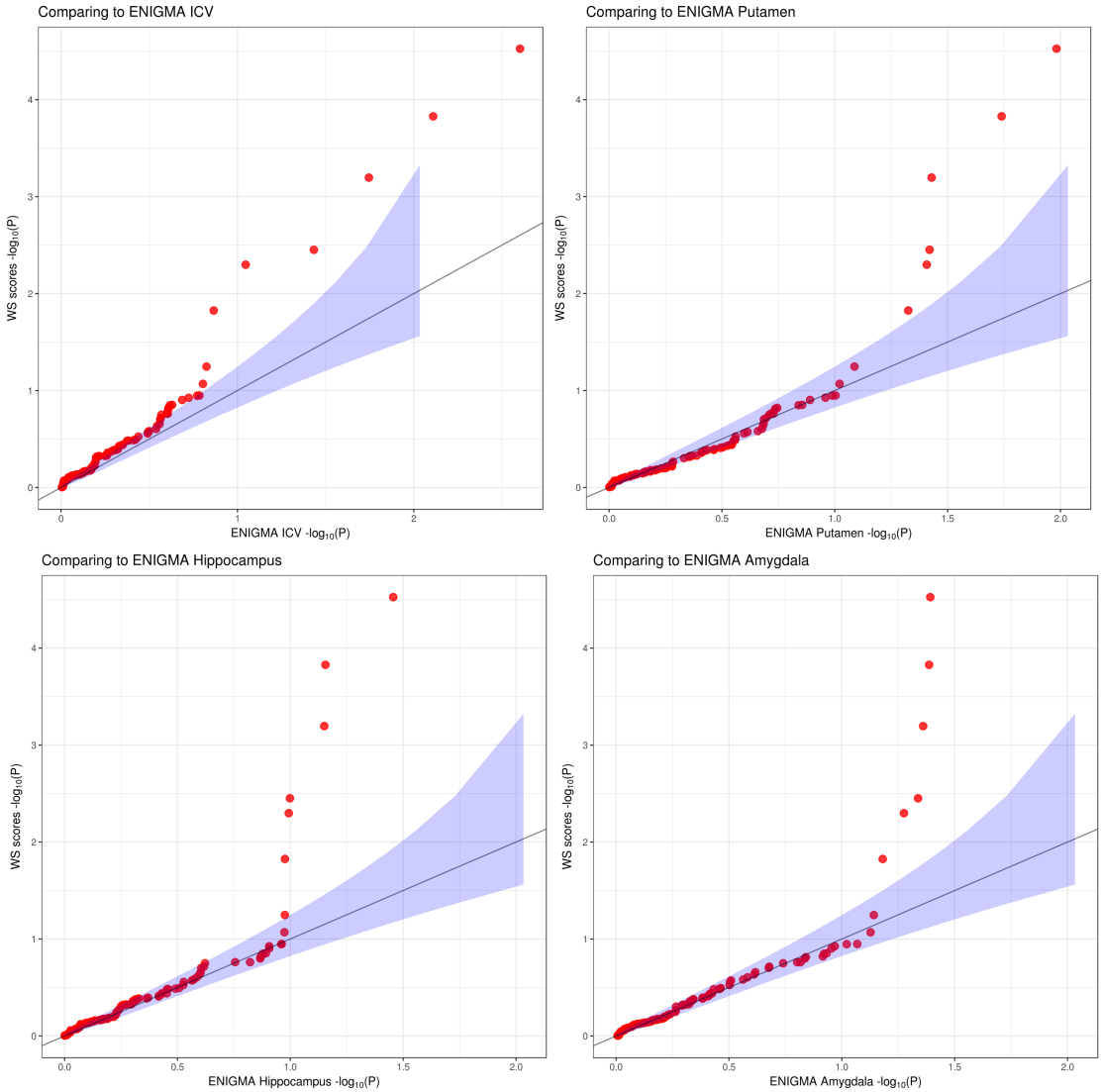

Figure 5.4: Local enrichment of genetic signals comparing to ENIGMA summary statistics. Quantile-quantile plots were based on the comparison between our results and summary statistics from ENIGMA study. Only SNP associations from the WS chromosomal regions were included in this analysis. Despite ENIGMA is almost 10 fold larger sample size than our current study, the genetic signals were enriched in our analyses as the tails of quantile-quantile plot significantly deviate away from the expected null. Upper left, compared to associations between SNPs of WS chromosomal regions and intra-cranial volumes (ICV) in ENIGMA. Upper right, compared to associations between SNPs of WS chromosomal regions and Putamen volumes in ENIGMA. Lower left, compared to associations between SNPs of WS chromosomal regions and hippocampal volumes in ENIGMA. Lower right, compared to associations between SNPs of WS chromosomal regions and amygdala volumes in ENIGMA.

Chapter 6

Determining the tree-structured topology of the human cortical surface from vertex-based genome-wide association study summary statistics

6.1 Introduction

Characterizing the influence of genetic variation on the organization of the human brain is critical for understanding the biological mechanisms controlling neural development and brain-related disease susceptibilities [122]. Although, structural features of the brain as measured by, e.g., magnetic resonance imaging (MRI), have been shown to be heritable [123, 124, 11], modularly organized [19, 10], and close to pathological process of diseases [96], determining which regions are under common genetic control, and to what degree, are not trivial to address. With the advent of genome-wide association studies (GWAS), it has been possible to identify genetic variants that are associated with

MRI-determined features of the brain [14]. However, due to the inherent multidimensional nature of data generated by MRI, which often involves thousands of sampling points (vertices/voxels) on cortical surface, researchers often reduce the data and define broad regions of interest (ROIs) to focus studies and reduce the burden of multiple hypothesis testing on statistical power for detecting associations. The highly multidimensional nature of MRI data is further compounded in genetic studies, particularly GWAS, where one may want to test millions of variants for association with MRI-derived phenotypes. Many strategies have been proposed for reducing this burden, including averaging over a set of MRI vertices from anatomically pre-defined regions [125, 12], using penalized regressions to jointly model SNP effects on clusters of vertices [20, 21], and applying polygenic scores from GWAS to specific MRI-derived phenotypes [126, 127]. These analysis methods all require the use of data from each individual, which can not only be hard to obtain if one wants to combine multiple data sets to increase power, but can also be computational challenging if one wants to relate all the genetic variants to each of the voxels (or subsets of them) simultaneously [14]. In addition, analysis methods, such as partial least squares and canonical correlation analysis methods, that try to relate very large sets of independent variables (such as genetic variants in a GWAS) to many dependent variables (such as individual voxels from an MRI analysis) are often very difficult to interpret, hard to generalize, and even more difficult to replicate given uncertainty in relevant parameter estimates without extremely large sample sizes. Since it is often the case that summary statistics of GWAS results, possibly applied to each voxel, are available or can be generated quite easily using standard GWAS analysis methods, analysis strategies that take require only summary statistics are an attractive alternative.

Leveraging GWAS summary statistics to look for patterns across studies focusing on different phenotypes is not entirely new. In fact, many methods to pool summary statistics from GWAS across multiple traits have been proposed [128, 129, 130, 131].

Unfortunately, many of the methods that combined summary statistics from GWAS studies rely on different assumptions about the relationships between the variants and the phenotypes of interest, must account for linkage disequilibrium (LD) between the variants (often Single Nucleotide Polymorphisms (SNPs)), and often have poor power to detect associations across the studies [132, 133, 134, 135]. Thus, if used in the characterization of genetically-mediation neuroanatomical patterns in the brain based on MRI data they could result in neuroanatomical topologies that are highly uncertain and lack confidence. This is particularly problematic given that most imaging genetics studies that combine GWAS data with MRI data since available summary statistics with sample sizes larger than 10,000 are derived from studies that focused on defined metrics and brain regions of interest, such as hippocampal volumes [13].

We have developed a novel and intuitive analytic framework that leverages summary statistics from voxel-based MRI GWAS to identify and characterize the genetically-mediated neuroanotomical topology of the human cortical surface. Our methods rely on the use a weighted pairwise Euclidean distance measure applied to z scores obtained from summary statistics for individual SNP associations from GWAS on individual voxels as a metrics for the genetic distance between voxels. These pairwise distances are then used in a unique hierarchical clustering scheme that can reveal a tree-structured topology of the voxels that can be refined to reveal groups of voxels that appear to be under common genetic control. We have validated the approach through simulation studies, we show that it can robustly recover pre-specified clusters of voxels even when the sample sizes for the GWAS and MRI data are modest. We ultimately applied our method to a large (n=1429) set of vertex-wise GWAS studies and revealed a compelling genetically-mediated topology of the human cortical surface. The method can easily be used to address other questions surrounding the organization of the brain and can also be used to identify genetically-mediated patterns in any setting in which a large number of

GWAS studies have been pursued on different phenotypes.

6.2 Material and methods

We describe the proposed method below in sections addressing: 1. The analytical technique for quantifying the distance between voxels based on GWAS summary statistic data; 2. The strategy for clustering the voxels; 3. The simulation studies use to assessment the power and robustness of the methods; and 4. Our strategy for applying to actual MRI voxel-based GWAS summary statistic data to identify a genetically-mediated neuroanatomical topology of the human cortical surface.

6.2.1 Weighted Euclidean distance for summary Z-statistics from voxel-based GWAS.

We start by noting that our formation of a Euclidean distance between summary-level Z-statistics associated with a SNP assessed in two different GWAS of two voxels is closely related to the polygenic model LD score regression and mixed-effects SNP-heritability models for exploring individual SNP-based genetic correlation methods as outlined in many different publications [134, 132, 133, 135]. Essentially, assuming that, for a given SNP j (j=1,,L), where L is the total number of SNPs in GWAS applied to two different phenotypes (i.e., vertex in our case), has an effect on the two vertices that follows bivariate normal distribution with mean vector and covariance matrix of the form:

$$N(0, \begin{bmatrix} \sigma^2_{1j} & \rho_{12j} \\ \rho_{12j} & \sigma^2_{2j} \end{bmatrix}) \tag{6.1}$$

Where p12j is the correlation between the voxels attributable to the SNP. The expected value of this Euclidean distance between individual Z-scores for vertex 1 (Z1)

and 2 (Z2) can then be expressed as the following:

$$\begin{aligned} & E[\sum_j (Z_{1j} - Z_{2j})^2] \\ & = \sum_j E[Z_{1j}^2] + E[Z_{2j}^2] - 2E[Z_{1j}, Z_{2j}] \\ & = B + n\sum_j (\sigma_{1j}^2 + \sigma_{2j}^2 - 2\rho_{12j}) \end{aligned} \tag{6.2}$$

Where B is a bias term and n is the sample size. Because the Z-statistics from the GWAS do not necessarily represent independent causal effect estimates of the SNPs but can be correlated due to potential underlying linkage-disequilibrium (LD) between the SNPs, we weighted each of the Z-scores for the two voxels based on their degree of LD in order to approximate the independent contribution of the SNPs [135]. To achieve this, we used the LDAK software to calculate the weights for each SNP [134]. The benefit of using the Euclidean distance measure outlined above is that we do not need to estimate the shared variance and covariance across SNPs to arrive at a measure of genetic correlation as with many other methods. Rather, we only need a distance metric to define and determine the degree of similarity across the genetic associations in the GWAS for each of the voxels in a pair.

6.2.2 Procedures to determine tree-structured genetic topologies

With the genetic distance defined for each pair of V vertices (where V is the total number of voxels considered in a study for which there are GWAS summary Z-statistics for each SNP), we can construct a V x V distance matrix that represents the Z-statistic-based GWAS distance between each pair of vertices. Hierarchical clustering can then be applied to this distance matrix to identify a tree-structured topology of the

voxels considered in the analyses. Because hierarchical clustering builds a tree structure iteratively by merging each vertex with its most similar neighbor, this building process when applied to the matrix of distances defined above is equivalent to finding the vertiex neighbor, i, of vertex j, based on:

$$i = \underset{i}{\operatorname{argmin}} \sum_{j} \sigma_{2j}^2 - 2\rho_{12j} \tag{6.3}$$

As such, traversing the resulting hierarchical tree from the top to the bottom represents moving from those vertices with the greatest shared (or similar) genetic influences, to those with the lowest shared (or least similar) genetic influences. To determine how many unique clusters of vertices with shared genetic association and association strength levels there might be, we used gap statistics [136].

6.2.3 Simulation studies

To simulate realistic settings in which there are some number of clusters of vertices with common genetic determinants, we had to make some assumptions because there are an infinite number of possible scenarios we could have explored. As a result, we focused on a few simple settings that demonstrate the effectiveness of the proposed technique to recapitulate known genetically-mediated topologies (i.e., clusters). We repeated simulations to assess the reliability and variability of the proposed techniques ability to recover a known topology. For each of repeated simulations, we randomly generated a voxel-based topology with two distinct genetically-mediated clusters where each of M total SNPs effect sizes (associated with the SNP ms (m=1,,M) Z-score) on the vertices were randomly assigned for each of the two components, k (k=1,2) as:

$$N(0, \frac{h_k^2}{m_k}) \tag{6.4}$$

We used HAPGEN2 [137] to randomly generate 1500 individuals with L total genotypes with realistic LD patterns derived from the LD patterns exhibited by the data on the European populations from 1000 genome project (phase 3 data) [138]. We pursued 1000 total simulations while varying the heritability of each the two topological components from 0.01 to 0.07. These ranges of heritability are consistent with SNP-based heritability estimates from individual-level random effects models [139]. We used Jaccard index [140] to evaluate the accuracy of recovered topologies for each simulation. We considered using an unweighted Euclidean distance measure in our simulation studies to determine what difference the weighting might make. In addition, in order to contrast the ability of the proposed method (with and without weighting) to recover known topologies with other methods designed to quantify genetic correlations between sets of phenotypes, we also applied the LD score regression method to the simulated data [132].

6.2.4 Empirical application to imaging genetic cohorts

To identify and characterize an actual genetically-mediated neuroanatomical topology of the human cortical surface, we applied our methods to data from three published imaging genetic cohort studies: PING [115], HUNT [112], and NCNG [113]. Combining data from these three independent cohorts, a total of n=1429 individuals of European descent with typical development were used in the analyses. We note that all three cohorts used the same MRI processing and quality control procedures for both the MRI measures, genotyping and genotype imputation strategies. In particular, the 3D reconstructed cortical surfaces were registered to a common spherical coordinates to ensure each vertex on cortical surface was compatible across all the individuals [8, 9]). To derive summary statistics from the GWAS on each vertex, we used a standard linear model relating genotype dosages for the imputed genotypes as well as covariates (the independent variables) to cortical surface voxels (the dependent variables). We chose two

widely used cortical surface measures for the vertex-specific GWAS: cortical surface area and cortical surface thickness. The covariates included in the association analyses for each SNP and voxel were age, age-squared, gender, and first five genetic principle components from a genetic relationship matrix (GRM) created for all subjects to accommodate any cryptic stratification and differences in ancestry among the individuals in the analyses. For the voxel-based analyses, we down sampled the voxels to 642 points per hemisphere using matrix calculations described in MatrixEQTL [141].

6.3 Results

6.3.1 Simulation Studies

The simulation results are depicted in Figure 6.1. The proposed procedure, using both weighted and unweighted Euclidean distances, reaches 100 percent accuracy in recovering the known topology. It is notable that in recovering the known topology by the propose methods there were false positives when heritability was 0.04 or greater. The weighted Euclidean distance provides better accuracy than the non-weighted Euclidean distance method when the heritability was less than 0.04. Interestingly, the use of the well-published LD score regression technique for characterizing the genetic correlation between voxels performed poorly relative to the proposed method.

6.3.2 Characterizing a genetically-mediated human cortical surface neuroanatomical topology

After applying the proposed method to the MRI data on the 1429 individuals from the three cohorts discussed in the Materials and Methods section, we uncovered a compelling tree-structured topology of cortical surface area, as illustrated in Figure

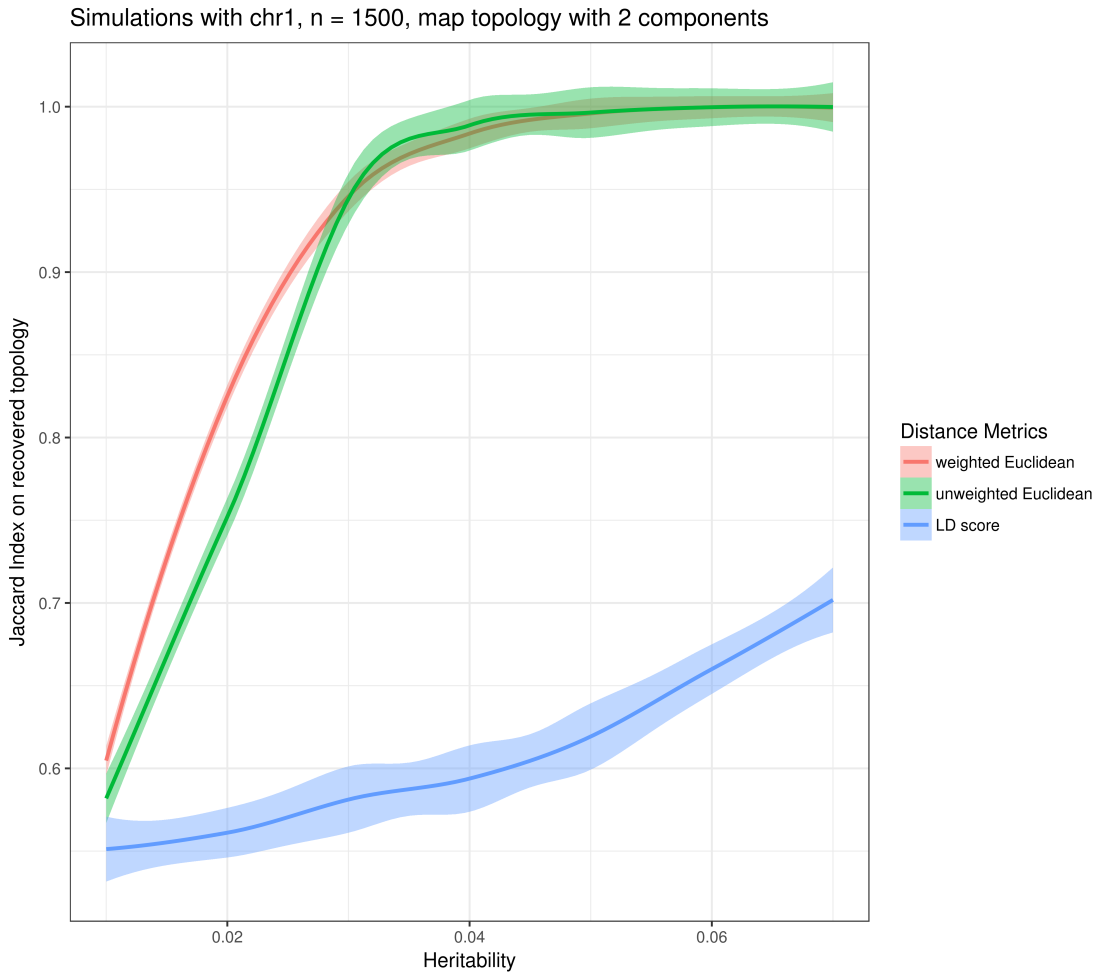

Figure 6.1: Simulation results from 1000 iterations with randomly generated two components topology. Each of iterations generated 1500 diploids of chromosome one based on the reference panel. The same hierarchical clustering processes were performed for each of different distance metrics.

6.2. Of 1284 sampled vertices, 14 clusters were identified based on gap statistics (Figure 6.2, lower left). In general, the genetic modules of cortical surface area were bilaterally symmetric. The first tree bifurcation divides subcortical regions, and the second differentiates the anterior-posterior axis (Figure 6.2, subfigure 1 and 2). In addition, lateralization was also observed in the final cluster results (Figure 6.2, lower right). The left hemisphere includes three additional clusters that corresponding to the Wernickes area (Figure 6.2, region A), Geschwinds area (Figure 6.2, region B), and

Brocas area (Figure 6.2, region H). The cluster results also grouped anterior cingulate, insular, and orbitofrontal cortex as one single genetic module (Figure 6.2, region G) which suggests that the recovered genetically-mediated regions of the cortical surface are consistent with a great deal of known biology. For the analysis of the cortical thickness data, Figure 6.3 illustrates the results. Four genetically-mediated clusters were identified (Figure 6.3, lower left). Similar to the cortical surface area, the first tree bifurcation differentiates the subcortical region (Figure 6.3, subfigure 1) and the identified clusters were bilaterally symmetric. The other clusters clearly suggest a top-down gradient m which again is intuitive (Figure 6.3, regions A, B, C).

6.4 Discussion

Using our simple and intuitive proposed approach for identifying and characterizing genetically-mediated neuroanatomical clustering in the cortical surface of the human brain, we identified a compelling hierarchical structure of cortical brain regions. Our proposed method differs from other methods for exploring the shared genetic determinants of multiple phenotypes in that it does not rely on estimating genetic correlation. Rather our approach relies on a weighted Euclidean distance of Z-statistics obtained from GWAS on individual vertices from cohort studies with MRI data. Our method can recover tree-structured genetically-mediated topologies even when the sample size is moderate to small. The neuroanatomical regions clustering together and the hierarchical relationships between them revealed by our method coincide with known functional domains, and suggests shared genetic influences on language and social processing centers of the brain.

We also contrasted our proposed method with a traditional method for identifying genetic correlations between different phenotypes and show that this method does not perform as well (Figure 6.1). This suggests that our proposed method could have broader

application than characterizing the genetically-mediated topology of human cortical surface. For example, our technique could be used to find more genetically homogeneous groups of traits in any GWAS settings (e.g., those considering the shared genetic influences of many different diseases or phenotypic features for which independent GWAS have been pursued). These more homogeneous groups could then be exploited by combining them to increase the sample size to identify individual loci that might influence them. In this light, improving statistical power for identifying causal loci by pooling multiple genetic association analyses depends on the underlying shared genetic effects across traits considered in an analysis [7]. Pooling methods for summary statistics [129, 131, 130] without a way to define the genetically homogeneous group could have limited benefit since the resulting combination of GWAS may include traits that are not likely to share genetic determinants. Our approach provides a way to generate the genetic clusters first using summary statistics and then picking out the traits that it makes most sense to combine to increase power to identify individual loci that influence them all.

In terms of the neuroanatomical topology of the human cortex that we identified, both the topologies for cortical surface area and thickness shared similar characteristics with the topologies derived from, e. g., twin studies. Bilateral symmetry, anterior-posterior differentiation in surface area, and top-down gradient in surface thickness have been noted in previous clustering results based on twin genetic correlations [19, 10]. However, the genetic lateralization of cortical surface area that we identified has not been shown before. Our results also suggest a genetic basis for language development, as the Wernicks area, Geschweinds area, and Brocas area have separate set of genetic influences. The regions related to social processing [98], i.e., anterior cingulate, insular, and orbitofrontal cortex, are also clustered as one genetically homogeneous module by our method. This suggests the genetic influences on the structural variation of cortical surface follow known functional domains of human cognition. It remains to be seen

which groups of genes are involved in each functional domain to a greater degree than others. Our results nevertheless provide a first glance into the genetic influence on the functional domains of human cognition.

Considering the modest sample size used in our empirical study of the human cortical surface area, our study is more of a proof-of-concept than a definitive assessment of the genetically-mediated topology of human cortex. A larger sample size would allow for greater precision in resolving genetically-mediated topological maps of the brain [94]. Given that the summary statistics are easier to share, as more researchers are willing to share summary statistics of multiple traits, our method has great promise in analyzing summary data provided to the community in the future. For example, the ENIGMA project has influenced the community to use the same MRI data processing pipeline to ensure consistency when meta-analyses are pursued with the data [14]. Voxel-wise GWAS will greatly benefit from similar thinking, using the same registration process to ensure comparable voxels when GWAS are performed. Our proposed method can be extended in at least one important way. The method can incorporate prior information about the SNP effect sizes, such as functional annotations [142] or pathogenic scores of SNP [143] to weight the SNPs when the distance function is computed between two vertices [135]. Ultimately, we have developed a simple and intuitive way of clustering vertices to identify important and biologically-relevant patterns and fell that its application can shed light on important aspects of the way the brain is organized. We also believe that the proposed method can motivate the development of more powerful strategies for analyzing high-dimensional data in efficient ways.

6.5 Acknowledgement

Chapter 6, in full, is being prepared for submission for publication. Chun Chieh Fan, Andrew J. Schork, Westly K. Thompson, Asta Kristine Hberg, Thomas Espeseth, Ole A. Andreassen, Anders M. Dale, Terry L. Jernigan, Nicholas J. Schork. The dissertation author was the primary investigator and author of this paper.

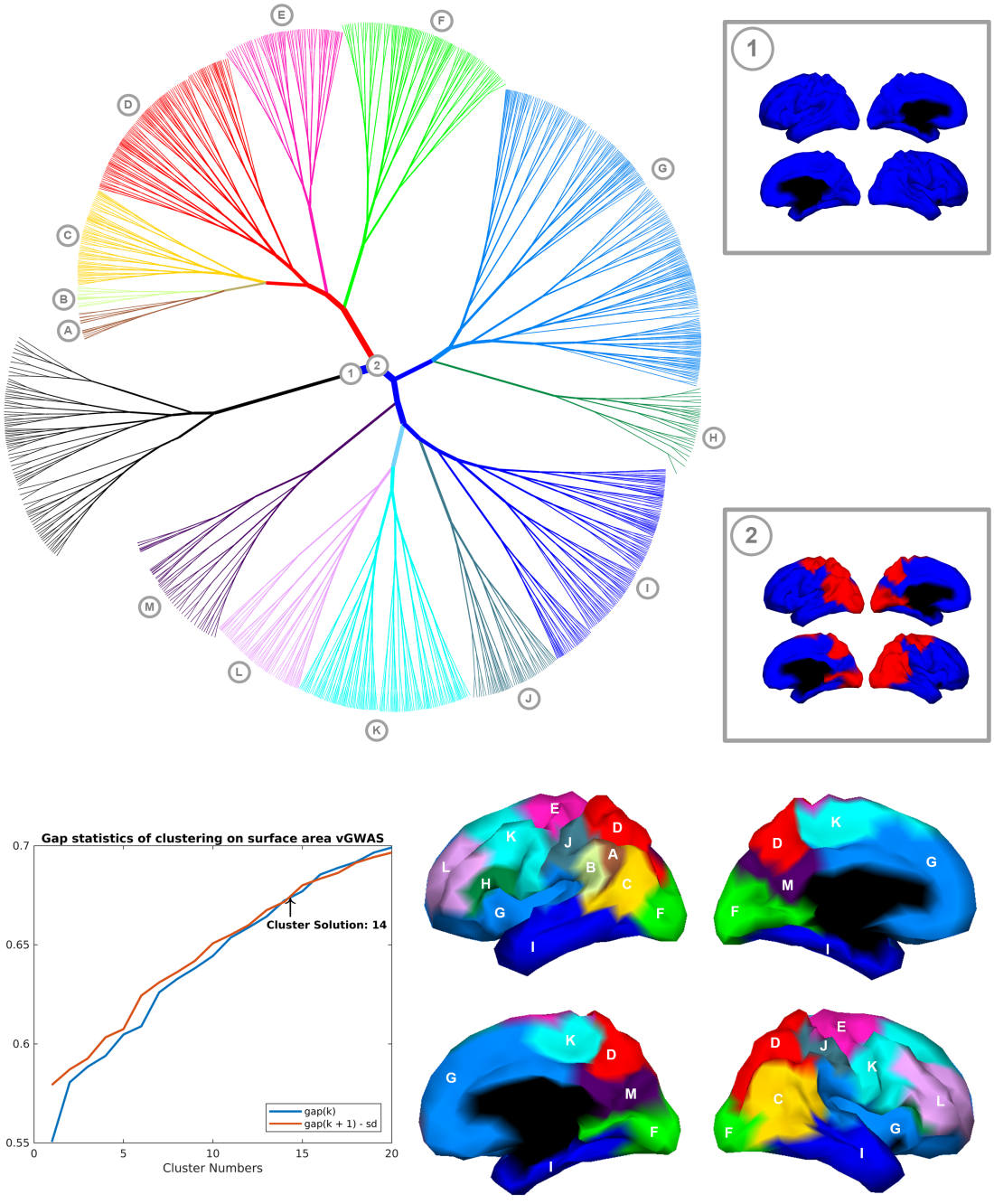

Figure 6.2: The tree structure is based on the dendrogram of hierarchical clustering while the cluster results were based on the gap statistics. Each of the four identified clusters except subcortical regions were labeled in different colors and labeled as A to C. The first two bifurcating points were demonstrated in subfigure 1 and 2. The final clusters were visualized in the lower right of the figure.

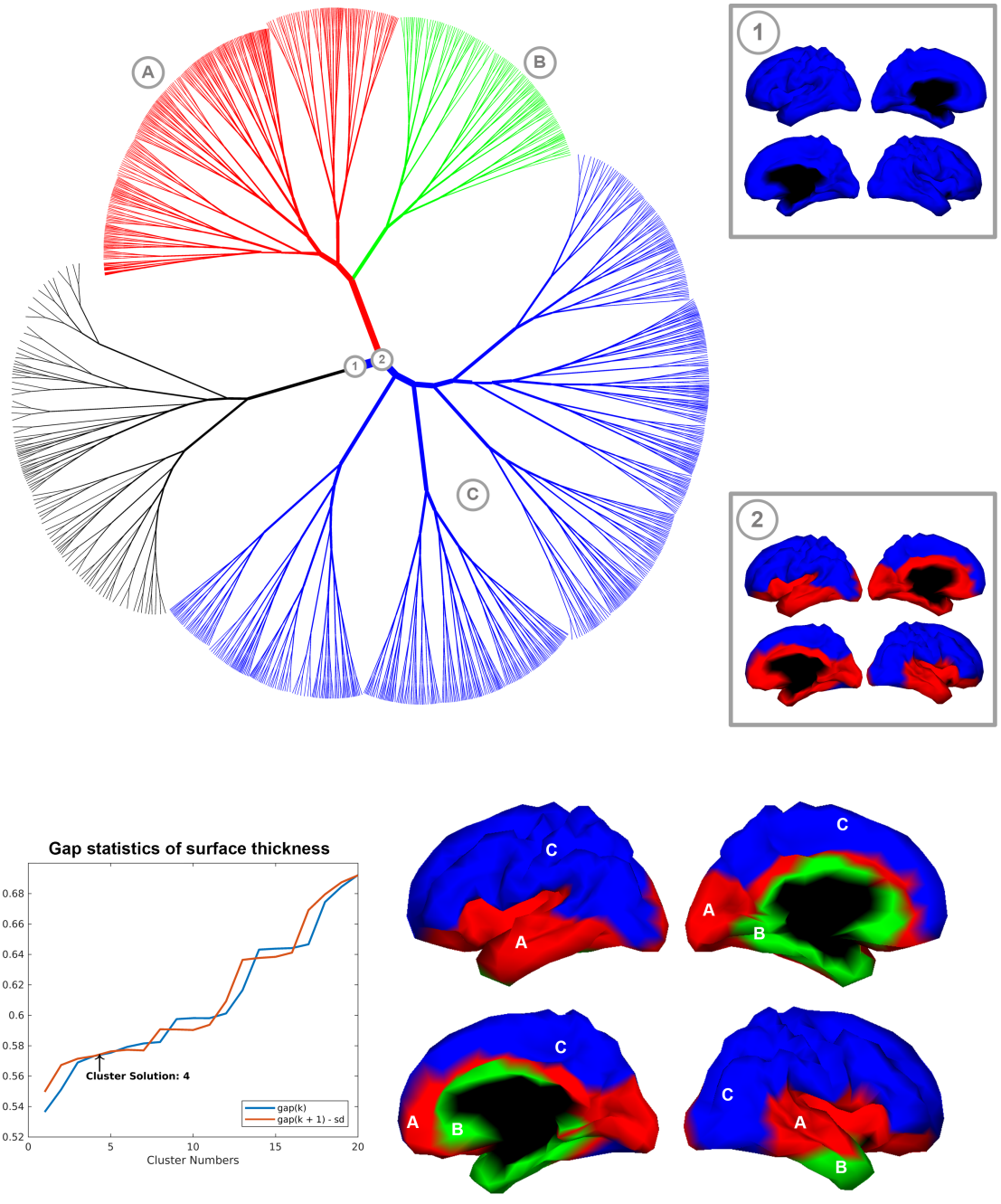

Figure 6.3: Tree-structured topology of cortical surface thickness. The tree structure is based on the dendrogram of hierarchical clustering while the cluster results were based on the gap statistics. Each of the four identified clusters except subcortical regions were labeled in different colors and labeled as A to C. The first two bifurcating points were demonstrated in subfigure 1 and 2. The final clusters were visualized in the lower right of the figure.

Appendix A

Final notes

It has been an exciting journey for studying genotype-phenotype mapping in the era of large-scale genomics. Besides the projects mentioned from chapter 2 to chapter 6, there are others that is ”genetically oriented” but out of the scope in this dissertation. In particular, because of the properties of genetic variants (rare mutation rates, randomization during meiosis, and upstream in the causal chains), the genetic association can be a good tool to understand the environmental factors and their underlying processess.

A.1 Spatial gene-by-environment mapping for schizophrenia reveals neighborhood of upbringing effects beyond urban-rural demarcations

Being born and residing in an urban setting during early life has been linked to an increased risk of schizophrenia. Apart from understanding the environmental factors that underlie this association, there is currently a dearth of information about regionally-varying risk factors mapped at a resolution superior to crude urban-rural categories. It

is also unclear whether spatial variation is related to genetic risks for schizophrenia, a disorder which has high heritability at the population level. We utilized a large-scale genetic case-cohort study (n=23,852) to critically explore the complex interplay between locale of upbringing effects and individuals genetic liabilities for schizophrenia. We applied a novel spatial mapping approach to estimate the spatial variation of locale effects (E) and gene-by-locale interactions (GxE) after taking into consideration urban-rural differences, genetic ancestry, and individuals genetic liabilities. Genetic liabilities were assessed using polygenic risk scores (PRS) derived from an independent genome-wide association study cohort (n = 150,064). We found significant contributions of spatially varying E and GxE beyond simple urban-rural differences. The E and GxE explained 10- and 5-fold more disease variance than the urban-rural categories, respectively. Within the boundaries of the capital city of Copenhagen, spatial variation was observed with odds ratios ranging from 0.63 to 1.90 for E and 0.72 to 1.39 for GxE. An interactive map can be found at (https://chunchiehfan.shinyapps.io/iPSYCH_Geo/). Our results indicate the locale of individuals upbringing has is associated with the risk of schizophrenia. This risk variation has finer resolution than a simple urban-rural demarcation. A series of sensitivity analyses suggest the locale of upbringing can substantially modulate individuals genetic susceptibilities to schizophrenia. Our results are a first look into spatial risk variation in the context of large-scale genetic studies, which could contribute to our understanding of modifiable risk factors of schizophrenia.

Bibliography

[1] Douglas S Falconer, Trudy FC Mackay, and Richard Frankham. Introduction to quantitative genetics (4th edn). *Trends in Genetics*, 12(7):280

[2] Eric S Lander and Nicholas J Schork. Genetic dissection of complex traits. *SCIENCE-NEW YORK THEN WASHINGTON-*, pages 2037–2037 0036–8075, 1994.

[3] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017/08/15.

[4] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

[5] Suzanne Sniekers, Sven Stringer, Kyoko Watanabe, Philip R Jansen, Jonathan R I Coleman, Eva Krapohl, Erdogan Taskesen, Anke R Hammerschlag, Aysu Okbay, Delilah Zabaneh, Najaf Amin, Gerome Breen, David Cesarini, Christopher F Chabris, William G Iacono, M Arfan Ikram, Magnus Johannesson, Philipp Koellinger, James J Lee, Patrik K E Magnusson, Matt McGue, Mike B Miller, William E R Ollier, Antony Payton, Neil Pendleton, Robert Plomin, Cornelius A Rietveld, Henning Tiemeier, Cornelia M van Duijn, and Danielle Posthuma. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat Genet*, 49(7):1107–1112, 07 2017.

[6] Irving I. Gottesman and Todd D. Gould. The endophenotype concept in psychiatry: Etymology and strategic intentions. *American Journal of Psychiatry*, 160(4):636–645, 2003. PMID: 12668349.

[7] David B Allison, Bonnie Thiel, Pamela St Jean, Robert C Elston, Ming C Infante, and Nicholas J Schork. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *The American Journal of Human Genetics*, 63(4):1190–1201, 1998.

[8] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis. i. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–94, 1999.

[9] B. Fischl, M. I. Sereno, and A. M. Dale. Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[10] Chi-Hua Chen, Mark Fiecas, E. D. Gutiérrez, Matthew S. Panizzon, Lisa T. Eyler, Eero Vuoksimaa, Wesley K. Thompson, Christine Fennema-Notestine, Donald J. Hagler, Terry L. Jernigan, Michael C. Neale, Carol E. Franz, Michael J. Lyons, Bruce Fischl, Ming T. Tsuang, Anders M. Dale, and William S. Kremen. Genetic topography of brain morphology. *Proceedings of the National Academy of Sciences*, 110(42):17089–17094, 2013.

[11] J. Eric Schmitt, Michael C. Neale, Bilqis Fassassi, Javier Perez, Rhoshel K. Lenroot, Elizabeth M. Wells, and Jay N. Giedd. The dynamic role of genetics on cortical patterning during childhood and adolescence. *Proceedings of the National Academy of Sciences*, 111(18):6774–6779, 2014.

[12] Chi-Hua Chen, Qian Peng, Andrew J. Schork, Min-Tzu Lo, Chun-Chieh Fan, Yunpeng Wang, Rahul S. Desikan, Francesco Bettella, Donald J. Hagler, Neurocognition Pediatric Imaging, Study Genetics, Initiative Alzheimer's Disease Neuroimaging, Lars T. Westlye, William S. Kremen, Terry L. Jernigan, Stephanie Le Hellard, Vidar M. Steen, Thomas Espeseth, Matt Huentelman, Asta K. Håberg, Ingrid Agartz, Srdjan Djurovic, Ole A. Andreassen, Nicholas Schork, Anders M. Dale, Neurocognition Pediatric Imaging, Study Genetics, and Initiative Alzheimer's Disease Neuroimaging. Large-scale genomics unveil polygenic architecture of human cortical surface area. *Nature Communications*, 6:7549, 2015.

[13] Derrek P. Hibar, Jason L. Stein, Miguel E. Renteria, Alejandro Arias-Vasquez, Sylvane Desrivieres, Neda Jahanshad, Roberto Toro, Katharina Wittfeld, Lucija Abramovic, Micael Andersson, Benjamin S. Aribisala, Nicola J. Armstrong, Manon Bernard, Marc M. Bohlken, Marco P. Boks, Janita Bralten, Andrew A. Brown, M. Mallar Chakravarty, Qiang Chen, Christopher R. K. Ching, Gabriel Cuellar-Partida, Anouk den Braber, Sudheer Giddaluru, Aaron L. Goldman, Oliver Grimm, Tulio Guadalupe, Johanna Hass, Girma Woldehawariat, Avram J. Holmes, Martine Hoogman, Deborah Janowitz, Tianye Jia, Sungeun Kim, Marieke Klein, Bernd Kraemer, Phil H. Lee, Loes M. Olde Loohuis, Michelle Luciano, Christine Macare, Karen A. Mather, Manuel Mattheisen, Yuri Milaneschi, Kwangsik Nho, Martina Papmeyer, Adaikalavan Ramasamy, Shannon L. Risacher, Roberto Roiz-Santianez, Emma J. Rose, Alireza Salami, Philipp G. Samann, Lianne Schmaal, Andrew J. Schork, Jean Shin, Lachlan T. Strike, Alexander Teumer, Marjolein M. J. van Donkelaar, Kristel R. van Eijk, Raymond K. Walters, Lars T. Westlye, Christopher D. Whelan, Anderson M. Winkler, Marcel P. Zwiers, Saud Alhusaini, Lavinia Athanasiu, Stefan Ehrlich, Marina M. H. Hakobjan, Cecilie B. Hartberg,

Unn K. Haukvik, Angelien J. G. A. M. Heister, David Hoehn, Dalia Kasperaviciute, David C. M. Liewald, Lorna M. Lopez, Remco R. R. Makkinje, Mar Matarin, Marlies A. M. Naber, D. Reese McKay, Margaret Needham, Allison C. Nugent, Benno Putz, Natalie A. Royle, Li Shen, Emma Sprooten, Daniah Trabzuni, Saskia S. L. van der Marel, Kimm J. E. van Hulzen, Esther Walton, Christiane Wolf, Laura Almasy, David Ames, Sampath Arepalli, Amelia A. Assareh, Mark E. Bastin, Henry Brodaty, Kazima B. Bulayeva, Melanie A. Carless, Sven Cichon, Aiden Corvin, Joanne E. Curran, Michael Czisch, Greig I. de Zubicaray, Allissa Dillman, Ravi Duggirala, Thomas D. Dyer, Susanne Erk, Iryna O. Fedko, Luigi Ferrucci, Tatiana M. Foroud, Peter T. Fox, Masaki Fukunaga, J. Raphael Gibbs, Harald H. H. Goring, Robert C. Green, Sebastian Guelfi, Narelle K. Hansell, Catharina A. Hartman, Katrin Hegenscheid, Andreas Heinz, Dena G. Hernandez, Dirk J. Heslenfeld, Pieter J. Hoekstra, Florian Holsboer, Georg Homuth, Jouke-Jan Hottenga, Masashi Ikeda, Clifford R. Jack Jr, Mark Jenkinson, Robert Johnson, Ryota Kanai, Maria Keil, Jack W. Kent Jr, Peter Kochunov, John B. Kwok, Stephen M. Lawrie, Xinmin Liu, Dan L. Longo, Katie L. McMahon, Eva Meisenzahl, Ingrid Melle, Sebastian Mohnke, Grant W. Montgomery, Jeanette C. Mostert, Thomas W. Muhleisen, Michael A. Nalls, Thomas E. Nichols, Lars G. Nilsson, Markus M. Nothen, Kazutaka Ohi, Rene L. Olvera, Rocio Perez-Iglesias, G. Bruce Pike, Steven G. Potkin, Ivar Reinvang, Simone Reppermund, Marcella Rietschel, Nina Romanczuk-Seiferth, Glenn D. Rosen, Dan Rujescu, Knut Schnell, Peter R. Schofield, Colin Smith, Vidar M. Steen, Jessika E. Sussmann, Anbupalam Thalamuthu, Arthur W. Toga, Bryan J. Traynor, Juan Troncoso, Jessica A. Turner, Maria C. Valdes Hernandez, Dennis van 't Ent, Marcel van der Brug, Nic J. A. van der Wee, Marie-Jose van Tol, Dick J. Veltman, Thomas H. Wassink, Eric Westman, Ronald H. Zielke, Alan B. Zonderman, David G. Ashbrook, Reinmar Hager, Lu Lu, Francis J. McMahon, Derek W. Morris, Robert W. Williams, Han G. Brunner, Randy L. Buckner, Jan K. Buitelaar, Wiepke Cahn, Vince D. Calhoun, Gianpiero L. Cavalleri, Benedicto Crespo-Facorro, Anders M. Dale, Gareth E. Davies, Norman Delanty, Chantal Depondt, Srdjan Djurovic, Wayne C. Drevets, Thomas Espeseth, Randy L. Gollub, Beng-Choon Ho, Wolfgang Hoffmann, Norbert Hosten, Rene S. Kahn, Stephanie Le Hellard, Andreas Meyer-Lindenberg, Bertram Muller-Myhsok, Matthias Nauck, Lars Nyberg, Massimo Pandolfo, Brenda W. J. H. Penninx, Joshua L. Roffman, Sanjay M. Sisodiya, Jordan W. Smoller, Hans van Bokhoven, Neeltje E. M. van Haren, Henry Volzke, Henrik Walter, Michael W. Weiner, Wei Wen, Tonya White, Ingrid Agartz, Ole A. Andreassen, John Blangero, Dorret I. Boomsma, Rachel M. Brouwer, Dara M. Cannon, Mark R. Cookson, Eco J. C. de Geus, Ian J. Deary, Gary Donohoe, Guillen Fernandez, Simon E. Fisher, Clyde Francks, David C. Glahn, Hans J. Grabe, Oliver Gruber, John Hardy, Ryota Hashimoto, Hilleke E. Hulshoff Pol, Erik G. Jonsson, Iwona Kloszewska, Simon Lovestone, Venkata S. Mattay, Patrizia Mecocci, Colm McDonald, Andrew M. McIntosh, Roel A. Ophoff, Tomas Paus, Zdenka Pausova, Mina Ryten, Perminder S. Sachdev, Andrew J. Saykin, Andy

Simmons, Andrew Singleton, Hilkka Soininen, Joanna M. Wardlaw, Michael E. Weale, Daniel R. Weinberger, Hieab H. H. Adams, Lenore J. Launer, Stephan Seiler, Reinhold Schmidt, Ganesh Chauhan, Claudia L. Satizabal, James T. Becker, Lisa Yanek, Sven J. van der Lee, Maritza Ebling, Bruce Fischl, W. T. Longstreth Jr, Douglas Greve, Helena Schmidt, Paul Nyquist, Louis N. Vinke, Cornelia M. van Duijn, Luting Xue, Bernard Mazoyer, Joshua C. Bis, Vilmundur Gudnason, Sudha Seshadri, M. Arfan Ikram, The Alzheimer/'s Disease Neuroimaging Initiative, The CHARGE Consortium, EPIGEN, IMAGEN, SYS, Nicholas G. Martin, Margaret J. Wright, Gunter Schumann, Barbara Franke, Paul M. Thompson, and Sarah E. Medland. Common genetic variants influence human subcortical brain structures. *Nature*, 520(7546):224–229, 04 2015.

[14] Paul M. Thompson, Ole A. Andreassen, Alejandro Arias-Vasquez, Carrie E. Bearden, Premika S. Boedhoe, Rachel M. Brouwer, Randy L. Buckner, Jan K. Buitelaar, Kazima B. Bulayeva, Dara M. Cannon, Ronald A. Cohen, Patricia J. Conrod, Anders M. Dale, Ian J. Deary, Emily L. Dennis, Marcel A. de Reus, Sylvane Desrivieres, Danai Dima, Gary Donohoe, Simon E. Fisher, Jean-Paul Fouche, Clyde Francks, Sophia Frangou, Barbara Franke, Habib Ganjgahi, Hugh Garavan, David C. Glahn, Hans J. Grabe, Tulio Guadalupe, Boris A. Gutman, Ryota Hashimoto, Derrek P. Hibar, Dominic Holland, Martine Hoogman, Hilleke E. Hulshoff Pol, Norbert Hosten, Neda Jahanshad, Sinead Kelly, Peter Kochunov, William S. Kremen, Phil H. Lee, Scott Mackey, Nicholas G. Martin, Bernard Mazoyer, Colm McDonald, Sarah E. Medland, Rajendra A. Morey, Thomas E. Nichols, Tomas Paus, Zdenka Pausova, Lianne Schmaal, Gunter Schumann, Li Shen, Sanjay M. Sisodiya, Dirk J. A. Smit, Jordan W. Smoller, Dan J. Stein, Jason L. Stein, Roberto Toro, Jessica A. Turner, Martijn P. van den Heuvel, Odile L. van den Heuvel, Theo G. M. van Erp, Daan van Rooij, Dick J. Veltman, Henrik Walter, Yalin Wang, Joanna M. Wardlaw, Christopher D. Whelan, Margaret J. Wright, and Jieping Ye. Enigma and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*, 145, Part B:389–408, 2017.

[15] Lindsay A Farrer, L Adrienne Cupples, Jonathan L Haines, Bradley Hyman, Walter A Kukull, Richard Mayeux, Richard H Myers, Margaret A Pericak-Vance, Neil Risch, and Cornelia M Van Duijn. Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease: a meta-analysis. *Jama*, 278(16):1349–1356

[16] Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research*, 17(10):1520–1528

[17] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, 9(3):e1003348

[18] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews. Genetics*, 17(7):392–406, 2016.

[19] Chi-Hua Chen, E. D. Gutierrez, Wes Thompson, Matthew S. Panizzon, Terry L. Jernigan, Lisa T. Eyler, Christine Fennema-Notestine, Amy J. Jak, Michael C. Neale, Carol E. Franz, Michael J. Lyons, Michael D. Grant, Bruce Fischl, Larry J. Seidman, Ming T. Tsuang, William S. Kremen, and Anders M. Dale. Hierarchical genetic organization of human cortical surface area. *Science*, 335(6076):1634–1636, 2012.

[20] Tian Ge, Thomas E. Nichols, Phil H. Lee, Avram J. Holmes, Joshua L. Roffman, Randy L. Buckner, Mert R. Sabuncu, and Jordan W. Smoller. Massively expedited genome-wide heritability analysis (megha). *Proceedings of the National Academy of Sciences*, 112(8):2479–2484, 2015.

[21] Lei Du, Heng Huang, Jingwen Yan, Sungeun Kim, Shannon L. Risacher, Mark Inlow, Jason H. Moore, Andrew J. Saykin, Li Shen, and for the Alzheimer's Disease Neuroimaging Initiative. Structured sparse canonical correlation analysis for brain imaging genetics: an improved graphnet method. *Bioinformatics*, 2016.

[22] Amy S Kelley, Kathleen McGarry, Rebecca Gorges, and Jonathan S Skinner. The burden of health care costs for patients with dementia in the last 5 years of lifeburden of health care costs for patients with dementia. *Annals of internal medicine*, 163(10):729–736 2015.

[23] Celeste M Karch, Carlos Cruchaga, and Alison M Goate. Alzheimer's disease genetics: from the bench to the clinic. *Neuron*, 83(1):11–26

[24] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, and Gary W Beecham. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458

[25] Rahul S Desikan, Andrew J Schork, Yunpeng Wang, Wesley K Thompson, Abbas Dehghan, Paul M Ridker, Daniel I Chasman, Linda K McEvoy, Dominic Holland, and Chi-Hua Chen. Polygenic overlap between c-reactive protein, plasma lipids, and alzheimer disease. *Circulation*, 131(23):2061–2069

[26] Adam C Naj, Gyungah Jun, Gary W Beecham, Li-San Wang, Badri Narayan Vardarajan, Jacqueline Buros, Paul J Gallins, Joseph D Buxbaum, Gail P Jarvik, and Paul K Crane. Common variants at ms4a4/ms4a6e, cd2ap, cd33 and epha1 are associated with late-onset alzheimer's disease. *Nature genetics*, 43(5):436–441

[27] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M Stadlan. Clinical diagnosis of alzheimer's disease report

of the nincdsadrda work group* under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34(7):939–939

[28] Duane L Beekly, Erin M Ramos, William W Lee, Woodrow D Deitrich, Mary E Jacka, Joylee Wu, Janene L Hubbard, Thomas D Koepsell, John C Morris, and Walter A Kukull. The national alzheimer's coordinating center (nacc) database: the uniform data set. *Alzheimer Disease Associated Disorders*, 21(3):249–258 0893–0341, 2007.

[29] Heiko Braak and Eva Braak. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica*, 82(4):239–259

[30] Suzanne S Mirra, A Heyman, D McKeel, SM Sumi, Barbara J Crain, LM Brownlee, FS Vogel, JP Hughes, G Van Belle, and L Berg. The consortium to establish a registry for alzheimer's disease (cerad) part ii. standardization of the neuropathologic assessment of alzheimer's disease. *Neurology*, 41(4):479–479

[31] Bradley T Hyman, Creighton H Phelps, Thomas G Beach, Eileen H Bigio, Nigel J Cairns, Maria C Carrillo, Dennis W Dickson, Charles Duyckaerts, Matthew P Frosch, and Eliezer Masliah. National institute on aging–alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease. *Alzheimer's Dementia*, 8(1):1–13

[32] W. Y. Yang, J. Novembre, E. Eskin, and E. Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet*, 44(6):725–31, 2012. Yang, Wen-Yun Novembre, John Eskin, Eleazar Halperin, Eran eng K25 HL080079/HL/NHLBI NIH HHS/ P01 HL030568/HL/NHLBI NIH HHS/ P01 HL28481/HL/NHLBI NIH HHS/ P01 HL30568/HL/NHLBI NIH HHS/ U01 DA024417/DA/NIDA NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. 2012/05/23 06:00 Nat Genet. 2012 May 20;44(6):725-31. doi: 10.1038/ng.2285.

[33] John P Klein, Hans C Van Houwelingen, Joseph G Ibrahim, and Thomas H *Handbook of survival analysis*. CRC Press, 2016.

[34] Kenneth J Rothman, Sander Greenland, and Timothy L *Modern epidemiology*. Lippincott Williams Wilkins, 2008.

[35] Ron Brookmeyer, Sarah Gray, and Claudia Kawas. Projections of alzheimer's disease in the united states and the public health impact of delaying disease onset. *American journal of public health*, 88(9):1337–1342 0090–0036, 1998.

[36] Valentina Escott-Price, Rebecca Sims, Christian Bannister, Denise Harold, Maria Vronskaya, Elisa Majounie, Nandini Badarinarayan, GERAD/PERADES, IGAP consortia, and Kevin Morgan. Common polygenic variation enhances risk prediction for alzheimer's disease. *Brain*, 138(12):3673–3684

[37] Jennifer S Yokoyama, Luke W Bonham, Renee L Sears, Eric Klein, Anna Karydas, Joel H Kramer, Bruce L Miller, and Giovanni Coppola. Decision tree analysis of genetic risk for clinically heterogeneous alzheimer's disease. *BMC neurology*, 15(1):47

[38] Elizabeth C Mormino, Reisa A Sperling, Avram J Holmes, Randy L Buckner, Philip L De Jager, Jordan W Smoller, Mert R Sabuncu, Michael Weiner, Paul Aisen, and Ronald Petersen. Polygenic risk of alzheimer disease is associated with early-and late-life processes. *Neurology*, 87(5):481–488

[39] Henna Martiskainen, Seppo Helisalmi, Jayashree Viswanathan, Mitja Kurki, Anette Hall, Sanna-Kaisa Herukka, Timo Sarajärvi, Teemu Natunen, Kaisa Kurkinen, and Jaakko Huovinen. Effects of alzheimer's disease-associated risk loci on cerebrospinal fluid biomarkers and disease progression: a polygenic risk score approach. *Journal of Alzheimer's Disease*, 43(2):565–573 2015.

[40] A Lacour, A Espinosa, E Louwersheimer, S Heilmann, I Hernández, S Wolfsgruber, V Fernandez, H Wagner, M Rosende-Roca, and A Mauleon. Genome-wide significant risk factors for alzheimer's disease: role in progression to dementia due to alzheimer's disease among subjects with mild cognitive impairment. *Molecular psychiatry*, 22(1):153, 2017.

[41] Vincent Chouraki, Christiane Reitz, Fleur Maury, Joshua C Bis, Celine Bellenguez, Lei Yu, Johanna Jakobsdottir, Shubhabrata Mukherjee, Hieab H Adams, and Seung Hoan Choi. Evaluation of a genetic risk score to improve risk prediction for alzheimer's disease. *Journal of Alzheimer's Disease*, 53(3):921–932 2016.

[42] M-X Tang, P Cross, H Andrews, DM Jacobs, S Small, K Bell, C Merchant, R Lantigua, R Costa, and Y Stern. Incidence of ad in african-americans, caribbean hispanics, and caucasians in northern manhattan. *Neurology*, 56(1):49–56

[43] Amit V Khera, Connor A Emdin, Isabel Drake, Pradeep Natarajan, Alexander G Bick, Nancy R Cook, Daniel I Chasman, Usman Baber, Roxana Mehran, and Daniel J Rader. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, 375(24):2349–2358 0028–4793, 2016.

[44] Nasim Mavaddat, Paul DP Pharoah, Kyriaki Michailidou, Jonathan Tyrer, Mark N Brook, Manjeet K Bolla, Qin Wang, Joe Dennis, Alison M Dunning, and Mitul Shah. Prediction of breast cancer risk based on profiling with common genetic variants. *JNCI: Journal of the National Cancer Institute*, 107(5 0027-8874), 2015.

[45] Ming-Xin Tang, Yaakov Stern, Karen Marder, Karen Bell, Barry Gurland, Rafael Lantigua, Howard Andrews, Lin Feng, Benjamin Tycko, and Richard Mayeux. The apoe-Ł 4 allele and the risk of alzheimer disease among african americans, whites, and hispanics. *Jama*, 279(10):751–755

[46] Christiane Reitz, Gyungah Jun, Adam Naj, Ruchita Rajbhandary, Badri Narayan Vardarajan, Li-San Wang, Otto Valladares, Chiao-Feng Lin, Eric B Larson, and Neill R Graff-Radford. Variants in the atp-binding cassette transporter (abca7), apolipoprotein e 4, and the risk of late-onset alzheimer disease in african americans. *Jama*, 309(14):1483–1492

[47] D. Falk. Interpreting sulci on hominin endocasts: old hypotheses and new findings. *Frontiers in Human Neuroscience*, 8, 2014. Ah4jx Times Cited:1 Cited References Count:42.

[48] Ralph L Holloway. The human brain evolving: a personal retrospective. *Annual review of Anthropology*, 37:1–19, 2008.

[49] Anders M. Fjell, Kristine Beate Walhovd, Timothy T. Brown, Joshua M. Kuperman, Yoonho Chung, Donald J. Hagler, Vijay Venkatraman, J. Cooper Roddey, Matthew Erhart, Connor McCabe, Natacha Akshoomoff, David G. Amaral, Cinnamon S. Bloss, Ondrej Libiger, Burcu F. Darst, Nicholas J. Schork, B. J. Casey, Linda Chang, Thomas M. Ernst, Jeffrey R. Gruen, Walter E. Kaufmann, Tal Kenet, Jean Frazier, Sarah S. Murray, Elizabeth R. Sowell, Peter van Zijl, Stewart Mostofsky, Terry L. Jernigan, and Anders M. Dale. Multimodal imaging of the self-regulating developing brain. *Proceedings of the National Academy of Sciences*, 109(48):19620–19625, 2012.

[50] K. B. Walhovd, A. M. Fjell, T. T. Brown, J. M. Kuperman, Y. H. Chung, D. J. Hagler, J. C. Roddey, M. Erhart, C. McCabe, N. Akshoomoff, D. G. Amaral, C. S. Bloss, O. Libiger, N. J. Schork, B. F. Darst, B. J. Casey, L. D. Chang, T. M. Ernst, J. Frazier, J. R. Gruen, W. E. Kaufmann, S. S. Murray, P. van Zijl, S. Mostofsky, A. M. Dale, and Pediat Imaging Neurocognition Gene. Long-term influence of normal variation in neonatal characteristics on human brain development. *Proceedings of the National Academy of Sciences of the United States of America*, 109(49):20089–20094, 2012. 054MF Times Cited:13 Cited References Count:58.

[51] Emiliano Bruner, José Manuel de la Cuétara, Michael Masters, Hideki Amano, and Naomichi Ogihara. Functional craniology and brain evolution: from paleontology to biomedicine. *Frontiers in Neuroanatomy*, 8:19, 2014. 24765064[pmid] Front Neuroanat.

[52] H. Reyes-Centeno, S. Ghirotto, F. Detroit, D. Grimaud-Herve, G. Barbujani, and K. Harvati. Genomic and cranial phenotype data support multiple modern human dispersals from africa and a southern route into asia. *Proc Natl Acad Sci U S A*, 111(20):7248–53, 2014.

[53] A. Manica, W. Amos, F. Balloux, and T. Hanihara. The effect of ancient population bottlenecks on human phenotypic variation. *Nature*, 448(7151):346–U6, 2007. 191GC Times Cited:127 Cited References Count:29.

[54] D. E. Lieberman, B. M. McBratney, and G. Krovitz. The evolution and development of cranial form in homosapiens. *Proc Natl Acad Sci U S A*, 99(3):1134–9, 2002. Lieberman, Daniel E McBratney, Brandeis M Krovitz, Gail eng 2002/01/24 10:00 Proc Natl Acad Sci U S A. 2002 Feb 5;99(3):1134-9. Epub 2002 Jan 22.

[55] Trygve E Bakken, Anders M Dale, and Nicholas J Schork. A geographic cline of skull and brain morphology among individuals of european ancestry. *Hum Hered*, 72(1):35–44, 2011.

[56] Timothy T Brown, Joshua M Kuperman, Yoonho Chung, Matthew Erhart, Connor McCabe, Donald J Hagler Jr, Vijay K Venkatraman, Natacha Akshoomoff, David G Amaral, Cinnamon S Bloss, B. J. Casey, Linda Chang, Thomas M Ernst, Jean A Frazier, Jeffrey R Gruen, Walter E Kaufmann, Tal Kenet, David N Kennedy, Sarah S Murray, Elizabeth R Sowell, Terry L Jernigan, and Anders M Dale. Neuroanatomical assessment of biological maturity. *Current Biology*, 22(18):1693–1698, 2012.

[57] D. Reich, N. Patterson, D. Campbell, A. Tandon, S. Mazieres, N. Ray, M. V. Parra, W. Rojas, C. Duque, N. Mesa, L. F. Garcia, O. Triana, S. Blair, A. Maestre, J. C. Dib, C. M. Bravi, G. Bailliet, D. Corach, T. Hunemeier, M. C. Bortolini, F. M. Salzano, M. L. Petzl-Erler, V. Acuna-Alonzo, C. Aguilar-Salinas, S. Canizales-Quinteros, T. Tusie-Luna, L. Riba, M. Rodriguez-Cruz, M. Lopez-Alarcon, R. Coral-Vazquez, T. Canto-Cetina, I. Silva-Zolezzi, J. C. Fernandez-Lopez, A. V. Contreras, G. Jimenez-Sanchez, M. J. Gomez-Vazquez, J. Molina, A. Carracedo, A. Salas, C. Gallo, G. Poletti, D. B. Witonsky, G. Alkorta-Aranburu, R. I. Sukernik, L. Osipova, S. A. Fedorova, R. Vasquez, M. Villena, C. Moreau, R. Barrantes, D. Pauls, L. Excoffier, G. Bedoya, F. Rothhammer, J. M. Dugoujon, G. Larrouy, W. Klitz, D. Labuda, J. Kidd, K. Kidd, A. Di Rienzo, N. B. Freimer, A. L. Price, and A. Ruiz-Linares. Reconstructing native american population history. *Nature*, 488(7411):370–+, 2012.

[58] The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010. 10.1038/nature09298.

[59] C. Y. Chen, S. Pollack, D. J. Hunter, J. N. Hirschhorn, P. Kraft, and A. L. Price. Improved ancestry inference using weights from external reference panels. *Bioinformatics*, 29(11):1399–1406, 2013.

[60] S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams. The genetic structure and history of africans and african americans. *Science*, 324(5930):1035–44, 2009. Tishkoff, Sarah A Reed, Floyd A

Friedlaender, Francoise R Ehret, Christopher Ranciaro, Alessia Froment, Alain Hirbo, Jibril B Awomoyi, Agnes A Bodo, Jean-Marie Doumbo, Ogobara Ibrahim, Muntaser Juma, Abdalla T Kotze, Maritha J Lema, Godfrey Moore, Jason H Mortensen, Holly Nyambo, Thomas B Omar, Sabah A Powell, Kweli Pretorius, Gideon S Smith, Michael W Thera, Mahamadou A Wambebe, Charles Weber, James L Williams, Scott M eng 1R01GM083606-01/GM/NIGMS NIH HHS/ F32 HG003801/HG/NHGRI NIH HHS/ F32 HG003801-01A1/HG/NHGRI NIH HHS/ F32HG03801/HG/NHGRI NIH HHS/ R01 GM076637/GM/NIGMS NIH HHS/ R01 GM076637-01/GM/NIGMS NIH HHS/ R01 GM083606/GM/NIGMS NIH HHS/ R01 GM083606-01/GM/NIGMS NIH HHS/ R01 HL065234/HL/NHLBI NIH HHS/ R01 HL065234-01/HL/NHLBI NIH HHS/ R01 HL65234/HL/NHLBI NIH HHS/ R01GM076637/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. New York, N.Y. 2009/05/02 09:00 Science. 2009 May 22;324(5930):1035-44. doi: 10.1126/science.1172257. Epub 2009 Apr 30.

[61] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, H. Tang, G. S. Barsh, D. M. Absher, D. A. Puts, J. Rocha, S. Beleza, R. W. Pereira, G. Baynam, P. Suetens, D. Vandermeulen, J. K. Wagner, J. S. Boster, and M. D. Shriver. Modeling 3d facial shape from dna. *PLoS Genet*, 10(3):e1004224, 2014. Claes, Peter Liberton, Denise K Daniels, Katleen Rosana, Kerri Matthes Quillen, Ellen E Pearson, Laurel N McEvoy, Brian Bauchet, Marc Zaidi, Arslan A Yao, Wei Tang, Hua Barsh, Gregory S Absher, Devin M Puts, David A Rocha, Jorge Beleza, Sandra Pereira, Rinaldo W Baynam, Gareth Suetens, Paul Vandermeulen, Dirk Wagner, Jennifer K Boster, James S Shriver, Mark D eng K99HG006446/HG/NHGRI NIH HHS/ Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't 2014/03/22 06:00 PLoS Genet. 2014 Mar 20;10(3):e1004224. doi: 10.1371/journal.pgen.1004224. eCollection 2014 Mar.

[62] W. W. Howells. *Skull shape and the map: craniometric analyses in the dispersion of modern homo*. Peabody Museum of Archaeology and Ethnology. Harvard University Press, 1989.

[63] J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–4, 2008. Li, Jun Z Absher, Devin M Tang, Hua Southwick, Audrey M Casto, Amanda M Ramachandran, Sohini Cann, Howard M Barsh, Gregory S Feldman, Marcus Cavalli-Sforza, Luigi L Myers, Richard M eng GM073059/GM/NIGMS NIH HHS/ GM28016/GM/NIGMS NIH HHS/ Research Support, N.I.H., Extramural New York, N.Y. 2008/02/23 09:00 Science. 2008 Feb 22;319(5866):1100-4. doi: 10.1126/science.1153717.

[64] Hugo Pan-Asian SNP Consortium, M. A. Abdulla, I. Ahmed, A. Assawamakin, J. Bhak, S. K. Brahmachari, G. C. Calacal, A. Chaurasia, C. H. Chen, J. Chen, Y. T. Chen, J. Chu, E. M. Cutiongco-de la Paz, M. C. De Ungria, F. C. Delfin, J. Edo, S. Fuchareon, H. Ghang, T. Gojobori, J. Han, S. F. Ho, B. P. Hoh, W. Huang, H. Inoko, P. Jha, T. A. Jinam, L. Jin, J. Jung, D. Kangwanpong, J. Kampuansai, G. C. Kennedy, P. Khurana, H. L. Kim, K. Kim, S. Kim, W. Y. Kim, K. Kimm, R. Kimura, T. Koike, S. Kulawonganunchai, V. Kumar, P. S. Lai, J. Y. Lee, S. Lee, E. T. Liu, P. P. Majumder, K. K. Mandapati, S. Marzuki, W. Mitchell, M. Mukerji, K. Naritomi, C. Ngamphiw, N. Niikawa, N. Nishida, B. Oh, S. Oh, J. Ohashi, A. Oka, R. Ong, C. D. Padilla, P. Palittapongarnpim, H. B. Perdigon, M. E. Phipps, E. Png, Y. Sakaki, J. M. Salvador, Y. Sandraling, V. Scaria, M. Seielstad, M. R. Sidek, A. Sinha, M. Srikummool, H. Sudoyo, S. Sugano, H. Suryadi, Y. Suzuki, K. A. Tabbada, A. Tan, K. Tokunaga, S. Tongsima, L. P. Villamor, E. Wang, Y. Wang, H. Wang, J. Y. Wu, H. Xiao, S. Xu, J. O. Yang, Y. Y. Shugart, H. S. Yoo, W. Yuan, G. Zhao, B. A. Zilfalil, and Consortium Indian Genome Variation. Mapping human genetic diversity in asia. *Science*, 326(5959):1541–5, 2009. Abdulla, Mahmood Ameen Ahmed, Ikhlak Assawamakin, Anunchai Bhak, Jong Brahmachari, Samir K Calacal, Gayvelline C Chaurasia, Amit Chen, Chien-Hsiun Chen, Jieming Chen, Yuan-Tsong Chu, Jiayou Cutiongco-de la Paz, Eva Maria C De Ungria, Maria Corazon A Delfin, Frederick C Edo, Juli Fuchareon, Suthat Ghang, Ho Gojobori, Takashi Han, Junsong Ho, Sheng-Feng Hoh, Boon Peng Huang, Wei Inoko, Hidetoshi Jha, Pankaj Jinam, Timothy A Jin, Li Jung, Jongsun Kangwanpong, Daoroong Kampuansai, Jatupol Kennedy, Giulia C Khurana, Preeti Kim, Hyung-Lae Kim, Kwangjoong Kim, Sangsoo Kim, Woo-Yeon Kimm, Kuchan Kimura, Ryosuke Koike, Tomohiro Kulawonganunchai, Supasak Kumar, Vikrant Lai, Poh San Lee, Jong-Young Lee, Sunghoon Liu, Edison T Majumder, Partha P Mandapati, Kiran Kumar Marzuki, Sangkot Mitchell, Wayne Mukerji, Mitali Naritomi, Kenji Ngamphiw, Chumpol Niikawa, Norio Nishida, Nao Oh, Bermseok Oh, Sangho Ohashi, Jun Oka, Akira Ong, Rick Padilla, Carmencita D Palittapongarnpim, Prasit Perdigon, Henry B Phipps, Maude Elvira Png, Eileen Sakaki, Yoshiyuki Salvador, Jazelyn M Sandraling, Yuliana Scaria, Vinod Seielstad, Mark Sidek, Mohd Ros Sinha, Amit Srikummool, Metawee Sudoyo, Herawati Sugano, Sumio Suryadi, Helena Suzuki, Yoshiyuki Tabbada, Kristina A Tan, Adrian Tokunaga, Katsushi Tongsima, Sissades Villamor, Lilian P Wang, Eric Wang, Ying Wang, Haifeng Wu, Jer-Yuarn Xiao, Huasheng Xu, Shuhua Yang, Jin Ok Shugart, Yin Yao Yoo, Hyang-Sook Yuan, Wentao Zhao, Guoping Zilfalil, Bin Alwi eng Historical Article Research Support, Non-U.S. Gov't New York, N.Y. 2009/12/17 06:00 Science. 2009 Dec 11;326(5959):1541-5. doi: 10.1126/science.1177074.

[65] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[66] L. L. Cavallisforza, P. Menozzi, and A. Piazza. Demic expansions and human-evolution. *Science*, 259(5095):639–646, 1993.

[67] L. Luca Cavalli-Sforza. The human genome diversity project: past, present and future. *Nat Rev Genet*, 6(4):333–340, 2005. 10.1038/nrg1596.

[68] Emiliano Bruner, Giorgio Manzi, and Juan Luis Arsuaga. Encephalization and allometric trajectories in the genus homo: Evidence from the neandertal and modern lineages. *Proc Natl Acad Sci U S A*, 100(26):15335–15340, 2003.

[69] C. C. Roseman. Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. *Proc Natl Acad Sci U S A*, 101(35):12824–9, 2004. Roseman, Charles C eng 2004/08/25 05:00 Proc Natl Acad Sci U S A. 2004 Aug 31;101(35):12824-9. Epub 2004 Aug 23.

[70] Sarah E. Medland, Neda Jahanshad, Benjamin M. Neale, and Paul M. Thompson. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nat Neurosci*, 17(6):791–800, 2014.

[71] C. A. Winkler, G. W. Nelson, and M. W. Smith. Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics, Vol 11*, 11:65–89, 2010. Brg14 Times Cited:55 Cited References Count:86 Annual Review of Genomics and Human Genetics.

[72] B. R. Pober. Williams-beuren syndrome. *N Engl J Med*, 362(3):239–52, 2010.

[73] T. F. Doyle, U. Bellugi, J. R. Korenberg, and J. Graham. "everybody in the world is my friend" hypersociability in young children with williams syndrome. *Am J Med Genet A*, 124A(3):263–73, 2004.

[74] M. A. Martens, S. J. Wilson, and D. C. Reutens. Research review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype. *Journal of Child Psychology and Psychiatry*, 49(6):576–608, 2008.

[75] T. L. Jernigan and U. Bellugi. Anomalous brain morphology on magnetic resonance images in williams syndrome and down syndrome. *Arch Neurol*, 47(5):529–33, 1990.

[76] C. Gaser, E. Luders, P. M. Thompson, A. D. Lee, R. A. Dutton, J. A. Geaga, K. M. Hayashi, U. Bellugi, A. M. Galaburda, J. R. Korenberg, D. L. Mills, A. W. Toga, and A. L. Reiss. Increased local gyrification mapped in williams syndrome. *Neuroimage*, 33(1):46–54, 2006.

[77] J. S. Kippenhan, R. K. Olsen, C. B. Mervis, C. A. Morris, P. Kohn, A. Meyer-Lindenberg, and K. F. Berman. Genetic contributions to human gyrification: sulcal morphometry in williams syndrome. *J Neurosci*, 25(34):7840–6, 2005.

[78] Shashwath A. Meda, Jennifer R. Pryweller, and Tricia A. Thornton-Wells. Regional brain differences in cortical thickness, surface area and subcortical volume in individuals with williams syndrome. *PLoS ONE*, 7(2):e31913, 2012.

[79] Andreas Meyer-Lindenberg, Philip Kohn, Carolyn B Mervis, J Shane Kippenhan, Rosanna K Olsen, Colleen A Morris, and Karen Faith Berman. Neural basis of genetically determined visuospatial construction deficit in williams syndrome. *Neuron*, 43(5):623–631, 2004.

[80] Paul M Thompson, Agatha D Lee, Rebecca A Dutton, Jennifer A Geaga, Kiralee M Hayashi, Mark A Eckert, Ursula Bellugi, Albert M Galaburda, Julie R Korenberg, and Debra L Mills. Abnormal cortical complexity and thickness profiles mapped in williams syndrome. *Journal of Neuroscience*, 25(16):4146–4158

[81] Günter P. Wagner and Jianzhi Zhang. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature Reviews. Genetics*, 12(3):204–213, 2011.

[82] Marilee A Martens, Sarah J Wilson, Paul Dudgeon, and David C Reutens. Approachability and the amygdala: insights from williams syndrome. *Neuropsychologia*, 47(12):2446–2453

[83] Liliana Capitao, Adriana Sampaio, Cassandra Sampaio, Cristiana Vasconcelos, Montse Férnandez, Elena Garayzábal, Martha E Shenton, and Óscar F Gonçalves. Mri amygdala volume in williams syndrome. *Research in developmental disabilities*, 32(6):2767–2772 0891–4222, 2011.

[84] Fumiko Hoeft, Li Dai, Brian W. Haas, Kristen Sheau, Masaru Mimura, Debra Mills, Albert Galaburda, Ursula Bellugi, Julie R. Korenberg, and Allan L. Reiss. Mapping genetically controlled neural circuits of social behavior and visuo-motor integration by a preliminary examination of atypical deletions with williams syndrome. *PLoS ONE*, 9(8):e104088, 2014.

[85] C. C. Fan, H. Bartsch, A. J. Schork, C. H. Chen, Y. Wang, M. T. Lo, T. T. Brown, J. M. Kuperman, Jr. Hagler, D. J., N. J. Schork, T. L. Jernigan, A. M. Dale, Neurocognition Pediatric Imaging, and Study Genetics. Modeling the 3d geometry of the cortical surface with genetic ancestry. *Curr Biol*, 25(15):1988–92, 2015.

[86] Mark A Eckert, Albert M Galaburda, Asya Karchemskiy, Alyssa Liang, Paul Thompson, Rebecca A Dutton, Agatha D Lee, Ursula Bellugi, Julie R Korenberg, and Debra Mills. Anomalous sylvian fissure morphology in williams syndrome. *NeuroImage*, 33(1):39–45, 2006.

[87] Brian D Mills, Janie Lai, Timothy T Brown, Matthew Erhart, Eric Halgren, Judy Reilly, Mark Appelbaum, and Pamela Moses. Gray matter structure and morphosyntax within a spoken narrative in typically developing children and children

with high functioning autism. *Developmental neuropsychology*, 38(7):461–480, 2013.

[88] D. C. Van Essen, D. Dierker, A. Z. Snyder, M. E. Raichle, A. L. Reiss, and J. Korenberg. Symmetry of cortical folding abnormalities in williams syndrome revealed by surface-based analyses. *J Neurosci*, 26(20):5470–83, 2006.

[89] David Wechsler. Wechsler adult intelligence scale–fourth edition (wais–iv). *San Antonio, TX: NCS Pearson*, 2008.

[90] Nathan White, Cooper Roddey, Ajit Shankaranarayanan, Eric Han, Dan Rettmann, Juan Santos, Josh Kuperman, and Anders Dale. Promo: Real-time prospective motion correction in mri using image-based tracking. *Magnetic Resonance in Medicine*, 63(1):91–105, 2010.

[91] Timothy T Brown, Joshua M Kuperman, Matthew Erhart, Nathan S White, J Cooper Roddey, Ajit Shankaranarayanan, Eric T Han, Dan Rettmann, and Anders M Dale. Prospective motion correction of high-resolution magnetic resonance imaging data in children. *Neuroimage*, 53(1):139–145

[92] J. Jovicich, S. Czanner, D. Greve, E. Haley, A. van der Kouwe, R. Gollub, D. Kennedy, F. Schmitt, G. Brown, J. Macfall, B. Fischl, and A. Dale. Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage*, 30(2):436–43, 2006.

[93] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging*, 17(1):87–97, 1998.

[94] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2009.

[95] S. Marenco, M. A. Siuta, J. S. Kippenhan, S. Grodofsky, W. L. Chang, P. Kohn, C. B. Mervis, C. A. Morris, D. R. Weinberger, A. Meyer-Lindenberg, C. Pierpaoli, and K. F. Berman. Genetic contributions to white matter architecture revealed by diffusion tensor imaging in williams syndrome. *Proc Natl Acad Sci U S A*, 104(38):15117–22, 2007.

[96] Andreas Meyer-Lindenberg, Carolyn B. Mervis, and Karen Faith Berman. Neural mechanisms in williams syndrome: a unique window to genetic influences on cognition and behaviour. *Nature Reviews. Neuroscience*, 7(5):380–393, 2006.

[97] Yalin Wang, Lei Yuan, Jie Shi, Alexander Greve, Jieping Ye, Arthur W Toga, Allan L Reiss, and Paul M Thompson. Applying tensor-based morphometry to parametric surfaces can improve mri-based disease diagnosis. *Neuroimage*, 74:209–230

[98] Ralph Adolphs. The neurobiology of social cognition. *Current opinion in neurobiology*, 11(2):231–239, 2001.

[99] Rebecca Saxe and Nancy Kanwisher. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4):1835–1842, 2003.

[100] May Tassabehji, Peter Hammond, Annette Karmiloff-Smith, Pamela Thompson, Snorri S Thorgeirsson, Marian E Durkin, Nicholas C Popescu, Timothy Hutton, Kay Metcalfe, and Agnes Rucka. Gtf2ird1 in craniofacial development of humans and mice. *Science*, 310(5751):1184–1187, 2005.

[101] EJ Young, T Lipina, E Tam, A Mandel, SJ Clapcote, AR Bechard, J Chambers, HTJ Mount, PJ Fletcher, and JC Roder. Reduced fear and aggression and altered serotonin metabolism in gtf2ird1targeted mice. *Genes, Brain and Behavior*, 7(2):224–234, 2008.

[102] Thanathom Chailangkarn, Cleber A. Trujillo, Beatriz C. Freitas, Branka Hrvoj-Mihic, Roberto H. Herai, Diana X. Yu, Timothy T. Brown, Maria C. Marchetto, Cedric Bardy, Lauren McHenry, Lisa Stefanacci, Anna Järvinen, Yvonne M. Searcy, Michelle DeWitt, Wenny Wong, Philip Lai, M. Colin Ard, Kari L. Hanson, Sarah Romero, Bob Jacobs, Anders M. Dale, Li Dai, Julie R. Korenberg, Fred H. Gage, Ursula Bellugi, Eric Halgren, Katerina Semendeferi, and Alysson R. Muotri. A human neurodevelopmental model for williams syndrome. *Nature*, 536(7616):338–343, 2016.

[103] Joan Stiles and Terry L Jernigan. The basics of brain development. *Neuropsychology Review*, 20(4):327–348, 12 2010.

[104] Aaron Alexander-Bloch, Jay N. Giedd, and Ed Bullmore. Imaging structural co-variance between human brain regions. *Nat Rev Neurosci*, 14(5):322–336, 05 2013.

[105] Susan M. Corley, Cesar P. Canales, Paulina Carmona-Mora, Veronica Mendoza-Reinosa, Annemiek Beverdam, Edna C. Hardeman, Marc R. Wilkins, and Stephen J. Palmer. Rna-seq analysis of gtf2ird1 knockout epidermal tissue provides potential insights into molecular mechanisms underpinning williams-beuren syndrome. *BMC Genomics*, 17:450, 2016. 2801[PII] 27295951[pmid] BMC Genomics.

[106] S. Desrivières, A. Lourdusamy, C. Tao, R. Toro, T. Jia, E. Loth, L. M. Medina, A. Kepa, A. Fernandes, B. Ruggeri, F. M. Carvalho, G. Cocks, T. Banaschewski, G. J. Barker, A. L. W. Bokde, C. Büchel, P. J. Conrod, H. Flor, A. Heinz, J. Gallinat, H. Garavan, P. Gowland, R. Brühl, C. Lawrence, K. Mann, M. L. P. Martinot, F. Nees, M. Lathrop, J. B. Poline, M. Rietschel, P. Thompson, M. Fauth-Bühler, M. N. Smolka, Z. Pausova, T. Paus, J. Feng, G. Schumann, and Imagen Consortium

the. Single nucleotide polymorphism in the neuroplastin locus associates with cortical thickness and intellectual ability in adolescents. *Molecular Psychiatry*, 20(2):263–274, 2015. 24514566[pmid] Mol Psychiatry.

[107] Maria Vounou, Eva Janousova, Robin Wolz, Jason L. Stein, Paul M. Thompson, Daniel Rueckert, Giovanni Montana, and Initiative Alzheimer's Disease Neuroimaging. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease. *Neuroimage*, 60(1):700–716, 2012.

[108] Peter M. Visscher. Sizing up human height variation. *Nat Genet*, 40(5):489–490, 2008. 10.1038/ng0508-489.

[109] Badam Enkhmandakh, Aleksandr V. Makeyev, Lkhamsuren Erdenechimeg, Frank H. Ruddle, Nyam-Osor Chimge, Maria Isabel Tussie-Luna, Ananda L. Roy, and Dashzeveg Bayarsaihan. Essential functions of the williams-beuren syndrome-associated tfii-i genes in embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1):181–186, 2009.

[110] Bridgett M. vonHoldt, Emily Shuldiner, Ilana Janowitz Koch, Rebecca Y. Kartzinel, Andrew Hogan, Lauren Brubaker, Shelby Wanser, Daniel Stahler, Clive D. L. Wynne, Elaine A. Ostrander, Janet S. Sinsheimer, and Monique A. R. Udell. Structural variants in genes associated with human williams-beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Science Advances*, 3(7), 2017.

[111] Xue Hua, Alex D. Leow, Suh Lee, Andrea D. Klunder, Arthur W. Toga, Natasha Lepore, Yi-Yu Chou, Caroline Brun, Ming-Chang Chiang, Marina Barysheva, Clifford R. Jack, Matt A. Bernstein, Paula J. Britson, Chadwick P. Ward, Jennifer L. Whitwell, Bret Borowski, Adam S. Fleisher, Nick C. Fox, Richard G. Boyes, Josephine Barnes, Danielle Harvey, John Kornak, Norbert Schuff, Lauren Boreta, Gene E. Alexander, Michael W. Weiner, Paul M. Thompson, and Initiative Alzheimer's Disease Neuroimaging. 3d characterization of brain atrophy in alzheimer's disease and mild cognitive impairment using tensor-based morphometry. *Neuroimage*, 41(1):19–34, 2008.

[112] Lasse-Marius Honningsvåg, Mattias Linde, Asta Håberg, Lars Jacob Stovner, and Knut Hagen. Does health differ between participants and non-participants in the mri-hunt study, a population based neuroimaging study? the nord-trøndelag health studies 1984-2009. *BMC medical imaging*, 12:23, 2012.

[113] Thomas Espeseth, Andrea Christoforou, Astri J. Lundervold, Vidar M. Steen, Stephanie Le Hellard, and Ivar Reinvang. Imaging and cognitive genetics: the norwegian cognitive neurogenetics sample. *Twin Research and Human Genetics*, 15(3):442–452, 2012.

[114] Lars M. Rimol, Ragnar Nesvåg, Don J. Hagler, Orjan Bergmann, Christine Fennema-Notestine, Cecilie B. Hartberg, Unn K. Haukvik, Elisabeth Lange, Chris J. Pung, Andres Server, Ingrid Melle, Ole A. Andreassen, Ingrid Agartz, and Anders M. Dale. Cortical volume, surface area, and thickness in schizophrenia and bipolar disorder. *Biological Psychiatry*, 71(6):552–560, 2012.

[115] Terry L. Jernigan, Timothy T. Brown, Donald J. Hagler Jr, Natacha Akshoomoff, Hauke Bartsch, Erik Newman, Wesley K. Thompson, Cinnamon S. Bloss, Sarah S. Murray, Nicholas Schork, David N. Kennedy, Joshua M. Kuperman, Connor McCabe, Yoonho Chung, Ondrej Libiger, Melanie Maddox, B. J. Casey, Linda Chang, Thomas M. Ernst, Jean A. Frazier, Jeffrey R. Gruen, Elizabeth R. Sowell, Tal Kenet, Walter E. Kaufmann, Stewart Mostofsky, David G. Amaral, and Anders M. Dale. The pediatric imaging, neurocognition, and genetics (ping) data repository. *NeuroImage*, 124, Part B:1149–1154, 2016.

[116] Chun Chieh Fan, Timothy T. Brown, Hauke Bartsch, Joshua M. Kuperman, Donald J. Hagler, Andrew Schork, Yvonne Searcy, Ursula Bellugi, Eric Halgren, and Anders M. Dale. Williams syndrome-specific neuroanatomical profile and its associations with behavioral features. *NeuroImage: Clinical*, 15:343–347, 2017.

[117] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, 2014.

[118] Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, Michael E. Goddard, and Peter M. Visscher. Common snps explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–569, 2010. 10.1038/ng.608.

[119] L. Dai, U. Bellugi, X. N. Chen, A. M. Pulst-Korenberg, A. Järvinen-Pasley, T. Tirosh-Wagner, P. S. Eis, J. Graham, D. Mills, Y. Searcy, and J. R. Korenberg. Is it williams syndrome? gtf2ird1 implicated in visual-spatial construction and gtf2i in sociability revealed by high resolution arrays. *American Journal of Medical Genetics. Part A*, 149A(3):302–314, 2009.

[120] A Antonell, M Del Campo, L F Magano, L Kaufmann, J Martínez de la Iglesia, F Gallastegui, R Flores, U Schweigmann, C Fauth, D Kotzot, and L A Pérez-Jurado. Partial 7q11.23 deletions further implicate gtf2i and gtf2ird1 as the main genes responsible for the williams–beuren syndrome neurocognitive profile. *Journal of Medical Genetics*, 47(5):312–320, 2010.

[121] Paulina Carmona-Mora, Jocelyn Widagdo, Florence Tomasetig, Cesar P. Canales, Yeojoon Cha, Wei Lee, Abdullah Alshawaf, Mirella Dottori, Renee M. Whan, Edna C. Hardeman, and Stephen J. Palmer. The nuclear localization pattern and

interaction partners of gtf2ird1 demonstrate a role in chromatin regulation. *Human Genetics*, 134(10):1099–1115, 2015.

[122] Neelroop N. Parikshak, Michael J. Gandal, and Daniel H. Geschwind. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat Rev Genet*, 16(8):441–458, 2015.

[123] Matthew S. Panizzon, Christine Fennema-Notestine, Lisa T. Eyler, Terry L. Jernigan, Elizabeth Prom-Wormley, Michael Neale, Kristen Jacobson, Michael J. Lyons, Michael D. Grant, Carol E. Franz, Hong Xian, Ming Tsuang, Bruce Fischl, Larry Seidman, Anders Dale, and William S. Kremen. Distinct genetic influences on cortical surface area and cortical thickness. *Cerebral Cortex (New York, NY)*, 19(11):2728–2735, 2009. 19299253[pmid] Cereb Cortex.

[124] William S. Kremen, Elizabeth Prom-Wormley, Matthew S. Panizzon, Lisa T. Eyler, Bruce Fischl, Michael C. Neale, Carol E. Franz, Michael J. Lyons, Jennifer Pacheco, Michele E. Perry, Allison Stevens, J. Eric Schmitt, Michael D. Grant, Larry J. Seidman, Heidi W. Thermenos, Ming T. Tsuang, Seth A. Eisen, Anders M. Dale, and Christine Fennema-Notestine. Genetic and environmental influences on the size of specific brain regions in midlife: The vetsa mri study. *NeuroImage*, 49(2):1213–1223, 2010.

[125] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980, 2006.

[126] Mert R. Sabuncu, Randy L. Buckner, Jordan W. Smoller, Phil Hyoun Lee, Bruce Fischl, and Reisa A. Sperling. The association between a polygenic alzheimer score and cortical thickness in clinically normal subjects. *Cerebral Cortex*, 22(11):2653–2661, 2012. 10.1093/cercor/bhr348.

[127] Bing Liu, Xiaolong Zhang, Yue Cui, Wen Qin, Yan Tao, Jin Li, Chunshui Yu, and Tianzi Jiang. Polygenic risk for schizophrenia influences cortical gyrification in 2 independent general populations. *Schizophrenia Bulletin*, 43(3):673–680, 2017. 10.1093/schbul/sbw051.

[128] Ole A Andreassen, Wesley K Thompson, Andrew J Schork, Stephan Ripke, Morten Mattingsdal, John R Kelsoe, Kenneth S Kendler, Michael C O'Donovan, Dan Rujescu, and Thomas Werge. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS genetics*, 9(4):e1003455, 2013.

[129] Xiaofeng Zhu, Tao Feng, Bamidele O Tayo, Jingjing Liang, J. Hunter Young, Nora Franceschini, Jennifer A Smith, Lisa R Yanek, Yan V Sun, Todd L Edwards,

Wei Chen, Mike Nalls, Ervin Fox, Michele Sale, Erwin Bottinger, Charles Rotimi, Yongmei Liu, Barbara McKnight, Kiang Liu, Donna K Arnett, Aravinda Chakravati, Richard S Cooper, and Susan Redline. Meta-analysis of correlated traits via summary statistics from gwass with an application in hypertension. *The American Journal of Human Genetics*, 96(1):21–36, 2015.

[130] Patrick Turley, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A. Furlotte, Patrik Magnusson, Sven Oskarsson, Magnus Johannesson, Peter M. Visscher, David Laibson, David Cesarini, Benjamin Neale, and Daniel J. Benjamin. Mtag: Multi-trait analysis of gwas. *bioRxiv*, 2017.

[131] Adrian Cortes, Calliope A Dendrou, Allan Motyer, Luke Jostins, Damjan Vukcevic, Alexander Dilthey, Peter Donnelly, Stephen Leslie, Lars Fugger, and Gil McVean. Bayesian analysis of genetic association across tree-structured routine healthcare data in the uk biobank. *Nat Genet*, 49(9):1311–1318, 09 2017.

[132] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, and Elise B Robinson. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 2015.

[133] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[134] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[135] Doug Speed, Na Cai, Ucleb Consortium the, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. Reevaluation of snp heritability in complex human traits. *Nat Genet*, advance online publication, 2017.

[136] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[137] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011. 10.1093/bioinformatics/btr341.

[138] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[139] Jian Yang, Teri A. Manolio, Louis R. Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M. Cunningham, Mariza de Andrade, Bjarke Feenstra, Eleanor Feingold, M. Geoffrey Hayes, William G. Hill, Maria Teresa Landi, Alvaro Alonso, Guillaume Lettre, Peng Lin, Hua Ling, William Lowe, Rasika A. Mathias, Mads Melbye, Elizabeth Pugh, Marilyn C. Cornelis, Bruce S. Weir, Michael E. Goddard, and Peter M. Visscher. Genome partitioning of genetic variation for complex traits using common snps. *Nat Genet*, 43(6):519–525, 2011. 10.1038/ng.823.

[140] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34–35, 1971. 10.1038/234034a0.

[141] Andrey A. Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012. 10.1093/bioinformatics/bts163.

[142] Andrew J. Schork, Wesley K. Thompson, Phillip Pham, Ali Torkamani, J. Cooper Roddey, Patrick F. Sullivan, John R. Kelsoe, Michael C. O'Donovan, Helena Furberg, Tobacco The, Consortium Genetics, Consortium The Bipolar Disorder Psychiatric Genomics, Consortium The Schizophrenia Psychiatric Genomics, Nicholas J. Schork, Ole A. Andreassen, and Anders M. Dale. All snps are not created equal: Genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. *PLOS Genetics*, 9(4):e1003449, 2013.

[143] Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–315, 2014.