MOLECULAR GENETIC INVESTIGATION OF TWINS, FAMILIES, AND POPULATIONS

Jeffrey John Beck

MOLECULAR GENETIC INVESTIGATION OF TWINS, FAMILIES, AND POPULATIONS

Jeffrey John Beck

Author: Jeffrey John Beck

Printing: Ridderprint, www.ridderprint.nl
Layout & cover design: Sara Terwisscha van Scheltinga, persoonlijkproefschrift.nl

ISBN: 978-94-6416-898-3

VRIJE UNIVERSITEIT

MOLECULAR GENETIC INVESTIGATION OF TWINS, FAMILIES, AND POPULATIONS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. C.M. van Praag,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op donderdag 16 december 2021 om 15.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Jeffrey John Beck

geboren te Marquette, Michigan, Verenigde Staten

promotor:	prof.dr. D.I. Boomsma
copromotoren:	dr. E.A. Ehli
	dr. J.J. Hottenga
promotiecommissie:	prof.dr. L. Henneman
	prof.dr. P. Scheet
	prof.dr. K. Eyster
	dr. A. Abdellaoui
	prof.dr. C.B. Lambalk
	prof.dr. A.H.M. Willemsen

Paranymphs:	Brandon N. Johnson
	Austin J. Van Asselt

ACKNOWLEDGMENTS

I gratefully acknowledge my parents, Allan and Patricia Beck, and my brother, Bradley Beck, and his family (Erin, Ella, Hally, Bobbie). Throughout my entire life and especially during graduate school, their unending compassion, unconditional support, and persistent encouragement motivated me never to give up and continuously pursue my dreams. I am forever appreciative of them and their unwavering patience and steadfast reassurance. I cannot possibly thank them enough.

Additionally, countless acquaintances, family members, friends, and teammates have supported me throughout my life, particularly during the completion of this thesis. Although I cannot specifically acknowledge them all, I am incredibly appreciative of every one of them.

I also want to extend a sincere thank you to my promoter, Dorret Boomsma, and my co-promoters, Jouke-Jan Hottenga and Erik Ehli. Each has been invaluable in helping me grow as an aspiring scientist in their unique ways, and I am very thankful for their supervision and guidance.

To my Dutch colleagues, I am incredibly appreciative of the hospitality and knowledge you imparted to me during numerous trips across the pond. You constantly invited me and made me feel welcome. I learned so much about Dutch culture and the Netherlands throughout these experiences. I am forever grateful for your compassion and the memories we shared.

Last but certainly not least, I would like to extend sincere appreciation to the Avera Institute for Human Genetics, the University of South Dakota, and Vrije Universiteit for fostering an educational and supportive atmosphere. For the last several years, these institutions have provided an environment that promoted my scientific development. Without the collaborative efforts of these institutions, none of the work presented in this thesis would have been possible.

TABLE OF CONTENTS

1

INTRODUCTION

TWINS AND TWIN GENETICS

Twins have intrigued human beings for centuries. The sameness exhibited by twins is fascinating, although the dissimilarities and manifestation of those differences are often even more compelling. Numerous literary texts and philosophical publications have drawn inspiration from and capitalized on the intriguing nature of twins. Some of those works highlight the unique and uncanny similarities of identical twins (e.g., *The Comedy of Errors* by William Shakespeare). In contrast, others emphasize the divergent appearance and behaviors of fraternal twins (e.g., *Twelfth Night* by William Shakespeare and biblical accounts of Jacob and Esau in the Book of Genesis).

Whether identical or fraternal, the defining characteristics of twins have also piqued the interest of scientists. For one, twins exemplify a form of a 'natural experiment,' in which the degree of genetic and environmental sharing are known or controlled for in varying amounts. In this way, twins are valuable study subjects when control for genetic background and early environmental influences is desired. Informed study designs utilize twins to estimate the contribution of genetic factors to disease, trait, or phenotype variation. Secondly, the biological and endocrinological processes fostering the twinning process also have great medical and scientific relevance. Enhanced understanding of the regulatory mechanisms underlying ovarian function, follicular growth, and maintenance of a multiple gestation pregnancy represent critical research areas on women's health and female (in)fertility. Our knowledge of the etiology of the twinning process has vastly improved in recent years, although a comprehensive understanding of identical and fraternal twins remains elusive.

Two kinds of twins exist - identical or monozygotic (MZ) twins and fraternal (non-identical) or dizygotic (DZ) twins. Both types of twins share prenatal and early environmental influences. The distinction of twin type is based on the number of independently fertilized zygotes during a single pregnancy; hence, the nomenclature monozygotic (i.e., one zygote) and dizygotic (i.e., two zygotes). MZ twins result from a postzygotic splitting event of a single fertilized zygote early in gestation resulting in a separation of cells into two or more embryos. MZ twins are therefore matched for genetic background. Alternatively, DZ twins derive from the release of two eggs, which are then independently fertilized by two sperm. Thus, DZ twins share the same amount of genetic material as ordinary siblings.

The difference in the number of fertilized zygotes dictates whether the twin pair is MZ or DZ. The universally accepted model for the generation of MZ twins is the 'fission model,' which conceptually defines the splitting of an embryo into two distinct entities within the first two weeks of development. The stage at which the splitting occurs is thought to determine membrane anatomy (chorionicity and amnionicity). Though widely adopted, the fission conjecture has been critically challenged, and alternative models have been proposed (i.e., the 'fusion model' – refer to chapter 2 for more details) [1]. The debate and remaining uncertainty around these hypotheses reiterate that the biological processes fundamental to the generation of MZ twins are still largely undetermined.

Regarding DZ twins, much more is understood concerning genetic and biological mechanisms leading to the development of two independent zygotes. Biologically, DZ twins arise from mechanisms that operate on the selection of developing ovarian follicles, where two follicles mature instead of one, releasing two oocytes that are then subsequently fertilized. A myriad of maternal factors are involved, including genetic disposition, hormonal (dys) regulation and coordination, as well as anatomical and physiological support. Despite the genetic influence on DZ twinning, it is still not possible to define the risk of having twins at an individual level.

Knowledge gaps of the twinning process persist, yet it is well understood that there are distinguishing characteristics in the mechanisms leading to MZ and DZ twins. Also evident is variation in the occurrence of MZ and DZ twins across worldwide populations. In general, MZ twinning seems to be a random yet consistent event, occurring in roughly 3 out of every 1000 maternities. Alternatively, rates of DZ twinning seem to fluctuate considerably with geography and time. Variation in DZ twinning rates have been studied extensively, with the highest rates observed in Sub-Saharan Africa (~23-40 per 1000 maternities), intermediate rates in Europe (10-20 per 1000 maternities), and the lowest rates occurring in Asia (~5-6 per 1000 maternities) [2, 3].

The various characteristics of twins and twin genetics are a central theme to this thesis. In chapter 2, an extensive review of the biology and genetics of MZ and DZ twinning is presented, highlighting what is known about twinning based on historical observations, epidemiological and clinical studies, and more recent molecular investigations. Chapter 3 overviews a molecular genetic analysis of DZ twinning in a large, multigenerational pedigree with multiple mothers of DZ twins. Chapter 4 utilizes genetic data from globally diverse twin-family populations to establish the degree of interpopulation genetic similarity using a custom-designed genotyping microarray, including genetic markers specific to DZ twinning and female fertility.

Twins are representative of the general population as participants in research projects since they tend to be born in all strata of society [4]. However, the inclusion of twin data into research projects raises additional questions. For example, to what extent can findings from genetic association studies of birth weight in singletons be generalized to twins? And can twins be included to boost sample sizes? Accordingly, chapter 4 focuses on the genetics underlying birth weight, a complex phenotype with maternal and fetal contributions, in twins compared to singletons.

Increasingly, data collection efforts for twin studies include (nuclear) family members in addition to the twin pair. Such family-based study designs offer advanced options for research and provide advantages over population-based approaches. Chapter 5 explores genetic ancestry estimates in twins and their family members as determined by popular bioinformatic methods.

THE NETHERLANDS TWIN REGISTER AT VRIJE UNIVERSITEIT

Scientific studies have exploited the differential genetic relatedness among twins by incorporating them in their designs since the early 1900s [5, 6]. Following the seminal twin studies, more elaborate study designs have been employed to partition trait variation or disease susceptibility into genetic and environmental components. The scientific merit of studies featuring twins was quickly realized, necessitating concerted efforts for gathering genetic and phenotypic data from twins and their family members.

Twin registries have increasingly become attractive resources for promoting twin and epidemiological studies by collecting biological samples and longitudinal data. The Netherlands Twin Register (NTR), maintained by the Department of Biological Psychology at Vrije Universiteit, is one example among global twin registers. Initiated in 1987 [7], the NTR has invited twins and their family members to enroll and participate in a wide variety of research studies related to behavior, development, and health. A primary research goal of the NTR is to analyze genetic and phenotypic data obtained from twins to disentangle the genetic and environmental contributions to cognitive and emotional development in childhood and adolescence as well as adult behavior, health, and well-being.

For more than thirty years [8], the NTR has curated a rich data resource for assessing genetic and nongenetic trait influences and contributing to gene-finding initiatives for complex human traits. Since its inception, more than 120,000 twins have enrolled in the twin register with nearly equal representation of their family members. Most twins and their families have participated in survey studies, with subsets generously donating some form of biological material. Longitudinal phenotyping, biological sample collection, and extended pedigrees with multigenerational representation have enabled research initiatives aimed at gene discovery, modeling causality, determining genetic inheritance, and studying population genetics.

An ongoing effort of the NTR is to collect biological samples from individuals who have provided phenotypic information. Whole blood samples are collected from adult twins to measure metabolic and immunological markers and extract genetic material (DNA and RNA) used in molecular genetic studies. In addition, adult twins and NTR participants who do not take part in biobank studies are asked to provide a swab of buccal epithelial cells, a far less invasive sample collection method. Buccal swabs are also the primary DNA collection method for young twins, their siblings, and parents. Large-scale sample collection is a primary motivation of the NTR, for which the scientific benefits are evident, especially when coupled with survey data. For twin participants themselves, receiving information on their zygosity is an extra incentive.

A primary interest of this thesis is the measurement of genetic variation from extracted DNA of the supplied biomaterial by NTR participants. Recent technological developments have enabled the direct measurement of hundreds of thousands of genetic variants at the DNA level at an affordable cost. The genetic variants, most often single nucleotide polymorphisms (SNPs), can be directly measured with polymerase chain reaction (PCR), microarray genotyping, or whole-genome sequencing. Regardless of the chosen technology, knowledge of an individual's genetic composition is attained.

Creating an extensive database of genetic measures linked with survey responses has established the NTR as a data-rich resource for studying human traits and diseases. Accordingly, the NTR has contributed to gene-discovery efforts for a considerable number of human phenotypes. In many of these studies, the NTR represents a single contribution to a concerted effort of many population-based twin registers from around the globe. Often, these efforts are organized by consortia dedicated to unraveling the genetic contributions to a particular trait or disease.

In this thesis, several aspects of the value of NTR genetic and phenotypic data are discussed. In chapter 3, a large, multigenerational Dutch pedigree with a rich history of DZ twinning is investigated using genotype and whole-genome sequence data. In chapter 4, population genetics of NTR participants are compared to other European ancestry-based populations for which data are routinely combined via consortia-led projects for gene discovery. In chapter 5, genetic data and birth weight information from twins enrolled in the NTR, and other worldwide twin registers were used to conduct a genome-wide association meta-analysis (GWAMA) on birth weight in twins. In chapter 6, genetic information from NTR participants was utilized to compare bioinformatic estimates of genetic ancestry in twins and their families.

THE AVERA TWIN REGISTER – A MIDWESTERN AMERICAN TWIN COHORT

Twin registers from around the world, especially prominent entities such as the NTR, have established the immense scientific importance of combined genetic and phenotypic data. The often-limited geographical scope of twin registers allows for meaningful comparisons of gene-trait associations between and within populations. Thus, an added incentive for establishing a population-based twin register is the opportunity to contribute to joint genetic analyses for gene discovery organized by global consortia.

The Avera Institute for Human Genetics (AIHG) in Sioux Falls, South Dakota, established the Avera Twin Register (ATR) in May of 2016 [9, 10]. The ATR is the first and only twin register in South Dakota. The primary goal of the ATR is to study the genetic and environmental influences on health, disease, and complex traits by harnessing the power of longitudinal biological sample collection and survey correspondence. Although twins and their family members have enrolled from across all United States regions, a majority come from Midwestern states, including South Dakota, North Dakota, Minnesota, Iowa, and Nebraska. In addition to serving as a prime research model for studying health and disease in a regional setting, another core goal of the ATR is to contribute to consortia-driven large-scale genetic studies focusing on the genetic underpinnings of complex traits. Furthermore, as a division of the molecular genetics lab at AIHG, the ATR prioritizes extensions of the twin design with advanced molecular assays of DNA, human microchimerism [11, 12], epigenetics (e.g., methylation) [13, 14], telomere repeat mass [15], and the gut microbiome [16, 17].

Chapter 4 assesses the degree of genetic similarity between the Midwestern American population exemplified by the ATR and global twin-family populations originating from the NTR and Australia. The work presented in chapter 4 represents the first application of genetic data collected on enrolled and consented participants of the ATR. Moreover, paired genotypic and phenotypic data for ATR participants were analyzed in chapter 5 to contribute to a global effort of gene discovery for birth weight in twins. This application represents the first usage of phenotypic data collected by the ATR.

THE NTR-AVERA COLLABORATION

Large-scale genetic investigations of complex traits have demonstrated the need for a collaborative research model to achieve the required sample sizes for answering complicated genetic questions. Often institutions/facilities are not fully equipped with both the laboratory instrumentation to systematically measure genetic variation and the computational equipment necessary to process and analyze the generated data. The disparity is exacerbated by the need for specialized knowledge and skill sets relevant to the laboratory setting and downstream bioinformatic/statistical analyses. Leveraging their expertise in wet-lab and dry-lab application, respectively, the AIHG and the NTR initiated collaboration in 2008 with the establishment of a formal agreement in May 2015. To this end, the molecular genetics laboratory at the AIHG has supplied the expertise and infrastructure necessary for performing a multitude of molecular genetic experiments.

An essential component of the collaborative agreement between the AIHG and the NTR is generating and analyzing genetic data and combining these data with health, lifestyle, and behavioral assessments to further gene discovery and contribute to projects aimed at improving physical and mental health. Microarray technologies have been the preferred method for directly measuring genetic variation. For the most part, the allure is due to the cost-effectiveness of microarrays compared to whole-genome sequencing. While not as comprehensive as sequencing-based approaches, microarrays provide genetic information in a more selective manner. Microarrays directly assay sites or regions of the human genome that are known to vary among individuals. The AIHG has offered the capability of generating genotype data with three primary microarray platforms, namely Affymetrix SNP 6.0, Affymetrix Axiom, and the Illumina Global Screening Array (GSA). Data from all three genotyping platforms were fundamental to NTR and ATR studies presented in chapters 3, 4, 5, and 6.

The NTR-Avera collaborative genotyping initiative has been extremely successful since its inception. The productive partnership has inspired fruitful contributions to several large-scale association studies spanning various human traits and disorders. Examples include aggression [18, 19], attention problems [20-25], brain structure and volume [26-29], depression [30-34], exercise behavior [35, 36], female fertility and twinning [37, 38], intelligence [39], substance use [40-42], personality [43], and a variety of others [44-47].

Exploiting the pertinent expertise of scientists at the AIHG and the NTR, the NTR-Avera partnership further enabled the design of customized DNA microarrays. The arrays were designed to enhance coverage of the human genome by utilizing suitable whole-genome reference sequences obtained from the Genome of the Netherlands project [48]. Additionally, the custom genotyping solutions were designed to include markers specific to pharmacogenomic responsiveness, cardiometabolic disease, psychiatric disorders, and other traits of particular interest, including fertility and twinning. The portfolio of custom-designed microarrays includes the Affymetrix Axiom-NL array [49] and its successor, the Illumina GSA. A detailed description of the GSA design and its first research application is provided in chapter 4.

In addition to DNA genotyping, the AIHG excels in performing additional molecular genetic methods, including, but not limited to, measurements of epigenetic signatures (i.e., DNA methylation) and whole-genome sequencing. The cost of sequencing has decreased by orders of magnitude since the completion of the Human Genome Project, reflecting Moore's Law (https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost). The more affordable price, together with the ever-growing desire to discover trait-associated genetic variants extending beyond those identified through SNP-based association studies, fosters novel research ideas that necessitate whole-genome sequence data. For example, detecting rare and functional variants associated with the DZ twinning process that cannot be identified through traditional common variant association studies. In chapter 3, the first results of a whole-genome sequencing project on a large Dutch pedigree with a rich history of DZ twinning are presented. The NTR ascertained the large pedigree for multiple mothers of spontaneous DZ twins with subsequent whole-genome sequencing at the AIHG on an Illumina Hi-Seq 2500. To date, this project represents the first NTR-Avera collaborative human whole-genome sequencing project.

SNP GENOTYPING

The human genome is enormous, comprising about three billion base pairs of DNA. The amount of biochemical individuality, or genetic variation, between any two humans is estimated at around 0.1% [50, 51]. On average, this means that about one base pair out of 1,000 will vary between any two individuals, equating to approximately 3 million total differences. However, the estimate may be higher if more complex forms of variation (e.g., copy-number variants) are considered. Given the vast number of potential genetic differences between humans, there exists a need for automated procedures to measure and analyze genetic variation. DNA microarray technologies have done just that, allowing for the assessment of thousands or even millions of genetic variants (i.e., SNPs) in a single experiment. Information obtained from these genotyping experiments is routinely utilized in population genetics, studies of pharmacogenomics, and precision medicine research.

Whereas the exact order of DNA base pairs can be determined in a targeted or whole-genome approach with sequencing methodologies, SNP genotyping more selectively defines which specific genetic variants an individual possesses. Given that greater than 10 million common genetic variants are likely to exist [52], SNP genotyping aims to measure a fraction of these variants directly, often between 100,000 and 1,000,000. Thus, these methods rely upon knowing where key variation exists within the genome. Identifying the minimal set of SNPs needed to genotype the entire human genome was the driving force behind establishing the International HapMap project [53].

Concerted whole-genome sequencing efforts like the 1000 Genomes Project [54] and the Genome of the Netherlands [48] have been invaluable for providing a genetic 'roadmap' of the human genome, highlighting important areas of genetic variation that exist between individuals and populations. These large-scale initiatives represent genetic reference databases used to help design and enable content selection for SNP genotyping microarrays. Additionally, these genetic reference panels indirectly enhance the information obtained from SNP genotyping experiments through a process called imputation. Imputation allows for the statistical inference of genotypes not directly measured. As a result, SNP genotyping followed by imputation yields datasets mimicking whole-genome sequencing experiments, but at a fraction of the cost.

While direct genotyping of SNPs throughout the genome provides a snapshot of the genetic variation per individual, it by no means is a comprehensive evaluation. Imputation of genotypic data is a commonly employed technique for bolstering the amount of genetic information that can be gleaned from a SNP genotyping experiment. A fundamental principle of imputation is the occurrence of non-random association of alleles at different locations within the genome, termed linkage disequilibrium (LD). Because of linked allelic information, direct measurement of all genetic markers is not necessary. Imputation leverages the direct measurement of carefully selected genetic variants in regions of little recombination, known as LD or haplotype blocks. Measurement of selected variants coupled with knowledge of associated alleles allows one to infer information about the genetic variants not directly assayed. The most informative markers constitute a core 'backbone' of genetic variants used for imputation in this context. The backbone is based on commonly utilized reference panels so that the remainder of genetic variants can be inferred from large population databases of whole-genome sequence data.

The technique of imputation has become an essential tool for geneticists performing genome-wide association studies. Imputation increases the power of genome-wide association scans by enhancing the number of genetic variants for association testing. Furthermore, imputation is particularly useful when aggregating genetic data obtained from different genotyping platforms comprised of other SNP markers, a persistent occurrence in the era of modern association studies. The studies described in chapters 4 and 5 were reliant upon genotype imputation.

Creating trustworthy technologies and accurate methods for interrogating regions of genetic variation is pivotal for obtaining reliable genotypes. Opposed to generating numerous reads per base in whole-genome sequencing (i.e., coverage depth), any given marker on a genotyping array is usually only measured once unless probes for a particular variant exist in replicate. Thus, the array design and experiment chemistry must be robust to provide accurate results. Fortuitously, pioneering biotechnology companies, such as Affymetrix (now part of Thermo Fisher Scientific), Illumina, and others, have devoted years of scientific expertise to optimizing genotyping solutions. The AIHG has routinely employed these products to facilitate large-scale genotyping projects. Furthermore, through the NTR-Avera collaborative partnership, scientists have worked hand-in-hand with biotechnology companies to design population-optimized custom genotyping arrays, such as the Axiom-NL array [49] and a customized Illumina Global Screening Array (described in chapter 4). These SNP microarrays were carefully designed to possess a core imputation backbone and additional content related to phenotypes of interest.

Design of a customized SNP genotyping microarray involves providing the array manufacturer with a list of targets in the form of genomic coordinates (i.e., the chromosome number with start and end base pair positions), SNP reference numbers (i.e., rsIDs), or gene regions (i.e., gene names and the number of bases upstream/downstream). The user-supplied targets dictate probe design by the manufacturer. Probes are short synthetically made DNA molecules, known as oligonucleotides, that interrogate a specific molecular region or site through complementary binding. Genomic target information is typically entered via a web application, which subsequently reports metrics regarding the likely predictive performance of each probe given the supplied target. Estimated probe performance is a function of the region that flanks the target and how difficult it is to design probes that will uniquely and reliably bind.

During array manufacturing, probe sequences complementary to the target are spotted or synthesized directly onto an immobilized glass or silicon surface (e.g., silica microbeads) using various technologies, including photolithography. The resulting product is a SNP microarray that somewhat resembles a microscope slide, capable of assaying hundreds of thousands or millions of genotypes per individual at once. Often the microarrays are combined into a bead chip or plate format, enabling the simultaneous genotyping of many individuals (e.g., 24 for Illumina GSA bead chips, 96 for Axiom-NL array plates).

Regardless of the array manufacturer, a SNP genotyping experiment typically involves six primary steps. Shown in Figure 1.1 is an example SNP genotyping workflow with the Axiom array. In a first step, high-quality DNA is extracted and amplified to make copies of the genetic material. Secondly, the whole genome amplified DNA is subjected to enzymatic or mechanical fragmentation procedures to cleave the DNA into pieces. In a third step, the fragmented DNA is precipitated. Fourthly, the precipitated fragmented DNA is resuspended. In the fifth step, DNA is hybridized to the microarray containing the probes, or complementary oligonucleotides. Washing removes any non-complementary DNA that does not bind. Lastly, in a final ligation/extension and staining step, DNA successfully hybridizing to the array will emit fluorescent signals that are imaged with sophisticated instrumentation. The fluorescent signals are then analyzed and converted into raw genotype calls for downstream analysis using array-specific, and often proprietary software.

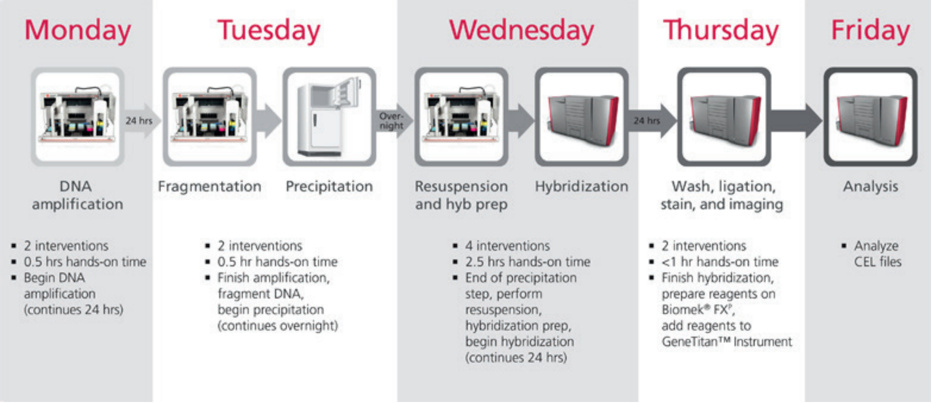

Monday
Tuesday
Wednesday
Thursday
Friday
24 hrs
Over-night
24 hrs
DNA amplification
Fragmentation
Precipitation
Resuspension and hyb prep
Hybridization
Wash, ligation, stain, and imaging
Analysis
▪ 2 interventions
▪ 0.5 hrs hands-on time
▪ Begin DNA amplification (continues 24 hrs)
▪ 2 interventions
▪ 0.5 hr hands-on time
▪ Finish amplification, fragment DNA, begin precipitation (continues overnight)
▪ 4 interventions
▪ 2.5 hrs hands-on time
▪ End of precipitation step, perform resuspension, hybridization prep, begin hybridization (continues 24 hrs)
▪ 2 interventions
▪ <1 hr hands-on time
▪ Finish hybridization, prepare reagents on Biomek® FXᴾ, add reagents to GeneTitan™ Instrument
▪ Analyze CEL files

Figure 1.1 - Workflow for the Axiom array (image obtained from https://www.affymetrix.com/products_services/instruments/specific/axiom_atp.affx)

Several aspects of SNP genotyping and the data obtained from the experiments were fundamental to the work presented in this thesis. During the first two years of my Ph.D. trajectory, I spent considerable time performing SNP genotyping experiments with the Axiom-NL array on thousands of NTR participants. As the principal performer of these genotyping experiments, I gained an immense understanding of the SNP genotyping process by troubleshooting and optimizing high-throughput strategies. The extensive genotyping effort produced substantial datasets that were integral to many elements of this thesis. In chapter 2, the genetics of twins are described, in which SNP genotyping is discussed as the most reliable strategy for determining the zygosity status of same-sex twin pairs. Also, chapter 2 provides an overview of research identifying the first genetic variants associated with DZ twinning, which were established through SNP genotyping experiments and subsequent imputation. In chapter 3, SNP genotype data were used to locate genetic regions co-segregating with being a mother of DZ twins. In chapter 4, a detailed description of the Illumina GSA design and its implementation is provided. Chapter 5 details the first genome-wide association study in the ATR and the findings of a GWAMA of twin birth weight, which utilized imputed and harmonized SNP genotype data on 42,212 twins from eight global population cohorts. Lastly, chapter 6 presents an empirical evaluation of ancestry estimation in twins and families based on genetic data from three distinct SNP genotyping arrays. The NTR has uniquely generated SNP genotype data on both members of a MZ twin pair.

GENOME-WIDE ASSOCIATION STUDIES

Over the years, significant scientific interest has been devoted to identifying associations between genotype and phenotype. Ultimately, these studies' underlying goal is to better understand trait or disease etiology to improve prediction, prevention, or treatment. Although there is a long history of study designs and strategies for elucidating genotype-phenotype associations (e.g., candidate gene studies, linkage studies in multi-generation pedigrees and sibling pairs), much of the effort in the last 10-15 years can be ascribed to genome-wide association studies (GWAS). In part, this is due to what is sometimes called the hypothesis-free nature underlying the GWAS design. GWAS do not require *a priori* knowledge or selection of interesting genetic variants related to a particular trait. However, they do test the hypothesis that a variant or multiple variants are associated with a trait. As opposed to candidate-gene driven approaches, a benefit of GWAS is that findings can often yield results that otherwise would not have been considered. This characteristic of GWAS has made it a popular and astonishingly successful study design, with greater than 270,000 trait-variant associations described in more than 5,000 publications to date (https://www.ebi.ac.uk/gwas/).

Despite the success and remarkable range of discoveries made by GWAS in recent years, challenges are still encountered when dealing with intrinsic limitations. Methodologically, GWAS utilize large samples of human genetic data (millions of SNPs) and phenotypic information to detect association by simultaneously testing the effect of genetic variants on a particular phenotype [55-58]. In this manner, the practice of performing millions of concurrent statistical tests necessitates stringent statistical correction to account for multiple testing and the potential for false positives. Thus, an omnipresent concern for GWAS is adequate statistical power to account for these corrections and the potentially small effects of individual SNPs.

The power to detect the associations between genetic variants and a trait depends on the chosen significance level, the experimental sample size, the effect size of the variant(s), the measurement of the phenotype, and several other factors. A genome-wide significance P-value threshold of $5x10^{-8}$ has become the standard in GWAS to account for the vast number of statistical tests being done [59, 60]. Although widely adopted, even more stringent thresholds have been suggested for studies using lower frequency genetic variants [61]. Thus, increasing sample sizes can enhance the power for detection. One sensible approach is to aggregate relevant data from as many resources as possible. In doing so, concerns are raised regarding systematic differences in allele frequencies that can occur by incorporating individuals from various (sub)populations, leading to spurious results and false positives. This phenomenon is termed population stratification and represents an important source of confounding that must be appropriately addressed in any GWAS [62].

GWAS and the essential concepts underlying their design represent a substantial portion of the work presented in this thesis. Broadly, the idea of population data aggregation for achieving adequate statistical power in a GWAS is empirically evaluated in chapter 4. Furthermore, a meta-analysis of GWAS results on birth weight in twins from worldwide twin registers is described in chapter 5. Chapter 6 utilizes genetic data to compare estimates of genetic ancestry, which are commonly employed steps in association studies and important indicators of population structure.

PRINCIPAL COMPONENTS ANALYSIS

In genetic association studies, principal component analysis (PCA) is an important method for summarizing genetic variation. PCA is a very general mathematical approach commonly utilized for dimensionality reduction of large data sets. PCA works by transforming many variables into a smaller group of uncorrelated variables containing all the information in the original data. The (much) smaller set of variables preserve most of the variation in the data and are called principal components (PCs). In general, PCA summarizes the most prominent patterns of variation in a data set containing numerous measurements.

In genetics, PCA condenses many individuals' genetic variation (typically tens or hundreds of thousands of genetic markers) into a relatively small number (often around 10) of PCs. The patterns of variation defined by PCA from genetic datasets of global populations have been shown to reflect ancestry differences and correlate with geography [63]. However, if implemented incorrectly, the validity of these relationships can be biased [64]. Regardless, PCA has been increasingly utilized to infer population (sub)structure from genetic data since its first adaptation for human genomic data in 1963 [65]. PCA is valuable in GWAS for adequately accounting for population structure. In practice, the PCs summarizing ancestry and other variation in the genotype data are used as covariates in association models.

In this thesis, PCA was instrumental for the work presented in several chapters. In chapter 4, PCA of Australian, Dutch, and Midwestern American individuals genotyped at the AIHG on the Illumina GSA was performed to visualize and compare population structure. PCA was also utilized to project study samples onto PCs calculated by a global reference set acquired from the Human Genome Diversity Project. The projection procedure allowed for a broader resolution assessment of the genetic similarity of the populations of interest. Chapter 5 utilized PCA and PCs to correct for cohort-specific metrics (e.g., genetic ancestry, genotyping platform, genotyping batch) in each GWAS of birth weight that contributed to the overall meta-analysis. Genomic PCs were also used as covariates in the predictive model of birth weight with polygenic scores. Chapter 6 applied PCA as a primary method for inferring genetic ancestry, in which comparisons of the resultant PCs were evaluated within twins and family members.

HYPOTHESES AND OBJECTIVES OF THE DISSERTATION

This dissertation broadly examines the genetics of the twinning phenomenon, twins, global twin-family populations, and the representativeness of twins in GWAS by employing various analytical approaches and study designs.

The second chapter of this thesis provides background information on the biology and genetics of twins. This chapter describes what is well understood and what remains mysterious regarding our scientific understanding of the twinning process. Key distinctions are made between the two types of twins, monozygotic (MZ) and dizygotic (DZ). The chapter's central theme is the unique characteristics of MZ and DZ twins, specifically differentiating their respective biology, epidemiology, genetics, and incidence. Moreover, twins and the processes underlying twinning have been extensively studied for years, yet a complete understanding of the mechanisms and contributory factors is still lacking. With that in mind, substantially more is known about the etiology of DZ twinning in part due to recent advances in molecular technologies and the illuminating power of gene discovery afforded by genetic association studies. Despite these developments and numerous scientific efforts, the processes underlying MZ twinning remain largely unresolved. Regardless of the lingering uncertainties, twins remain a precious resource for studying genetics and complex traits, especially in the 'omics' era [66].

In chapters 3 through 6, specific scientific questions related to twinning and the broader field of human genetics are addressed with data from twins, their families, and the populations they represent. The feasibility of these studies was contingent on the availability of sizable genotypic data sets from worldwide twin and family cohorts.

The focus of chapter 3 is the genetic influences of DZ twinning. Throughout many years' worth of twin studies on behavioral traits, many researchers realized the strong tendency for DZ twinning to run in families. This recognition prompted many segregation and pedigree analyses to illuminate the genetics of DZ twinning. Here, we expand upon these studies by identifying a multigenerational pedigree with many mothers of DZ twins to further uncover genetic factors associated with DZ twinning. The study's objective was to use pedigree-based genotypic and sequence data to identify genetic variants influencing a mother's propensity to conceive DZ twins. The project intended to discover novel rare/structural variants extending beyond the common variants with established DZ twinning associations. We hypothesized that we could identify large genetic regions of interest co-segregating in mothers of DZ twins by analyzing genetic data from available family members through linkage analysis. Furthermore, we expected that the common areas would harbor rare variants of large effect and that substantially influence DZ twinning. More broadly, determining whether the identified variants are pedigree specific or possessed by a larger group of mothers of DZ twins necessitated evaluation against population-matched controls.

Chapter 4 leverages the vast amount of information contained within genotypic data to evaluate the genetic similarity of global populations. First, I provide a detailed description of the design of a genotyping microarray that facilitated the work in this chapter and much of the remaining thesis content. Implementation of the microarray enabled the generation of a wealth of genotypic data representative of individuals from three globally distinct populations, namely Australian, Dutch, and Americans from the Midwestern region of the United States. We hypothesized to find comparable estimates of genetic similarity between the populations since they each have ancestral origins from Europe. Comparisons were augmented with worldwide reference data and an auxiliary genetic dataset generated from the same microarray obtained from a Nigerian population.

Chapter 5 builds on the findings of the previous chapter in that genetic data from European ancestry-based populations were aggregated to investigate the genetics of birth weight in twins. Birth weight is an important indicator of overall health, and critical links between low birth weight and higher risks of perinatal morbidity and mortality have been established [67-71]. Furthermore, a wealth of evidence has demonstrated an impact of birth weight and diseases in adulthood [72], including type 2 diabetes [73], cardiovascular disease [74, 75], high blood pressure [76-79], psychological distress [80], and body mass index [81, 82]. For these reasons, birth weight has been studied extensively and is known to be influenced by genetic and environmental factors [83]. Genetically, variation in birth weight is complicated by the effects of fetal and maternal genes [84-86]. Although complex, most of what is understood about birth weight genetics has been established by studies of singletons. Twins are often excluded due to their, on average, lower birth weight and the uncertainty of differing genetic influences when compared to singletons. There has only been one published GWAS of the genetics of birth weight in twins, precisely 4,593 female twins from the United Kingdom, in which one genome-wide significant signal was identified [87]. While the findings provided the first insight into the genetic factors influencing birth weight in twins, much remained to be determined concerning how the genetic component of birth weight compares in twins and singletons. While differences in average birth weight between twins and singletons exist, we hypothesized that the common genetic effects influencing birth weight are similar between the groups. We additionally aimed to identify novel genetic variants associated with birth weight in twins by meta-analyzing GWAS results supplied by eight twin cohorts. An indication of the genetic overlap was determined by comparing meta-analysis results to those previously reported for singletons.

Chapter 6 analyzes genetic information from NTR twins and family members to examine estimates of genetic ancestry. Ancestry inference is pivotal in association studies since systematic differences between groups can confound real association signals leading to spurious results. Therefore, ancestry estimation strategies are routinely employed to mitigate or eliminate the effects of population stratification by including ancestry-specific covariates in the association models or excluding outliers. Another method for diminishing the impact of population stratification is to employ a family-based design in which genetic relatedness is accounted for, and ancestry is essentially under control. However, remaining ambiguities surround this approach in that a comprehensive assessment of genetic ancestry estimation has not been performed within families and between sets of twins. We utilized genotypic data of many NTR participants to address this uncertainty, including independently genotyped MZ twins, DZ twins, siblings, and parents to estimate genetic ancestry within families. For this project, genotypic data was supplied by three different microarrays (Affymetrix 6, Affymetrix Axiom-NL, and Illumina GSA), enabling additional evaluation as a function of the genotyping platform. This aspect of the study is particularly relevant in studies where data are combined across study cohorts and genotyping platforms. The array differences have the potential to impact ancestry estimates. We aimed to address this concern by estimating genetic ancestry using standard algorithmic (i.e., PCA) and model-based approaches. Both methods have their inherent benefits and limitations, and the resulting estimates reflect different parameterizations of genetic ancestry. The hypothesis that genetic ancestry estimates would show fewer differences between more closely related family members, independent of the genotyping array, was thoroughly tested in this manner.

Chapter 7 summarizes the main findings of the preceding chapters and provides overall conclusions of the work. In what follows, I present my perspective on the future of twin studies and population genetics. Finally, in chapter 8, I offer concise summaries of the studies presented in chapters 2-6.

REFERENCES

1. Herranz, G., *The timing of monozygotic twinning: a criticism of the common model.* Zygote, 2013. **23**(1): p. 27-40.
2. Bulmer, M.G., *The Biology of Twinning in Man.* 1970: Clarendon. 206.
3. Hall, J.G., *Twinning.* Lancet, 2003. **362**(9385): p. 735-43.
4. Martin, N., D. Boomsma, and G. Machin, *A twin-pronged attack on complex traits.* Nat Genet, 1997. **17**(4): p. 387-92.
5. Poll, H., *Über Zwillingsforschung als Hilfsmittel menschlicher Erbkunde.* Zeitschrift für Ethnologie, 1914. **46**: p. 87-105.
6. Siemens, H.W., *Die Zwillingspathologie.* Mol. Gen. Genet., 1924. **35**: p. 311-312.
7. Boomsma, D.I., J.F. Orlebeke, and G.C. van Baal, *The Dutch Twin Register: growth data on weight and height.* Behav Genet, 1992. **22**(2): p. 247-51.
8. Ligthart, L., et al., *The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs.* Twin Res Hum Genet, 2019. **22**(6): p. 623-636.
9. Kittelsrud, J., et al., *Avera Twin Register: Growing through Online Consenting and Survey Collection.* Twin Research and Human Genetics, 2019(Special Issue on Twin Registers).
10. Kittelsrud, J., et al., *Establishment of the Avera Twin Register in the Midwest USA.* Twin Res Hum Genet, 2017. **20**(5): p. 414-418.
11. Peters, H.E., et al., *Low prevalence of male microchimerism in women with Mayer-Rokitansky-Kuster-Hauser syndrome.* Hum Reprod, 2019. **34**(6): p. 1117-1125.
12. Johnson, B.N., et al., *Male microchimerism in females: a quantitative study of twin pedigrees to investigate mechanisms.* Hum Reprod, 2021.
13. Odintsova, V.V., et al., *Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles.* Frontiers in Psychiatry, 2021. **12**: p. 1141.
14. van Dongen, J., et al., *Identical twins carry a persistent epigenetic signature of early genome programming.* Nature Communications, In press.
15. Finnicum, C.T., et al., *Relative Telomere Repeat Mass in Buccal and Leukocyte-Derived DNA.* PLoS One, 2017. **12**(1): p. e0170765.
16. Finnicum, C.T., et al., *Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk.* BMC Microbiol, 2019. **19**(1): p. 230.
17. Finnicum, C.T., et al., *Metataxonomic Analysis of Individuals at BMI Extremes and Monozygotic Twins Discordant for BMI.* Twin Res Hum Genet, 2018. **21**(3): p. 203-213.
18. Pappa, I., et al., *A genome-wide approach to children's aggressive behavior: The EAGLE consortium.* Am J Med Genet B Neuropsychiatr Genet, 2016. **171**(5): p. 562-72.
19. Ip, H.F., et al., *Genetic Association Study of Childhood Aggression across raters, instruments and age.* bioRxiv, 2021: p. 854927.
20. Ehli, E.A., et al., *De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on Attention Problems.* Eur J Hum Genet, 2012. **20**(10): p. 1037-43.
21. Groen-Blokhuis, M.M., et al., *A prospective study of the effects of breastfeeding and FADS2 polymorphisms on cognition and hyperactivity/attention problems.* Am J Med Genet B Neuropsychiatr Genet, 2013. **162B**(5): p. 457-65.
22. de Zeeuw, E.L., et al., *Polygenic scores associated with educational attainment in adults predict educational achievement and ADHD symptoms in children.* Am J Med Genet B Neuropsychiatr Genet, 2014. **165B**(6): p. 510-20.
23. Abdellaoui, A., et al., *CNV Concordance in 1,097 MZ Twin Pairs.* Twin Res Hum Genet, 2015. **18**(1): p. 1-12.
24. Middeldorp, C.M., et al., *A Genome-Wide Association Meta-Analysis of Attention-Deficit/Hyperactivity Disorder Symptoms in Population-Based Pediatric Cohorts.* J Am Acad Child Adolesc Psychiatry, 2016. **55**(10): p. 896-905 e6.
25. Demontis, D., et al., *Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder.* Nat Genet, 2019. **51**(1): p. 63-75.
26. Hibar, D.P., et al., *Common genetic variants influence human subcortical brain structures.* Nature, 2015. **520**(7546): p. 224-9.
27. Adams, H.H., et al., *Novel genetic loci underlying human intracranial volume identified through genome-wide association.* Nat Neurosci, 2016. **19**(12): p. 1569-1582.
28. Hibar, D.P., et al., *Novel genetic loci associated with hippocampal volume.* Nat Commun, 2017. **8**: p. 13624.
29. Satizabal, C.L., et al., *Genetic architecture of subcortical brain structures in 38,851 individuals.* Nat Genet, 2019. **51**(11): p. 1624-1636.
30. Middeldorp, C.M., et al., *The genetic association between personality and major depression or bipolar disorder. A polygenic score analysis using genome-wide association data.* Transl Psychiatry, 2011. **1**: p. e50.
31. Benke, K.S., et al., *A genome-wide association meta-analysis of preschool internalizing problems.* J Am Acad Child Adolesc Psychiatry, 2014. **53**(6): p. 667-676 e7.
32. Okbay, A., et al., *Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses.* Nat Genet, 2016.
33. Mbarek, H., et al., *Genome-Wide Significance for PCLO as a Gene for Major Depressive Disorder.* Twin Res Hum Genet, 2017. **20**(4): p. 267-270.
34. Wray, N.R., et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nat Genet, 2018. **50**(5): p. 668-681.
35. Huppertz, C., et al., *A twin-sibling study on the relationship between exercise attitudes and exercise behavior.* Behav Genet, 2014. **44**(1): p. 45-55.
36. Mee, D.J.v.d., et al., *Dopaminergic Genetic Variants and Voluntary Externally Paced Exercise Behavior.* Medicine & Science in Sports & Exercise, 2018. **50**: p. 70P 708.
37. Barban, N., et al., *Genome-wide analysis identifies 12 loci influencing human reproductive behavior.* Nat Genet, 2016.
38. Mbarek, H., et al., *Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility.* Am J Hum Genet, 2016. **98**(5): p. 898-908.
39. Franic, S., et al., *Intelligence: shared genetic basis between Mendelian disorders and a polygenic trait.* Eur J Hum Genet, 2015. **23**(10): p. 1378-83.
40. Baumert, J., et al., *No evidence for genome-wide interactions on plasma fibrinogen by smoking, alcohol consumption and body mass index: results from meta-analyses of 80,607 subjects.* PLoS One, 2014. **9**(12): p. e111156.
41. Minica, C.C., et al., *Genome-wide association meta-analysis of age at first cannabis use.* Addiction, 2018. **113**(11): p. 2073-2086.

42. Liu, M., et al., *Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use.* Nat Genet, 2019. **51**(2): p. 237-244.

43. van den Berg, S.M., et al., *Meta-analysis of Genome-Wide Association Studies for Extraversion: Findings from the Genetics of Personality Consortium.* Behav Genet, 2016. **46**(2): p. 170-82.

44. Gieger, C., et al., *New gene functions in megakaryopoiesis and platelet formation.* Nature, 2011. **480**(7376): p. 201-8.

45. Ibrahim-Verbaas, C.A., et al., *GWAS for executive function and processing speed suggests involvement of the CADM2 gene.* Mol Psychiatry, 2016. **21**(2): p. 189-197.

46. Sabater-Lleal, M., et al., *Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease.* Circulation, 2013. **128**(12): p. 1310-24.

47. Wain, L.V., et al., *Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure.* Nat Genet, 2011. **43**(10): p. 1005-11.

48. Genome of the Netherlands, C., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet, 2014. **46**(8): p. 818-25.

49. Ehli, E.A., et al., *A method to customize population-specific arrays for genome-wide association testing.* Eur J Hum Genet, 2017. **25**(2): p. 267-270.

50. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nature Reviews Genetics, 2006. **7**(2): p. 85-97.

51. Ku, C.S., et al., *The discovery of human genetic variations and their use as disease markers: past, present and future.* J Hum Genet, 2010. **55**(7): p. 403-15.

52. International HapMap, C., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.

53. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

54. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**: p. 68 - 74.

55. Visscher, P.M. and G.W. Montgomery, *Genome-wide association studies and human disease: from trickle to flood.* JAMA, 2009. **302**(18): p. 2028-9.

56. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.

57. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.

58. Claussnitzer, M., et al., *A brief history of human disease genetics.* Nature, 2020. **577**(7789): p. 179-189.

59. Pe'er, I., et al., *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genet Epidemiol, 2008. **32**(4): p. 381-5.

60. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

61. Fadista, J., et al., *The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants.* Eur J Hum Genet, 2016. **24**(8): p. 1202-5.

62. Price, A.L., et al., *New approaches to population stratification in genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 459-63.

63. Novembre, J., et al., *Genes mirror geography within Europe.* Nature, 2008. **456**(7218): p. 98-101.

64. Elhaik, E., *Why most Principal Component Analyses (PCA) in population genetic studies are wrong.* bioRxiv, 2021: p. 2021.04.11.439381.

65. Cavalli-Sforza, L.L. and A.W.F. Edwards. *Analysis of Human Evolution*. in *Genetics Today, Proceedings of the 11th International Congress of Genetics*. 1963. the Hague, the Netherlands: Pergamon, New York.

66. van Dongen, J., et al., *The continuing value of twin studies in the omics era.* Nat Rev Genet, 2012. **13**(9): p. 640-53.

67. Battaglia, F.C. and L.O. Lubchenco, *A practical classification of newborn infants by weight and gestational age.* J Pediatr, 1967. **71**(2): p. 159-63.

68. Wilcox, A.J. and I.T. Russell, *Birthweight and perinatal mortality: II. On weight-specific mortality.* Int J Epidemiol, 1983. **12**(3): p. 319-25.

69. Wilcox, A.J., *Birth weight and perinatal mortality: the effect of maternal smoking.* Am J Epidemiol, 1993. **137**(10): p. 1098-104.

70. McIntire, D.D., et al., *Birth weight in relation to morbidity and mortality among newborn infants.* N Engl J Med, 1999. **340**(16): p. 1234-8.

71. Wilcox, A.J., *On the importance--and the unimportance--of birthweight.* Int J Epidemiol, 2001. **30**(6): p. 1233-41.

72. Barker, D.J. and P.M. Clark, *Fetal undernutrition and disease in later life.* Rev Reprod, 1997. **2**(2): p. 105-12.

73. Whincup, P.H., et al., *Birth weight and risk of type 2 diabetes: a systematic review.* JAMA, 2008. **300**(24): p. 2886-97.

74. Eriksson, M., et al., *Birth weight and cardiovascular risk factors in a cohort followed until 80 years of age: the study of men born in 1913.* J Intern Med, 2004. **255**(2): p. 236-46.

75. Wang, S.F., et al., *Birth weight and risk of coronary heart disease in adults: a meta-analysis of prospective cohort studies.* J Dev Orig Health Dis, 2014. **5**(6): p. 408-19.

76. Law, C.M. and A.W. Shiell, *Is blood pressure inversely related to birth weight? The strength of evidence from a systematic review of the literature.* J Hypertens, 1996. **14**(8): p. 935-41.

77. RG, I.J., C.D. Stehouwer, and D.I. Boomsma, *Evidence for genetic factors explaining the birth weight-blood pressure relation. Analysis in twins.* Hypertension, 2000. **36**(6): p. 1008-12.

78. Jarvelin, M.R., et al., *Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort.* Hypertension, 2004. **44**(6): p. 838-46.

79. Gamborg, M., et al., *Birth weight and systolic blood pressure in adolescence and adulthood: meta-regression analysis of sex- and age-specific results from 20 Nordic studies.* Am J Epidemiol, 2007. **166**(6): p. 634-45.

80. Cheung, Y.B., et al., *Birthweight and psychological distress in adult twins: a longitudinal study.* Acta Paediatr, 2004. **93**(7): p. 965-8.

81. Sorensen, H.T., et al., *Relation between weight and length at birth and body mass index in young adulthood: cohort study.* BMJ, 1997. **315**(7116): p. 1137.

82. Johansson, M. and F. Rasmussen, *Birthweight and body mass index in young adulthood: the Swedish young male twins study.* Twin Res, 2001. **4**(5): p. 400-5.

83. Yokoyama, Y., et al., *Genetic and environmental factors affecting birth size variation: a pooled individual-based analysis of secular trends and global geographical differences using 26 twin cohorts.* Int J Epidemiol, 2018. **47**(4): p. 1195-1206.

84. Warrington, N.M., et al., *Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors.* Nat Genet, 2019. **51**(5): p. 804-814.

85. Moen, G.H., et al., *Mendelian randomization study of maternal influences on birthweight and future cardiometabolic risk in the HUNT cohort.* Nat Commun, 2020. **11**(1): p. 5404.

86. Srivastava, A.K., et al., *Haplotype-based heritability estimations reveal gestational duration as a maternal trait and fetal size measurements at birth as fetal traits in human pregnancy.* bioRxiv, 2020: p. 2020.05.12.079863.

87. Metrustry, S.J., et al., *Variants close to NTRK2 gene are associated with birth weight in female twins.* Twin research and human genetics : the official journal of the International Society for Twin Studies, 2014. **17**(4): p. 254-61.

2

BIOLOGY AND GENETICS OF DIZYGOTIC AND MONOZYGOTIC TWINNING

Published as
Beck, J.J., Bruins, S., Mbarek, H., Davies, G.E., Boomsma, D.I. (2021) *Biology and Genetics of Dizygotic and Monozygotic Twinning.* In: Khalil, A., Lewi, L., Lopriore, E. (*Eds*) *Twin and higher-order pregnancies*, Springer Nature, doi.org 10.1007/978-3-030-47652-6

ABSTRACT

This chapter summarizes what is known and what remains unknown about the human twinning process. The interest of this chapter is a description of the processes underlying twinning, with a specific focus on the biological and genetic aspects. While the mechanisms and contributory factors to dizygotic twinning are becoming well established, much remains unknown about the etiology of monozygotic twinning. Here we provide an overview of the incidence of twinning across the globe and present what is known about the influences of twinning based on the findings of historical, epidemiological, and more recent molecular studies.

Keywords

- Dizygotic twins
- Monozygotic twins
- Genetics
- Assisted reproductive technologies
- Twinning rates

Definitions

1. Zygosity – The number of zygotes that become fertilized leading to a multiple birth, or the genetic makeup of the pregnancy.
2. Dizygotic twins – Non-identical or fraternal twins that are the result of two independent ova that are fertilized by two separate spermatozoa.
3. Monozygotic twins – Identical twins that arise from a single fertilized ovum.

Learning Objectives

1. Define and differentiate between the biological mechanisms that give rise to monozygotic and dizygotic twins.
2. Compare and contrast the composition of fetal membranes of monozygotic and dizygotic twins.
3. Describe the genetic and non-genetic factors that are associated with spontaneous dizygotic twinning events.
4. Explain the strengths and weaknesses of zygosity assessment methods.
5. Describe the differences in rates of monozygotic and dizygotic twins.
6. Introduce the first genome-wide association study of dizygotic twinning.

INTRODUCTION

Twins and higher-order multiples have piqued the interest of humankind for many centuries. The remarkable similar resemblance often ascribed to twins has been observed in many literary texts and philosophical works. Twins with similar outward appearances but with noticeable personality differences have been well characterized throughout history. Observations of twins are noted as far back in time as the Biblical accounts of Jacob and Esau, by philosophers Augustine of Hippo and Aristotle, and by poets like Shakespeare (e.g., *The Comedy of Errors* and *Twelfth Night*).

From a scientific perspective, it was in the nineteenth century that the Scottish obstetrician J Matthews Duncan recognized and documented that two types of twins existed, now commonly referred to as identical and fraternal (non-identical) twins [1]. Sir Francis Galton was the first the recognize the value of studying twins to elucidate the genetic contribution to variation in human traits [2]; however, he was not aware of the distinction between monozygotic (MZ or identical) and dizygotic (DZ or fraternal) twins. Even in the early twentieth century, the existence of two types of twins was debated. The famous statistician Sir Ronald Fisher (who was the second of twins himself) proved mathematically that it was highly unlikely that there was more than one type of twin [3]. Nevertheless, the idea put forth by Galton was to compare trait concordance in twins in an attempt at disentangling the genetic (nature) and environmental (nurture) influences. Galton's proposal laid the foundation for modern era twin studies aimed at discerning the genetic contribution of complex traits and etiology of disease.

The theoretical basis of quantitative genetics proved to be fundamental to the creation and application of the classical twin design, which builds on the – by now firmly established – differential genetic relatedness of MZ and DZ twin pairs. Realizing the enormous potential of quantitative genetic theory to studies that apply the classical twin design for studying human traits, large twin registries were established in the 1950s [4], although studies of twins had already been done in Russia [5-7] and elsewhere. The first scientific studies of twins in medicine were by Poll [8] and Siemens, who investigated the different levels of similarity between MZ and DZ twins for mole counts. Findings from their work suggested that MZ twin pairs were nearly genetically identical, whereas DZ twin pairs shared on average 50% of their genetic variation; this was reflected in the twice as large resemblance for mole counts in MZ than in DZ twin pairs. By now, twin registries are increasingly popular and have proven strengths in longitudinal data collection and inclusion of biological samples to evaluate the genetic variation in susceptibility to disease.

Remarkably, despite the rapid gain in knowledge about the importance of genetic variation due to advancements in genotyping technology combined with the development of powerful linkage and genetic association studies, there is no comprehensive understanding of the twinning process. For example, there are no estimates for the heritability of twinning. A number of factors influencing the DZ twinning process are well described, and the first genetic factors for DZ twinning have been characterized [10]. However, the heritability of DZ twinning remains unknown, and the knowledge regarding the etiology of MZ twinning is even more limited.

This chapter summarizes the current body of knowledge surrounding the epidemiological, biological, and genetic aspects of human twinning. Included are explanations of the biological mechanisms of twinning, descriptions of the underlying genetic contributions to the twinning process, and details on the frequency of twinning within populations.

ZYGOSITY, CHORIONICITY, PLACENTATION OF TWINNING

Zygosity refers to the genetic makeup of the pregnancy or the number of zygotes that become fertilized leading to a multiple birth. Twins, in rarer cases, triplets, quadruplets (four), quintuplets (five) arising from a single fertilized ovum are termed monozygotic (MZ). The first identical quintuplets known to have survived their infancy were the Canadian Dionne quintuplets, all five of which survived to adulthood (Fig. 2.1). Alternatively, twins or multiples originating from two or more ova that are fertilized by separate spermatozoa are called dizygotic (DZ) [11] or trizygotic in the case of triplets. Nowadays, higher-order births of non-identical twins often are the result of assisted reproductive technologies (ART).

Figure 2.1 - The Dionne quintuplets. The Dionne quintuplets, born May 28, 1934, outside of Callander, Ontario, Canada. Despite being born 2 months premature, all five quintuplets survived to adulthood. (Source: https://commons.wikimedia.org/wiki/File:Dionne_quintuplets.jpg)

Multiple gestation pregnancies are inherently high risk to both the mother and the developing fetuses. Certain twin pregnancies, specifically those possessing a single chorion (monochorionic), exhibit an even higher risk for numerous pre- and perinatal complications. Therefore, information about the status of the developing fetal membranes is helpful for monitoring and improving the outcome of a multiple gestation pregnancy.

In early-stage human development, the membranes of the placenta begin to form around day 4. Examination of placental membranes by ultrasound imaging serves as a non-invasive method for determining zygosity/twin status [12-15]. According to the traditional models of twinning (Fig. 2.2), DZ twins have distinct placentas and fetal membranes and are therefore dichorionic diamniotic, although fused membranes are possible. Approximately two-thirds of all MZ twins share one placenta with monochorionic diamniotic membranes, while roughly one-third have completely distinct placentas and membranes (dichorionic diamniotic). Only about 1% of MZ twins have one set of membranes and one placenta, making them monochorionic monoamniotic. The latter case poses the highest risk for pre- and perinatal morbidity and mortality.

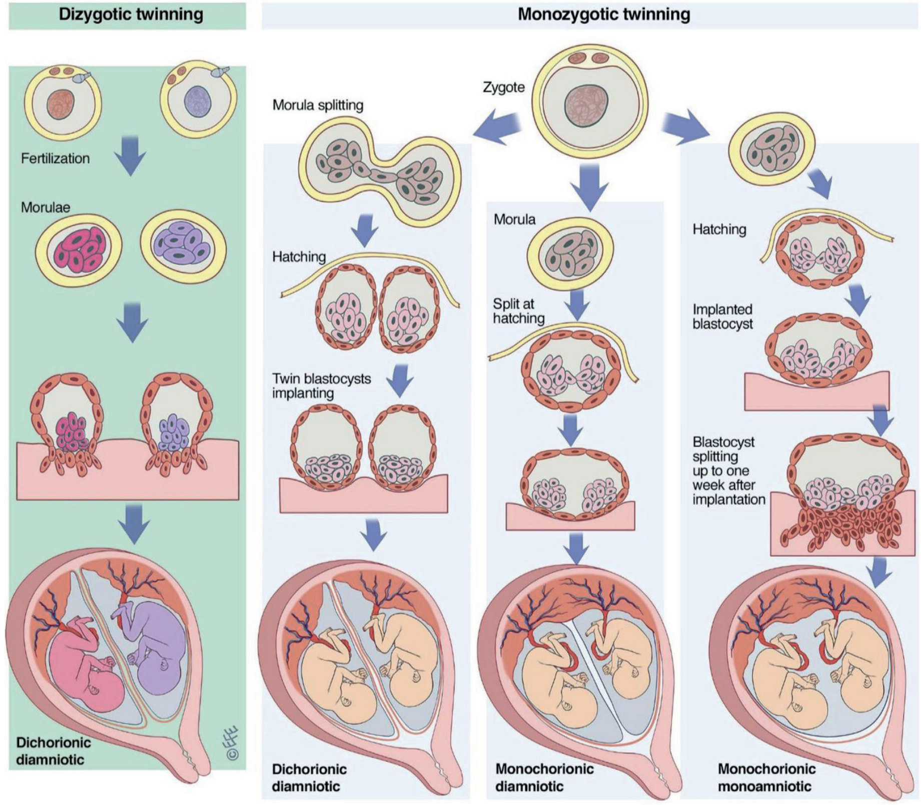


Fig. 2.2 - The traditional model of twinning. The formation of the two main types of twins according to the traditional model of twinning. DZ twins are the result of two distinct fertilization events and are dichorionic and diamniotic. MZ twins result from the post-zygotic splitting of a single embryo early in gestation with varying numbers of fetal membranes depending on the timing of embryo splitting. (This figure was obtained from McNamara et al. [16])

The placenta and membranes of a multiple gestation pregnancy represent the first means of identifying some MZ twins due to the presence of a single chorion. Although MZ twins will be of the same sex, not all same-sex twins are MZ. Therefore, for all same-sex twin pairs, DNA typing is the most useful and reliable method for determining zygosity [17]. Challenging the assumption that a single chorion indicates monozygosity, it is also possible for DZ twins to possess monochorionic diamniotic placentas. Thus, the dogma of a single chorion being synonymous with monozygosity is no longer proper due to chimeric DZ twins [18], a phenomenon in which one individual is composed of cells from two or more zygotes.

In normal embryogenesis, the chorionic membrane begins to form at about day 3. It is believed that if zygote separation takes place early, typically between day 1 and day 3, the result is MZ twins with separate placentas and membranes (dichorionic diamniotic). Alternatively, monochorionic diamniotic MZ twins result after the chorion has formed (day 3) but before the amnion has formed (typically between day 6 and day 8). Therefore, postzygotic splitting resulting in monochorionic diamniotic MZ twins typically occurs between day 3 and day 8. MZ twins with monochorionic monoamniotic membranes likely arise by splitting between day 8 and day 13. Conjoined twins are thought to arise after the beginning of the formation of the primitive streak, likely after day 14. However, the timing and mechanism(s) are not clear and empirical evidence is rare.

ZYGOSITY DETERMINATION

Physical Appearance

For research purposes, twin zygosity is often and most easily determined from questionnaires regarding the similarity of physical characteristics. For instance, twins that are equal on most physical features and frequently confused for one another are typically judged to be MZ. Alternatively, twins that differ on two or more physical characteristics and/or are not often mistaken for one another are often classified as DZ. As a whole, zygosity assessment from survey responses of physical traits corresponds rather well with zygosity determination through DNA typing [19, 20]. Apart from DNA testing, other basic rules routinely employed for zygosity determination are if the twins are opposite-sex – DZ, of discordant blood groups – DZ, or possess a single, non-fused, placenta – MZ (please note the presence of two placentas does not imply DZ).

Placentation

Placental examination of a twin birth is common to establish the type of chorion and infer zygosity. DZ twins, except for chimeric twins, have two placentas with two chorions and two amnions. Thus, most DZ twins have dichorionic diamniotic placentation. Although the placentas of DZ twins can sometimes appear fused, yet they are functionally independent and with no inter-placental communication. Placentation in MZ twins is thought to vary depending on the timing of postzygotic splitting following a single fertilization event. Dichorionic diamniotic MZ twins (~33%) are formed if the split occurs early, on days 1–3, up to the morula stage. Monochorionic diamniotic MZ twins (~66%) result if the split happens between days 3 and 8, during which blastocyst hatching starts. Monochorionic monoamniotic MZ twins (~1%) occur if the split occurs between days 8 and 13. If no split has occurred by day 13, conjoined twins form [11, 16]. Examples of varying numbers of fetal membranes are also observed in triplets and higher-order multiples, as shown in Fig. 2.3 (images adapted from Lamb et al. [21]).

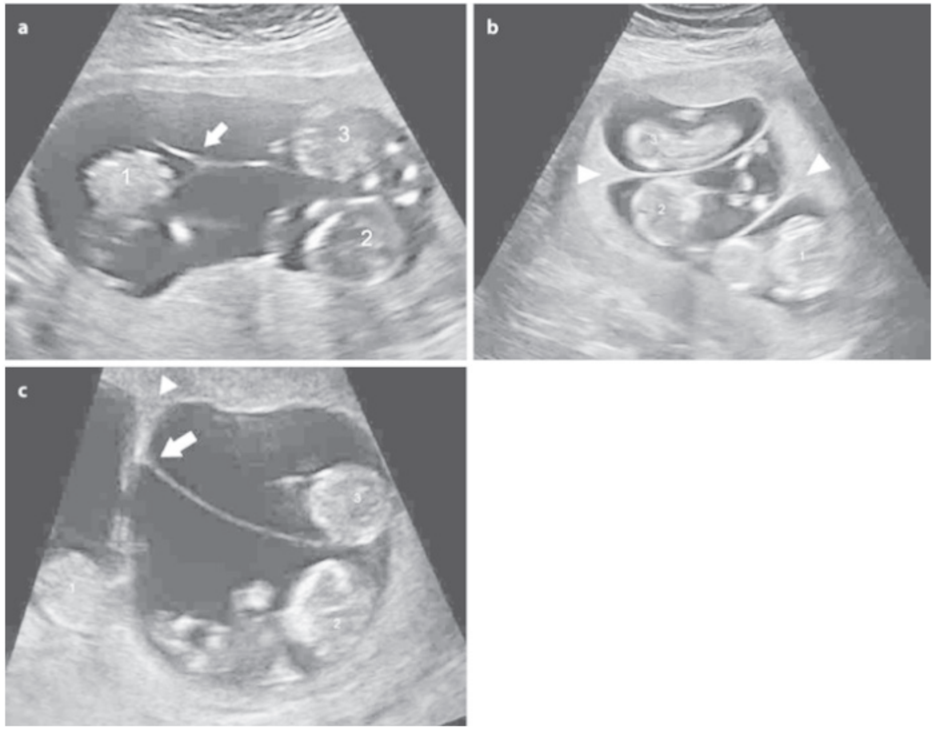


Fig. 2.3 - Ultrasound images of triplets with varying numbers of fetal membranes. a Monochorionic and therefore monozygotic triplets at 12 weeks of gestational age. The arrow indicates the meeting point of three amniotic membranes. Numbers indicate the three fetuses. b Trichorionic triplets at 12 weeks gestational age. The arrowheads indicate the separation between each fetus. These three fetuses do not share their placentas. This set of triplets can be trizygotic, dizygotic (one identical pair), or monozygotic. Numbers indicate the three fetuses. c. Dichorionic, triamniotic triplets at 13 weeks gestational age. The arrowhead indicates the separation of the chorionic membranes, which proves that this fetus does not share a placenta with fetuses 2 or 3. The arrow indicates the amniotic membranes of fetuses 2 and 3, which are a monozygotic pair. At this point in time, it is unsure if fetus 1 shares zygosity with fetuses 2 and 3. Numbers indicate the three fetuses. (This figure was adapted from Lamb et al. [21] with written permission from Cambridge University Press)

DNA Typing

The most robust and reliable way of determining zygosity is by DNA typing. The advent of cost-effective DNA genotyping has allowed for the opportunity to accurately determine zygosity through quantitative measures of allele sharing between twins when biological samples are available [22, 23]. Genetically, MZ twins will share (close to) 100% of their alleles, and on average, DZ twins will share 50% of their alleles: similar to the allele sharing pattern of siblings. The recommended minimum number of single nucleotide polymorphisms (SNPs) needed to assess zygosity is around 50 [23]; however, utilization of approximately 20,000–30,000 SNPs bolsters confidence in zygosity determination. Procedurally, after genotyping and quality control, an optimal number of SNPs are selected, and allele sharing in all pairs is determined. Sharing is reported as a proportion of markers for which a pair shares zero alleles (Z_0), one allele (Z_1), and two alleles (Z_2). From the proportions, total allele sharing (represented by $\hat{\pi}$) is calculated with the following formula: $\hat{\pi} = Z_2 + 0.5 * Z_1$. MZ pairs are identified by finding pairs with a $\hat{\pi} > 0.90$, allowing for some measurement error. DZ pairs are defined as pairs with a $\hat{\pi}$ and Z_1 between ~0.30 and ~0.70 [22].

ETIOLOGY OF TWINNING

Genetic Causes of MZ Twins

There have been several reports of families in which MZ twinning occurs more frequently than expected [24-29], although there is no compelling evidence to support an underlying genetic contribution to MZ twinning. Instances of familial MZ twinning from both maternal and paternal lineages have been documented, yet it has also been suggested that independent of the sex of the parent transmitting the gene, a single gene is responsible for MZ twinning [28, 30]. Additional evidence suggests that there is no paternal effect on familial MZ twinning [31].

More recently, a gene thought to likely play a role in MZ twinning is *PITX2*. The *PITX2* gene was found as a candidate for monozygotic twinning in a molecular screening of an "experimental twinning" model in chickens [32]. *PITX2* encodes a protein that acts as a transcription factor, regulating the expression of genes involved with the formation of the embryonic axis.

Generation of MZ Twins

The universally accepted model of MZ twinning, frequently referred to as the "fission model," rests on the hypothesis of postzygotic splitting of the conceptus within the first 2 weeks of development [16]. As exemplified by the model, the number of fetuses, chorions, and amnions results from the timing of embryo division. An alternative model of MZ twinning, sometimes called the "fusion model," challenges the traditional postzygotic splitting conjecture. The fusion model has been suggested due to criticisms of the traditional model lacking scientific evidence and the lack of specification of cleavage initiating factors [33]. The proposed alternate theory is based on two premises: (1) – MZ twinning occurs during the first cleavage division, resulting in twin zygotes, and (2) – the structure of the fetal membranes is dependent on the various modes of fusion of the fetal membranes within the zona pellucida. Despite the two theories, the embryological processes that govern MZ twinning are still largely unknown and up for debate [34].

Genetic Causes of DZ Twins

DZ twinning is a complex trait that is likely under the influence of multiple genes. In the last decade, astonishing progress in characterizing the genes responsible for DZ twinning has been made. However, a comprehensive understanding of the genetic factors underlying the human tendency to conceive DZ twins is still lacking. Bearing this in mind, there are a number of genes with known roles in ovulation, female fertility, and DZ twinning in humans [35]. For example, mutations resulting in amino acid changes of the follicle-stimulating hormone receptor (FSHR) protein have been shown to influence DZ twinning [36]. Additionally, a variant within the promoter region of the *FSHR* gene was found to segregate with DZ twinning in a large family [37]; however, other studies have not replicated the involvement of this gene [38]. Several other candidate gene studies have provided evidence suggesting the involvement of other genes in the DZ twinning process, namely, serpin family A member 1 (*SERPINA1*) commonly referred to as alpha-1-antitrypsin [39, 40], peroxisome proliferator-activated receptor gamma (*PPARG*) [41], and the fragile X "premutation" (*FRAXA*) [42, 43], although results were not replicated in future studies.

Linkage studies in family-based study designs have not provided evidence of enhanced genetic sharing among affected family members over chromosomal regions harboring the previously described candidate genes [37, 44]. However, linkage studies have indicated chromosomal regions that may possess novel candidate genes for DZ twinning [37, 44, 45]. For example, a study of 525 Australian and Dutch families of DZ twinning demonstrated the presence of new candidate DZ twinning genes on chromosomes 6, 12, and 20 [37]. The results reaffirmed the notion that DZ twinning is a complex phenotype influenced by numerous genes. Mutations in growth differentiation factor-9 (*GDF9*) appear to influence DZ twinning in humans, albeit such mutations appear to be rare. Screening in large numbers of families with a rich history of DZ twinning revealed a two-base deletion in *GDF9* in heterozygous form resulting in a loss-of-function mutation in three families [46, 47]. It was also discovered that overall genetic variation in GDF9 is more prevalent in mothers of DZ twins compared to controls [47]. Findings from non-human studies (e.g., sheep) of DZ twinning have implicated bone morphogenetic protein 15 (*BMP15*) and bone morphogenetic protein receptor 1B (*BMPR1B*) in the DZ twinning process; however, similar effects have not been found when studying *BMP15* [48] or *BMPR1B* [49] in humans. Both *GDF9* and *BMP15* are expressed in the oocyte and are essential for follicular development, and have been implicated in premature ovarian failure [50]. *BMPR1B* is expressed in multiple cell types of the ovary and is the cognate receptor for *BMP15*. Surprisingly, while heterozygous mutations (one copy present) in *GDF9* and *BMP15* increase twinning rates, homozygous mutations (two copies present) result in female infertility.

More recently, genome-wide association studies (GWAS) have made it possible to scan the entire human genome for SNPs associated with a trait of interest in humans (e.g., twinning). In 2016, the first meta-analysis of GWAS in European-ancestry populations (Netherlands, Australia, Minnesota [United States of America]) identified the first common genetic variants associated with spontaneous DZ twinning [51]. Three statistically significant SNPs were found: in the upstream region of the follicle-stimulating hormone beta subunit (*FSHB*) gene, within an intron of the mothers against decapentaplegic homolog 3 (*SMAD3*) gene, and in an intergenic region on chromosome 1. The two SNPs near *FSHB* and within *SMAD3* were replicated in an independent Icelandic cohort. The former encodes the beta subunit of FSH, while the gene product of the latter is a transcription factor involved in gonadal responsiveness to FSH. Sixty-three candidate genes of DZ twinning [52] were also tested, but only *FSHB* was associated with DZ twinning in gene-based tests. Interestingly, the replicated SNPs associated with twinning were also found to be associated with higher serum FSH levels, and with multiple aspects of female fertility, including earlier age at menarche, earlier age at first child, higher lifetime parity, earlier age at menopause, and later age at last child. Polygenic risk scores for DZ twinning were found to be significantly associated with DZ twinning in the independent Icelandic cohort, with a higher likelihood of having children, higher lifetime number of children, and an earlier age at first child. Together these findings corroborate the link between fertility and DZ twinning.

Generation of DZ Twins

Mechanisms leading to dizygotic twins operate on the selection of developing follicles within the ovary when instead of one ovum being released mid-cycle, two follicles mature, and both oocytes are released for fertilization. The subsequent fertilization of two eggs by two sperm during a pregnancy results in DZ twins. The processes of ovarian folliculogenesis and dominant follicle selection are governed by both circulating and intra-ovarian concentrations of FSH. Spontaneous DZ twinning tends to run in families and is associated with elevated concentrations of FSH in the mother [53]. FSH amounts seem to vary with geography, season, ethnic origin, and increasing parity, and are increased in tall, heavy, and older mothers with a peak at around 37 years of age [54].

Following this logic, it has been proposed that age-dependent twinning may also be due to natural selection favoring double ovulation events in response to declining fertility with increasing age [55]. It has also been documented that mothers of DZ twins have an increased number of FSH pulses during the early follicular phase, without a concurrent luteinizing hormone (LH) pulse [56].

It is well known that improved nutrition is a contributing factor for increases in multiple ovulation (i.e., twinning frequency) in other species [57, 58], yet this has not been demonstrated in humans. In fact, human twinning rates do not appear to reflect the average nutritional status as established from longitudinal studies of countries experiencing extended periods of starvation, such as the Dutch hunger winter [59]. In general, above a specific yet undetermined threshold, nutrition seems to be of minimal importance for twinning and reproduction in general.

INCIDENCE OF TWINS

In the early twentieth century, Wilhelm Weinberg postulated a method, referred to as the "Weinberg differential rule," for approximating population estimates of the number of MZ versus DZ twins [60]. Weinberg's proposal assumed that all MZ twins and half of DZ twins would be of the same sex, with the other half of DZ twins being of the opposite sex. Therefore, he suggested multiplying the number of opposite-sex twins by two, which served as an estimate for the total number of DZ twins. The excess of same-sex twins would then be the number of MZ twins. Thought of in a different way, subtracting the total number of opposite-sex twins from the number of same-sex twins would yield an estimate of the number of MZ twins. Weinberg's calculation has been widely adopted because it serves as a simple method for estimating twinning frequency in populations; however, it rests on the assumption that the frequency of same-sex twins is the same as that of opposite-sex twin pairs, which may not always be the case [61, 62].

The rate of twinning includes stillbirths (≥28 weeks) and live births and is defined as the number of twin maternities per 1000 maternities. Differences in twinning rates between geographical regions have been studied extensively. In the 1970s, Bulmer studied twinning frequencies in three distinct geographical regions: Europe/North Africa, Sub-Saharan Africa, and Asia [63]. Bulmer found that the highest rate of twinning occurred in Sub-Saharan Africa (~23 per 1000 maternities), while the lowest rate occurred in Asia (~5–6 per 1000 maternities). Twinning rates exhibit considerable temporal and spatial variation (see Fig. 2.4). Provided that the MZ twinning rate is known to be fairly constant around the world, the variation in the overall twinning rate is generally attributed to the variation in DZ twinning rates.

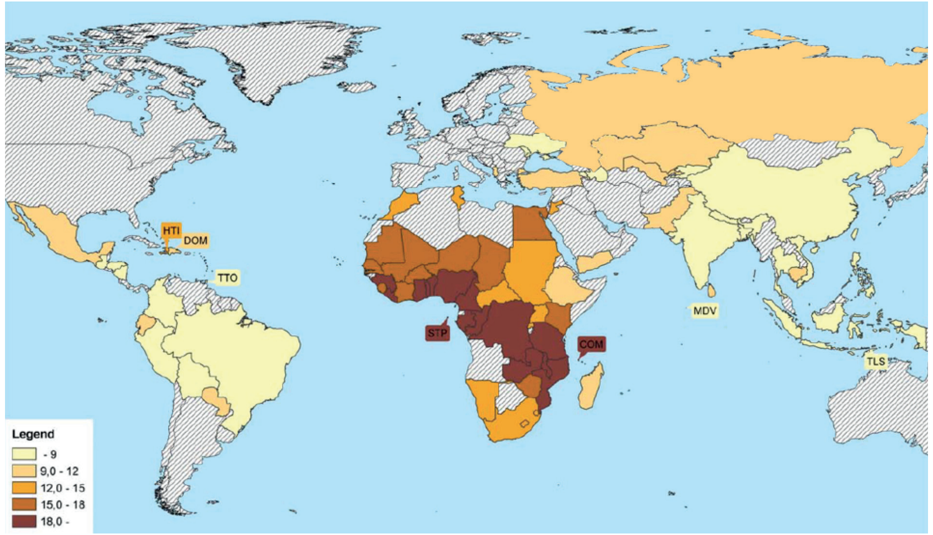


Fig. 2.4 - Rates of twinning worldwide. Heatmap showing the number of twins per 1000 births in 77 countries. Huge variation in twinning rates can be observed across the different regions of the developing world. (The figure was adapted from Smits and Monden [64] by Veronika Odintsova to include 2011 Russian twinning rates.)

Global twinning rates have predominantly been reported from Western countries, with less known about twinning rates in Eastern countries. In 2017, data received from open resources of national statistics of the Ministry of Health of the Russian Federation reported a multiple birth rate of 12.27 births per 1000 deliveries (alive or stillbirth) as seen in Table 2.1 [65]. The overall twinning rate during this time was reported to be 12.09 per 1000, and the rate of higher-order multiples occurred at a rate of 0.18 per 1000.

Table 2.1 - Multiple births in Russian Federation in 2017

		Total number			Proportion in all deliveries			Birth rate per 1000		
	All deliveries, including born outside clinic	Multiple (all)	Twins	Triplets and higher order	Multiple (all)	Twins	Triplets and higher order	Multiple (all)	Twins	Triplets and higher order
Russian Federation	1,649,782	20,239	19,938	301	1.23	1.21	0.02	12.27	12.09	0.18

Table 2.1 was created and provided by Veronika Odintsova based on Russian national statistics [65].

Incidence of MZ Twins

Worldwide and across all races, MZ twin birth rates occur at a constant rate of approximately 4 in every 1000 pregnancies [66]. The remarkable consistency in MZ twinning rates among all populations suggests that identical twinning is an occurrence that is not influenced by genetics. Unlike DZ twinning, the incidence of MZ twins is independent of maternal age, height, weight, or parity [67]. Although MZ twinning appears to be a sporadic event, instances of familial MZ twinning of varying modes of inheritance have been reported, with one report of autosomal dominant inheritance with variable penetrance [68]. The introduction of assisted reproductive technologies (ART) greatly enhances rates of DZ twinning and, to some extent, MZ twinning [69]. The increase in MZ twinning rates due to ART has been attributed to mechanical forces affecting the zona pellucida or to the effects of incubation media and late implantation in in vitro fertilization procedures [70-72].

Incidence of DZ Twins

DZ twinning is common, yet large regional differences in DZ twinning rates exist around the world. The rate of DZ twins ranges from approximately 6 per 1000 maternities in Asia to 10–20 per 1000 in Europe and the United States, to as high as 40 per 1000 in certain regions of Africa [11]. DZ twinning rates also vary substantially over time. In the United States, the observed incidence of twin births increased by a factor of 1.9 between 1971 and 2009 [73]. A considerable portion of the increase is attributable to fertility treatments, with an estimated 36% of all twins born in the United States in 2011 resulting from ART [73, 74]. DZ twinning rates peaked in the mid-2000s, following the number of ART pregnancies. In developed countries, with technological advancements and careful monitoring, DZ twinning rates due to ART have dropped substantially. In contrast to MZ twinning, spontaneous (i.e., no ART) DZ twinning is also very dependent on many maternal factors, including age, nationality, parity, height, weight, and family history [35, 54].

Triplets and Higher-Order Multiples

The pattern of variation in triplet rates across the world is remarkably similar to that observed for twin births. That is, rates of triplet births are highest in African countries, intermediate in European populations, and lowest in Asian countries. More generally, the rate of triplets is in accordance with Hellin's law [75], which states that there is on average one twin maternity per N singleton maternities and that there is one X-tuplet maternity per $N^{(X-1)}$ [76, 77]. Thus, if the number of twin maternities is one in N singleton maternities, then the number of triplets is one in $N^{(3-1)}$. It follows that quadruplets (see Fig. 2.5) would occur at a rate of $N^{(4-1)}$. According to this logic, if there is one twin in every 80.05 births, this predicts that triplets occur at a rate of about 1 in every 6408 births, which is only 4% higher than the incidence actually observed [64].

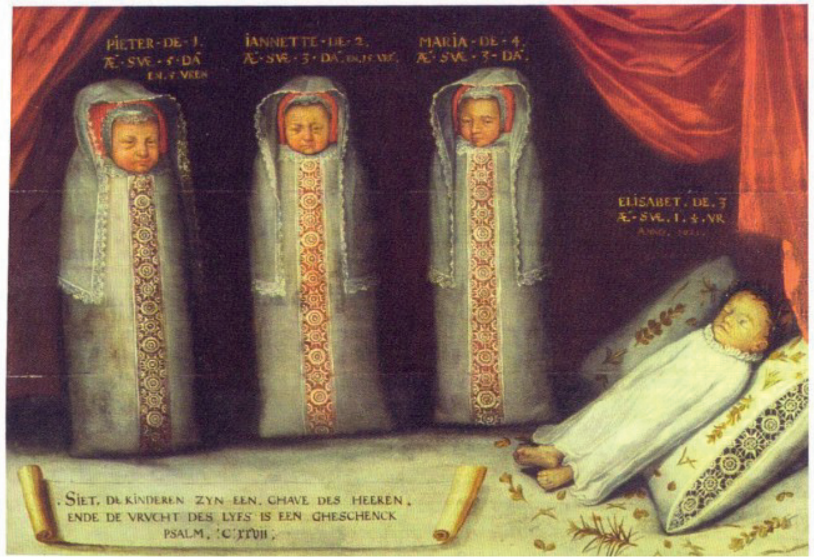


Fig. 2.5 - Painting of Dutch quadruplets. "Vierling Costerus" – Painting of quadruplets born on June 9, 1621, in Dordrecht, the Netherlands. Remarkably, the birth of the first (Pieter) and the last (Maria) was separated by 53 hours, indicating an extremely difficult birthing process. Prior to the quad birth of one boy and three girls as pictured here, the parents mentioned the birth of twins 7 years before. Sadly, due to the high infant mortality rate in the seventeenth century, about 40–50% of children did not reach the age of 18, and the chance of survival was even smaller for multiple births. In the case of the quadruplets pictured here, one died an hour and a half after birth (Elisabet), and two others deceased within the first year. (This figure was obtained by written permission from the Dordrecht Museum. The Noordbrabants Museum in 's-Hertogenbosch received this painting as a gift in 1925 and presented it on loan to the Dordrechts Museum in 1986. Painter and client are unknown.)

FACTORS AFFECTING TWINNING

Many of the risk factors for DZ twinning are rather well established and include assisted reproductive technologies (ART), higher maternal age, parity, body composition, and smoking [11, 35, 78]. For MZ twinning, there is little to no agreement regarding the involved risk factors and/or causes [79]. Given that the two types of twinning are biologically distinct phenomena, it is not surprising that many of the factors involved in DZ twinning are not, or to a lesser extent, found in MZ twinning. Below we describe established risk factors involved in multiple pregnancy and multiple birth.

Assisted Reproductive Technologies (ART)

Assisted reproductive technologies, especially in vitro fertilization (IVF) and ovulation induction (OI), are well-established risk factors for DZ and, to some extent, MZ twinning [79]. Ovulation-inducing agents such as clomiphene citrate, human pituitary gonadotropins, and human menopausal gonadotropins are known to increase ovulation rate and hence the probability of multiple pregnancy [80]. In the case of IVF, increased multiple pregnancy and multiple birth are due to the transfer of multiple embryos [11, 81]. Although less well understood, there is also a slight increase in MZ twinning after IVF and other ART, with estimates ranging from a two- to 12-fold increase in MZ twinning rate after ART procedures [79] and a two- to fivefold increase in MZ twinning following IVF [82, 83]. Notably, OI agents are often used in concert with IVF; thus, the increased chance of multiple pregnancy following IVF cannot be merely attributed to multiple embryo transfer [35].

Other Risk Factors

Higher maternal age is another well-established risk factor for DZ twinning [35, 63]; however, conflicting reports have been made for MZ twinning [63, 84, 85]. Paradoxically, while fertility decreases with age, the spontaneous twinning rate increases. Both polyovulation and embryonic survival rates increase with maternal age, suggesting that increased ovulation rate and decreased spontaneous abortion of potentially unhealthy offspring could act together as an insurance system to produce one last round of reproduction, with twinning being merely a by-product [55, 86]. Additionally, increased parity (the number of maternities prior to twin pregnancy) is associated with a higher risk of DZ, but not MZ twinning [11, 87], independent of maternal age [35, 88]. Moreover, maternal body composition has been reported rather consistently in relation to DZ twinning, with both obesity and tall stature increasing the risk of DZ twinning [53, 85, 89]. Furthermore, an unexpected relation between maternal smoking and a higher probability of DZ twinning is sometimes observed, although the mechanism behind this observation remains unclear [78, 90, 91].

ENDOCRINOLOGY OF DZ TWINNING

Mothers of spontaneous DZ twins have a predisposition to multiple ovulation events due to interference with the selection of a single dominant follicle. Multiple follicle growth and subsequent multiple ovulation events have been observed in mothers of hereditary dizygotic twins [92, 93]. Follicular recruitment, selection, and dominance is controlled by a complex regulatory network within the hypothalamic-pituitary-ovarian axis (Fig. 2.6). The two main pituitary-derived hormones essential for reproductive function are FSH and LH and are secreted in response to the pulsatile secretion of gonadotropin-releasing hormone. FSH is the main hormone controlling follicular growth, and its secretion is controlled by the main secretory products of the large dominant follicle(s), namely, estradiol and inhibin. Circulating concentrations of FSH, other intra-ovarian factors (e.g., GDF9 and BMP15), and their cognate receptors physiologically regulate ovarian folliculogenesis and ovulation quota. Transcriptional regulators of *FSH*, such as SMAD3, also regulate gonadal responsiveness to FSH. Ongoing development of a single follicle takes place when a certain threshold level of plasma FSH is marginally exceeded [94, 95]. When FSH levels are much higher than the threshold level or exceed the threshold for an extended duration, multiple follicle growth can result [96]. In accordance with the endocrine model of dizygotic twinning [97], high levels of pituitary gonadotropins (i.e., FSH) are responsible for increased multiple ovulation in mothers of DZ twins. A number of studies, but not all [92], have shown increased levels of plasma gonadotropins in mothers of DZ twins [56, 98-100].

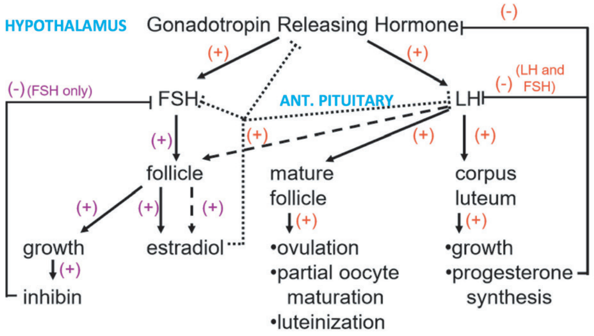


Fig. 2.6 - Hormonal feedback and endocrine regulation of female reproductive physiology. + signs denote positive feedback, whereas blunted arrows indicate negative feedback. (This figure was obtained by written permission from the lecture materials of Dr. Kathleen Eyster (University of South Dakota, United States))

NON-HUMAN TWINNING

Studies of non-human species have revealed several genes that contribute to DZ twinning. For example, in sheep, which are typically uniparous, certain breeds have higher incidences of multiple births [101]. Several genes, namely, *GDF9*, *BMP15*, and *BMPR1B*, have been confirmed to influence twinning rates in sheep by increasing follicle development and oocyte maturation [49, 102-104]. Major genes that increase ovulation rate and litter size in sheep and humans have been shown to have implications in other species. Evidence exists for genetic effects on twinning in cattle [105], the marmoset monkey [106], and hormone-induced ovulation rate in mice [107].

Genes specific to MZ twinning remain elusive. Animal studies (originally in rabbit and roe deer) have suggested that MZ twinning results from disturbances to developmental thresholds and that delayed fertilization/implantation play a role [63]. These hypotheses have been further tested in nine-banded armadillos (*Dasypus novemcinctus*), which bear obligate MZ quadruplets each and every time they breed [108-111].

SEX RATIO

The sex ratio is defined as the ratio of males to total births. For DZ twins and singletons, the ratio is 0.514, meaning a slight excess of males [63, 112]. For spontaneous MZ twins, triplets, and quadruplets, the sex ratio is lower (0.496) due to a slight excess of females [62, 112]. The value is even lower for monoamniotic twins, including conjoined twins, with a sex ratio of 0.2 [112]. There does not seem to be an excess of males in aborted twins. Dichorionic MZ twins exhibit the smallest increase in females, whereas monochorionic diamniotic MZ twins show the largest increase; thus, the rise may be due to later twinning events. Monochorionic monoamniotic twins and conjoined twins show an even greater increase in the number of females.

DISCUSSION AND CONCLUSION

Twinning is a common and multifactorial phenomenon, and elements of the twinning process remain poorly uncharacterized. Improved understanding of the underlying biological and genetic aspects of MZ and DZ twins and of the twinning process as a whole has been enhanced by the development of molecular and cytogenetic techniques. The influences on DZ twinning are well studied, including contributions of numerous maternal (age, height, weight, parity, family history), environmental (ART), and associated genetic factors (most notably common variants near *FSHB* and within *SMAD3*). However, despite the robust associations with DZ twinning, the ability to predict DZ twinning events remains imprecise due to the myriad of genetic and non-genetic contributions. Likewise, comprehensive models of MZ twinning lack compelling evidence and are challenged by instances of atypical twinning. MZ twinning is likely influenced by some equivocal combination of non-genetic and genetic factors that likely result in delayed fertilization, embryo development, implantation, or some form of mechanical disruption of the early embryo. Further elucidation of the mechanisms by which twinning processes occur will have significant merit for predicting, managing, and improving the outcomes of multiple gestation pregnancies.

Review Questions

1. What are the etiological similarities and differences of monozygotic and dizygotic twins?
2. What is the role played by assisted reproductive technologies in the incidence of twinning over time?

Multiple-Choice Questions

1. What is the most conclusive method for determining the zygosity status of twins?

(a) Examination of placentation
(b) Sex status (same or opposite sex)
(c) Evaluation of survey responses
(d) DNA testing
(e) Assessment of outward physical appearance

Answer: (d). Although the other methods are convenient and relatively robust approaches for determining the zygosity status of twins, only DNA testing provides a quantitative and accurate assessment of allele sharing for all twins, including same-sex twin pairs.

2. What is another term for dizygotic twins?

(a) Identical twins
(b) Fraternal twins
(c) Typical twins
(d) Atypical twins

Answer: (b). Dizygotic, or non-identical twins, develop from separate ova and are therefore genetically distinct. Thus, because their genetic relatedness is the same as other sibling pairs, dizygotic twins are commonly referred to as fraternal twins.

3. The rarest form of monozygotic twins (with the exception of conjoined twins) exhibit which of the following fetal membrane states?

(a) Dichorionic diamniotic
(b) Dichorionic monoamniotic
(c) Monochorionic diamniotic
(d) Monochorionic monoamniotic

Answer: (d). Monochorionic monoamniotic monozygotic twins represent about 1% of all twin pregnancies. Approximately 66% of monozygotic twins are monochorionic diamniotic, whereas 33% are dichorionic diamniotic.

4. Genome-wide association studies have identified common genetic variants associated with spontaneous dizygotic twinning and female fertility in which genes?

(a) *FSHB* and *SMAD3*
(b) *GDF9* and *BMP15*
(c) *BMP4* and *WFIKKN1*
(d) *FSHR* and *BMPR1B*

Answer: (a). Single nucleotide polymorphisms near *FSHB* and within *SMAD3* are significantly associated with a higher rate of spontaneous dizygotic twinning and several other aspects of female fertility (e.g., earlier age at menarche, earlier age at first child, and higher lifetime parity).

5. In order of highest to lowest incidence, which of the following captures the large regional differences observed for dizygotic twinning?

(a) Asia, Africa, Europe
(b) Africa, Europe, Asia
(c) Europe, Asia, Africa
(d) Europe, Africa, Asia

Answer: (b). Whereas monozygotic twinning rates are relatively constant worldwide (~3 per 1000 births), large regional differences exist in dizygotic twinning rates with the highest incidence in African populations (~40 per 1000 births), followed by European populations (~10–20 per 1000 births), and the lowest incidence in Asian populations (~6 per 1000 births).

6. Which of the following is not a known non-genetic risk factor for spontaneous dizygotic twinning?

(a) Parity
(b) Maternal age
(c) Nutritional status
(d) Smoking status
(e) Body mass index
(f) Height

Answer: (c). Spontaneous dizygotic twinning is associated with parity, as well as increased maternal age, increased body mass index, increased height, and smoking status prior to pregnancy. Nutritional status is not known to be a direct contributor to dizygotic twinning, as longitudinal studies in countries that experienced periods of starvation demonstrated consistent rates of twinning (e.g., Dutch hunger winter [59]).

ACKNOWLEDGMENTS

We are grateful to Veronika Odintsova for the generous contribution of the multiple birth data from the Russian Federation, which was obtained from the Federal Research Institute for Health Organization and Informatics, Ministry of Health of Russian Federation, Moscow, Russia.

REFERENCES

1. Duncan, J.M., *Fertility, Sterility and Allied Topics*. Second ed. 1871, New York: William Wood and Co. 498.
2. Galton, F., *The History of Twins, as a Criterion of the Relative Powers of Nature and Nurture*, in *Fraser's Magazine*. 1875. p. 566-576.
3. Fisher, R.A., *The Genesis of Twins.* Genetics, 1919. **4**(5): p. 489-99.
4. Skytthe, A., et al., *The Danish Twin Registry.* Scand J Public Health, 2011. **39**(7 Suppl): p. 75-8.
5. Levit, S.G., *Twin investigations in the U.S.S.R.* Character and Personality, 1935. **3**: p. 188-193.
6. Sukhanov, S., *Opsihozah u bliznetsov [On psychosis in twins].* Klinicheskij jurnal, 1900. **4**: p. 341-352.
7. Yudin, T.I., *O shodstve psihoza u bratjev i sester. [On similarity of psychosis in brothers and sisters].* Sovremennaya psihiatria, 1907. **10**: p. 337-342.
8. Poll, H., *Über Zwillingsforschung als Hilfsmittel menschlicher Erbkunde.* Zeitschrift für Ethnologie, 1914. **46**: p. 87-105.
9. Siemens, H.W., *Die Zwillingspathologie.* Mol. Gen. Genet., 1924. **35**: p. 311-312.
10. Mbarek, H., et al., *Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility.* Am J Hum Genet, 2016. **98**(5): p. 898-908.
11. Hall, J.G., *Twinning.* Lancet, 2003. **362**(9385): p. 735-43.
12. Benirschke, K., *The placenta in twin gestation.* Clin Obstet Gynecol, 1990. **33**(1): p. 18-31.
13. Benirschke, K. and C.K. Kim, *Multiple pregnancy. 1.* N Engl J Med, 1973. **288**(24): p. 1276-84.
14. Benirschke, K. and C.K. Kim, *Multiple pregnancy. 2.* N Engl J Med, 1973. **288**(25): p. 1329-36.
15. Benirschke, K. and E. Masliah, *The placenta in multiple pregnancy: outstanding issues.* Reprod Fertil Dev, 2001. **13**(7-8): p. 615-22.
16. McNamara, H.C., et al., *A review of the mechanisms and evidence for typical and atypical twinning.* Am J Obstet Gynecol, 2016. **214**(2): p. 172-191.
17. Cutler, T.L., et al., *Why Accurate Knowledge of Zygosity is Important to Twins.* Twin Res Hum Genet, 2015. **18**(3): p. 298-305.
18. Umstad, M.P., et al., *Chimaeric twins: why monochorionicity does not guarantee monozygosity.* Aust N Z J Obstet Gynaecol, 2012. **52**(3): p. 305-7.
19. Wray, N.R., et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nat Genet, 2018. **50**(5): p. 668-681.
20. Forget-Dubois, N., et al., *Diagnosing zygosity in infant twins: physical similarity, genotyping, and chorionicity.* Twin Res, 2003. **6**(6): p. 479-85.
21. Lamb, D.J., et al., *Effects of chorionicity and zygosity on triplet birth weight.* Twin Res Hum Genet, 2012. **15**(2): p. 149-57.
22. Odintsova, V.V., et al., *Establishing a Twin Register: An Invaluable Resource for (Behavior) Genetic, Epidemiological, Biomarker, and 'Omics' Studies.* Twin Res Hum Genet, 2018. **21**(3): p. 239-252.
23. Hannelius, U., et al., *Large-scale zygosity testing using single nucleotide polymorphisms.* Twin Res Hum Genet, 2007. **10**(4): p. 604-25.
24. Cyranoski, D., *Developmental biology: Two by two.* Nature, 2009. **458**(7240): p. 826-9.
25. Barban, N., et al., *Genome-wide analysis identifies 12 loci influencing human reproductive behavior.* Nat Genet, 2016.
26. Machin, G., *Familial monozygotic twinning: a report of seven pedigrees.* Am J Med Genet C Semin Med Genet, 2009. **151C**(2): p. 152-4.
27. Shapiro, L.R., L. Zemek, and M.J. Shulman, *Familial monozygotic twinning: an autosomal dominant form of monozygotic twinning with variable penetrance.* Prog Clin Biol Res, 1978. **24 Pt B**: p. 61-3.
28. Shapiro, L.R., L. Zemek, and M.J. Shulman, *Genetic etiology for monozygotic twinning.* Birth Defects Orig Artic Ser, 1978. **14**(6A): p. 219-22.
29. St Clair, J.B. and M.D. Golubovsky, *Paternally derived twinning: a two century examination of records of one Scottish name.* Twin Res, 2002. **5**(4): p. 294-307.
30. Michels, V.V. and V.M. Riccardi, *Twin recurrence and amniocentesis: male and MZ heritability factors.* Birth Defects Orig Artic Ser, 1978. **14**(6A): p. 201-11.
31. Lichtenstein, P., B. Kallen, and M. Koster, *No paternal effect on monozygotic twinning in the Swedish Twin Registry.* Twin Res, 1998. **1**(4): p. 212-5.
32. Torlopp, A., et al., *The transcription factor Pitx2 positions the embryonic axis and regulates twinning.* Elife, 2014. **3**: p. e03743.
33. Herranz, G., *The timing of monozygotic twinning: a criticism of the common model.* Zygote, 2013. **23**(1): p. 27-40.
34. Denker, H.W., *Comment on G. Herranz: The timing of monozygotic twinning: a criticism of the common model. Zygote (2013).* Zygote, 2013. **23**(2): p. 312-4.
35. Hoekstra, C., et al., *Dizygotic twinning.* Hum Reprod Update, 2008. **14**(1): p. 37-47.
36. Al-Hendy, A., et al., *Association between mutations of the follicle-stimulating-hormone receptor and repeated twinning.* Lancet, 2000. **356**(9233): p. 914.
37. Painter, J.N., et al., *A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins.* Hum Reprod, 2010. **25**(6): p. 1569-80.
38. Montgomery, G.W., et al., *Mutations in the follicle-stimulating hormone receptor and familial dizygotic twinning.* Lancet, 2001. **357**(9258): p. 773-4.
39. Boomsma, D.I., et al., *Protease inhibitor (Pi) locus, fertility and twinning.* Hum Genet, 1992. **89**(3): p. 329-32.
40. Lieberman, J., N.O. Borhani, and M. Feinleib, *Twinning as a heterozygous advantage for alpha1-antitrypsin deficiency.* Prog Clin Biol Res, 1978. **24 Pt B**: p. 45-54.
41. Duffy, D., et al., *IBD sharing around the PPARG locus is not increased in dizygotic twins or their mothers.* Nat Genet, 2001. **28**(4): p. 315.
42. Fryns, J.P., *The female and the fragile X. A study of 144 obligate female carriers.* Am J Med Genet, 1986. **23**(1-2): p. 157-69.
43. Kenneson, A. and S.T. Warren, *The female and the fragile X reviewed.* Semin Reprod Med, 2001. **19**(2): p. 159-65.
44. Derom, C., et al., *Genome-wide linkage scan for spontaneous DZ twinning.* Eur J Hum Genet, 2006. **14**(1): p. 117-22.

45. Busjahn, A., et al., *A region on chromosome 3 is linked to dizygotic twinning.* Nat Genet, 2000. **26**(4): p. 398-9.

46. Montgomery, G.W., et al., *A deletion mutation in GDF9 in sisters with spontaneous DZ twins.* Twin Res, 2004. **7**(6): p. 548-55.

47. Palmer, J.S., et al., *Novel variants in growth differentiation factor 9 in mothers of dizygotic twins.* J Clin Endocrinol Metab, 2006. **91**(11): p. 4713-6.

48. Zhao, Z.Z., et al., *Variation in bone morphogenetic protein 15 is not associated with spontaneous human dizygotic twinning.* Hum Reprod, 2008. **23**(10): p. 2372-9.

49. Luong, H.T., et al., *Variation in BMPR1B, TGFRB1 and BMPR2 and control of dizygotic twinning.* Twin Res Hum Genet, 2011. **14**(5): p. 408-16.

50. Dixit, H., et al., *Genes governing premature ovarian failure.* Reprod Biomed Online, 2010. **20**(6): p. 724-40.

51. Mbarek, H., C.V. Dolan, and D.I. Boomsma, *Two SNPs Associated With Spontaneous Dizygotic Twinning: Effect Sizes and How We Communicate Them.* Twin Res Hum Genet, 2016. **19**(5): p. 418-21.

52. Harris, R.A., et al., *Evolutionary genetics and implications of small size and twinning in callitrichine primates.* Proc Natl Acad Sci U S A, 2014. **111**(4): p. 1467-72.

53. Nylander, P.P., *The factors that influence twinning rates.* Acta Genet Med Gemellol (Roma), 1981. **30**(3): p. 189-202.

54. Campbell, D.M., A.J. Campbell, and I. MacGillivray, *Maternal characteristics of women having twin pregnancies.* J Biosoc Sci, 1974. **6**(4): p. 463-70.

55. Hazel, W.N., et al., *An age-dependent ovulatory strategy explains the evolution of dizygotic twinning in humans.* Nat Ecol Evol, 2020.

56. Lambalk, C.B., et al., *Increased levels and pulsatility of follicle-stimulating hormone in mothers of hereditary dizygotic twins.* J Clin Endocrinol Metab, 1998. **83**(2): p. 481-6.

57. Montgomery, G.W. and H. Hawker, *Seasonal reproduction in ewes selected on seasonal changes in wool growth.* J Reprod Fertil, 1987. **79**(1): p. 207-13.

58. Hunter, M.G., et al., *Endocrine and paracrine control of follicular development and ovulation rate in farm species.* Anim Reprod Sci, 2004. **82-83**: p. 461-77.

59. Eriksson, A.W., et al., *Twinning rate in Scandinavia, Germany and The Netherlands during years of privation.* Acta Genet Med Gemellol (Roma), 1988. **37**(3-4): p. 277-97.

60. Weinberg, W., *Beiträge zur physiologie und der pathologie der mehrlingsgeburten beim menschen.* Arch Physiol, 1902. **88**: p. 346-30.

61. Cameron, A.H., *The Birmingham twin survey.* Proc R Soc Med, 1968. **61**(3): p. 229-34.

62. James, W.H., *Excess of like sexed pairs of dizygotic twins.* Nature, 1971. **232**(5308): p. 277-8.

63. Bulmer, M.G., *The Biology of Twinning in Man*. 1970: Clarendon. 206.

64. Smits, J. and C. Monden, *Twinning across the Developing World.* PLoS One, 2011. **6**(9): p. e25239.

65. Polikarpov A.V., A.G.A., Golubev N.A., Turina E.M., Ogryzko E.V., Shelepova, E.A., *Key Indicators of the Health of the Mother and Child, the Activities of the Service for Childhood and Obstetrics in the Russian Federation.* Moscow: The Ministry of Health of the Russian Federation, 2018: p. 164.

66. Tong, S., D. Caddy, and R.V. Short, *Use of dizygotic to monozygotic twinning ratio as a measure of fertility.* Lancet, 1997. **349**(9055): p. 843-5.

67. Bressers, W.M., et al., *Increasing trend in the monozygotic twinning rate.* Acta Genet Med Gemellol (Roma), 1987. **36**(3): p. 397-408.

68. Harvey, M.A., R.M. Huntley, and D.W. Smith, *Familial monozygotic twinning.* J Pediatr, 1977. **90**(2): p. 246-7.

69. Alikani, M., et al., *Monozygotic twinning following assisted conception: an analysis of 81 consecutive cases.* Hum Reprod, 2003. **18**(9): p. 1937-43.

70. Abusheikha, N., et al., *Monozygotic twinning and IVF/ICSI treatment: a report of 11 cases and review of literature.* Hum Reprod Update, 2000. **6**(4): p. 396-403.

71. Milki, A.A., et al., *Incidence of monozygotic twinning with blastocyst transfer compared to cleavage-stage transfer.* Fertil Steril, 2003. **79**(3): p. 503-6.

72. Steinman, G., *Mechanisms of twinning. VI. Genetics and the etiology of monozygotic twinning in in vitro fertilization.* J Reprod Med, 2003. **48**(8): p. 583-90.

73. Kulkarni, A.D., et al., *Fertility treatments and multiple births in the United States.* N Engl J Med, 2013. **369**(23): p. 2218-25.

74. Fauser, B.C., P. Devroey, and N.S. Macklon, *Multiple birth resulting from ovarian stimulation for subfertility treatment.* Lancet, 2005. **365**(9473): p. 1807-16.

75. Hellin, D., *Die Ursache der Multiparitat der uniparen Tiere uberhaupt und der Zwillingsschwangerschaft beim Menschen insbesondere (The causes ofmultiple maternities among uniparous animals and in man).* Seitz & Schauer, 1895.

76. Fellman, J. and A.W. Eriksson, *On the history of Hellin's law.* Twin Res Hum Genet, 2009. **12**(2): p. 183-90.

77. Fellman, J. and A.W. Eriksson, *Statistical analyses of Hellin's law.* Twin Res Hum Genet, 2009. **12**(2): p. 191-200.

78. Hoekstra, C., et al., *Body composition, smoking, and spontaneous dizygotic twinning.* Fertil Steril, 2010. **93**(3): p. 885-93.

79. Aston, K.I., C.M. Peterson, and D.T. Carrell, *Monozygotic twinning associated with assisted reproductive technologies: a review.* Reproduction, 2008. **136**(4): p. 377-86.

80. Lamont, J.A., *Twin pregnancies following induction of ovulation: a literature review.* Acta Genet Med Gemellol (Roma), 1982. **31**(3-4): p. 247-53.

81. Ananth, C.V. and S.P. Chauhan, *Epidemiology of twinning in developed countries.* Semin Perinatol, 2012. **36**(3): p. 156-61.

82. Sills, E.S., M.J. Tucker, and G.D. Palermo, *Assisted reproductive technologies and monozygous twins: implications for future study and clinical practice.* Twin Res, 2000. **3**(4): p. 217-23.

83. Edwards, R.G., L. Mettler, and D.E. Walters, *Identical twins and in vitro fertilization.* J In Vitro Fert Embryo Transf, 1986. **3**(2): p. 114-7.

84. Steinman, G., *Mechanisms of twinning. II. Laterality and intercellular bonding in monozygotic twinning.* J Reprod Med, 2001. **46**(5): p. 473-9.

85. Bortolus, R., et al., *The epidemiology of multiple births.* Hum Reprod Update, 1999. **5**(2): p. 179-87.

86. Varella, M., Fernandes, E., Arantes, J., Acquaviva, T., Lucci, T., Hsu, R., David, V., Bussah, V., Valentova, J., Segal, N., Otta, E., *Twinning as an Evolved Age-dependent Physiological Mechanism: Evidence from Large Brazilian Samples*, in *Multiple Pregnancy (New Challenges)*. 2018, United Kingdom: IntechOpen.

87. Hankins, G.V. and G.R. Saade, *Factors influencing twins and zygosity.* Paediatr Perinat Epidemiol, 2005. **19 Suppl 1**: p. 8-9.

88. Tong, S. and R.V. Short, *Dizygotic twinning as a measure of human fertility.* Hum Reprod, 1998. **13**(1): p. 95-8.
89. Basso, O., et al., *Risk of twinning as a function of maternal height and body mass index.* JAMA, 2004. **291**(13): p. 1564-6.
90. Olsen, J., B. Bonnelykke, and J. Nielsen, *Tobacco smoking and twinning.* Acta Med Scand, 1988. **224**(5): p. 491-4.
91. Parazzini, F., et al., *Coffee and alcohol intake, smoking and risk of multiple pregnancy.* Hum Reprod, 1996. **11**(10): p. 2306-9.
92. Gilfillan, C.P., et al., *The control of ovulation in mothers of dizygotic twins.* J Clin Endocrinol Metab, 1996. **81**(4): p. 1557-62.
93. Martin, N.G., et al., *Excessive follicular recruitment and growth in mothers of spontaneous dizygotic twins.* Acta Genet Med Gemellol (Roma), 1991. **40**(3-4): p. 291-301.
94. Schoemaker, J., et al., *The FSH threshold concept in clinical ovulation induction.* Baillieres Clin Obstet Gynaecol, 1993. **7**(2): p. 297-308.
95. Brown, J.B., *Pituitary control of ovarian function--concepts derived from gonadotrophin therapy.* Aust N Z J Obstet Gynaecol, 1978. **18**(1): p. 46-54.
96. Baird, D.T., *A model for follicular selection and ovulation: lessons from superovulation.* J Steroid Biochem, 1987. **27**(1-3): p. 15-23.
97. Milham, S., Jr., *Pituitary Gonadotrophin and Dizygotic Twinning.* Lancet, 1964. **2**(7359): p. 566.
98. Martin, N.G., et al., *Pituitary-ovarian function in mothers who have had two sets of dizygotic twins.* Fertil Steril, 1984. **41**(6): p. 878-80.
99. Martin, N.G., et al., *Elevation of follicular phase inhibin and luteinizing hormone levels in mothers of dizygotic twins suggests nonovarian control of human multiple ovulation.* Fertil Steril, 1991. **56**(3): p. 469-74.
100. Nylander, P.P., *Serum levels of gonadotrophins in relation to multiple pregnancy in Nigeria.* J Obstet Gynaecol Br Commonw, 1973. **80**(7): p. 651-3.
101. Montgomery, G.W., K.P. McNatty, and G.H. Davis, *Physiology and molecular genetics of mutations that increase ovulation rate in sheep.* Endocr Rev, 1992. **13**(2): p. 309-28.
102. Demars, J., et al., *Genome-wide association studies identify two novel BMP15 mutations responsible for an atypical hyperprolificacy phenotype in sheep.* PLoS Genet, 2013. **9**(4): p. e1003482.
103. Reader, K.L., et al., *Booroola BMPR1B mutation alters early follicular development and oocyte ultrastructure in sheep.* Reprod Fertil Dev, 2012. **24**(2): p. 353-61.
104. Vage, D.I., et al., *A missense mutation in growth differentiation factor 9 (GDF9) is strongly associated with litter size in sheep.* BMC Genet, 2013. **14**: p. 1.
105. Komisarek, J. and Z. Dorynek, *Genetic aspects of twinning in cattle.* J Appl Genet, 2002. **43**(1): p. 55-68.
106. Marmoset Genome, S. and C. Analysis, *The common marmoset genome provides insight into primate biology and evolution.* Nat Genet, 2014. **46**(8): p. 850-7.
107. Spearow, J.L., *Major genes control hormone-induced ovulation rate in mice.* J Reprod Fertil, 1988. **82**(2): p. 787-97.
108. Enders, A.C., *Implantation in the nine-banded armadillo: how does a single blastocyst form four embryos?* Placenta, 2002. **23**(1): p. 71-85.

109. Prodohl, P.A., et al., *Molecular documentation of polyembryony and the micro-spatial dispersion of clonal sibships in the nine-banded armadillo, Dasypus novemcinctus.* Proc Biol Sci, 1996. **263**(1377): p. 1643-9.
110. Storrs, E.E. and R.J. Williams, *A study of monozygous quadruplet armadillos in relation to mammalian inheritance.* Proc Natl Acad Sci U S A, 1968. **60**(3): p. 910-4.
111. Blickstein, I. and L.G. Keith, *On the possible cause of monozygotic twinning: lessons from the 9-banded armadillo and from assisted reproduction.* Twin Res Hum Genet, 2007. **10**(2): p. 394-9.
112. James, W.H., *Sex ratio and placentation in twins.* Ann Hum Biol, 1980. **7**(3): p. 273-6.

3

PEDIGREE BASED ANALYSIS OF HUMAN DIZYGOTIC TWINNING USING WHOLE GENOME SEQUENCE DATA

This chapter summarizes an ongoing project

ABSTRACT

Spontaneous human dizygotic (DZ) twinning runs in families and is known to be influenced by numerous genetic and non-genetic factors, though the physiological pathways and complete genetic origin are unknown. Genetic data from large trait-rich pedigrees may enhance the ability to identify novel variants associated with DZ twinning. In this manner, we analyzed whole-genome genotype and sequence data from selected members of a large multigenerational pedigree with a rich history of DZ twinning to identify rare/functional variants underlying the trait. Non-parametric linkage analysis was performed to define genomic regions co-segregating with being a mother of DZ twins, but no strong linkage peaks were observed. Haplotypes were estimated and combined with genetic variants from whole-genome sequence data of selected mothers of DZ twins revealing large shared genomic regions on chromosomes 1, 3, 6, 11. We hypothesize that these areas are regions of interest containing rare variants with substantive effects on DZ twinning. Whether the variants are pedigree-specific or characteristic of a larger cohort of population-matched mothers of DZ twins will necessitate screening and further examination. In addition to contributions of common variants associated with DZ twinning, rare variant identification has the potential to elucidate novel genetic biomarkers indexing fertility and the prediction of DZ twinning.

Keywords: dizygotic twinning, whole-genome sequencing, genotyping, pedigree analysis

INTRODUCTION

Human spontaneous dizygotic (DZ) twinning occurs when two or more oocytes are released and fertilized during a single pregnancy. DZ twinning is considered a complex trait influenced by environmental and genetic factors. DZ twinning is common, affecting approximately 1-4% of women worldwide, and tends to run in families [1]. In addition to family history, increased parity and gravidity also increase the risk of spontaneous DZ twinning [2, 3]. Mothers of DZ twins (MoDZT) are taller, have increased BMI, are often overweight, and smoke more frequently before the twin pregnancy [4]. Rates of DZ twinning vary considerably with geographic location and time. Regionally, large prevalence differences exist, with the lowest and highest rates reported in Asia (~5-6 per 1000 maternities) and Sub-Saharan Africa (~23 per 1000 maternities), respectively [2, 3, 5, 6]. Together, these observations suggest DZ twinning is a heritable trait with an underlying polygenic inheritance. Over the years, many have attempted to illuminate the genetic basis of DZ twinning through hormone and ultrasound studies, segregation and pedigree analyses, candidate-gene approaches, and linkage projects [reviewed in ref 7]. In the end, only portions of the genetic complexity of DZ twinning have been explained, providing the opportunity to explore its genetic origin with innovative study designs.

Bulmer initially postulated that DZ twinning was due to a recessive gene with low penetrance and a gene frequency of 50% [2]. Results from subsequent pedigree-based analysis contradicted the recessive model, stating that the phenotype of ‘having DZ twins’ is consistent with an autosomal monogenic dominant model with a gene frequency of 3.5% and a female lifetime penetrance of 10% [8]. Subsequent linkage scans for DZ twinning parameterized their models accordingly and, in the end, concluded that the mode of inheritance is more complex than originally expected [9]. Further evidence for complex inheritance was demonstrated by expanded linkage efforts of affected sister pairs (at least two sisters who were both mothers of spontaneous DZ twins) from over 500 families from Australia, New Zealand, Utah, and the Netherlands, which did not return any strong linkage signals [10]. Others have added that various non-genetic factors also influence DZ twinning [3, 4], fostering additional support for the hypothesis that the mode of inheritance of DZ twinning is likely complex and unlikely to be a simple dominant or recessive trait.

The ongoing search for common genetic variants explaining DZ twinning inheritance was propelled forward by the feasibility of quantifying genetic variation at a large scale with SNP microarrays. A landmark meta-analysis of genome-wide association studies of 1,908 mothers of DZ twins and 12,953 controls identified and replicated an association of DZ twinning with common genetic variants in *FSHB* and *SMAD3* [11]. Since, additional efforts have demonstrated replication of these associations and have extended the search to uncover the genetics of multiple births [12]. Still, identification of rare and low-frequency genetic variants with substantial effect, accounting for more than a tiny fraction of variation in DZ twinning, has remained elusive [13]. Opposed to the common genetic variation captured by microarrays and imputed datasets, one approach for rare variant identification is to analyze whole-genome sequence data. Sequence data obtained from large informative pedigrees can be examined to search for possible highly penetrant driver variants.

Here, we employed such a design to identify rare and/or functional variants associated with DZ twinning using combined within-family linkage information and whole-genome sequence data. We identified a large pedigree with a rich history of DZ twinning, containing 18 MoDZT. With DNA extracted from samples provided by 17 individuals (4 males, 13 females [11 of which are MoDZT]), we performed genotyping and whole-genome sequencing experiments to generate datasets for linkage analysis and variant identification. We reasoned that large genetic regions shared by the most distantly affected MoDZT contain novel variants with considerable effect, leading to an enhanced understanding of biological pathways important for the DZ twinning process.

METHODS

Pedigree description

A large Dutch pedigree with a rich history of spontaneous DZ twinning (i.e., no use of assisted reproductive technologies) was ascertained. There are 21 sets of DZ twins and 18 MoDZT spanning multiple generations (Figure 3.1). Samples were collected from 17 individuals (4 males and 13 females). Of the four male samples, two were part of a same-sex DZ twin pair. Of the 13 females, 11 are MoDZT, one of which is part of an opposite-sex DZ twin pair. Two of the MoDZT gave birth to two sets of DZ twins. DNA was extracted in the Netherlands and sent to the Avera Institute for Human Genetics (AIHG) for SNP genotyping and sequencing.

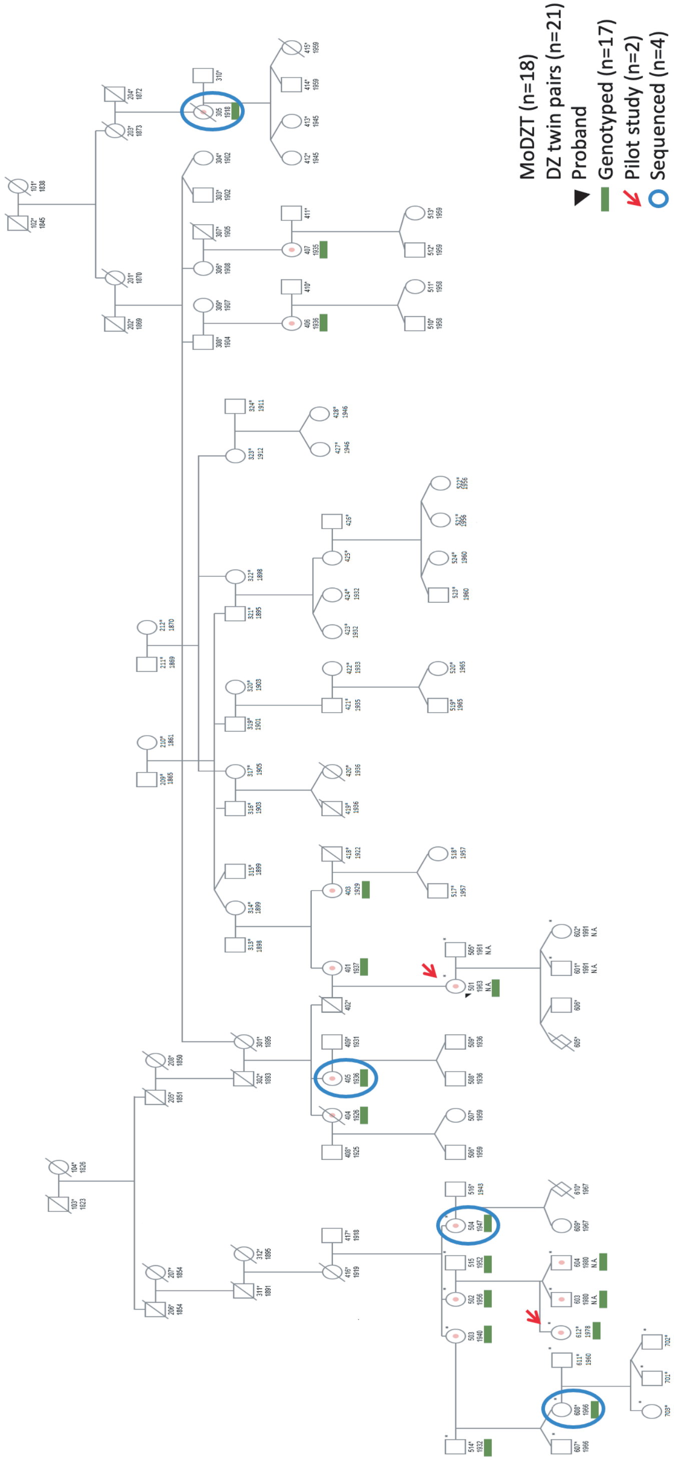


Figure 3.1 – Large Dutch pedigree ascertained for multiple MoDZT

Sample quality control and genotyping

Briefly, DNA purity was assessed with a Nanodrop spectrophotometer. DNA quantity was measured with a double-stranded DNA dye method using a Qubit Fluorometer. All samples were of sufficient quality and quantity for downstream genotyping and were normalized to a concentration of 50ng/uL. SNP genotyping was done on the Illumina GSA according to the manufacturer's protocol. Input for sample target preparation was 200ng of high-quality genomic DNA. Genotype calls were made with GenomeStudio2.0 and exported in PLINK file format for downstream analysis.

We observed more than expected allele sharing between individual 305 and many of the genotyped MoDZT from the far left-hand side (paternal side from proband) of the pedigree. Individual 305 is related to the individuals in that cluster only through a marriage of individuals 302 and 301, so would be expected to show minimal allele sharing with any member of that cluster, akin to the allele sharing of two unrelated individuals. We extracted a subset of SNPs from the whole-genome sequence data of individual 305 to re-calculate genome-wide IBD sharing. The same pattern of allele sharing was observed.

Linkage analysis

Analysis was performed with Merlin software [18] with a grid size of 0.1. SNPs with Mendelian inconsistencies, minor allele frequency (MAF)<0.01, and substantial deviation from Hardy-Weinberg Equilibrium (p<0.00001) were excluded prior to analysis. Genotypes for key individuals for which samples were unavailable were set to missing. Mothers without twins were assigned an unknown status rather than unaffected because it is possible that these mothers possess the genetic regions of interest but did not (yet) express the trait. For this reason, mothers without twins could not definitively be specified as unaffected.

In the pedigree, two subfamilies were defined with respect to the proband (sample number 501 marked by the black triangle in Figure 3.1). The first cluster contained individuals 202, 201, 302, 301, 308, 309, 306, 307, 313, 314, 401, 403, 404, 405, 402, 406, 407, 501. The second cluster was represented by individuals 416, 417, 514, 503, 502, 515, 504, 608, 147, 603, 604. Sample 305 was omitted because of absent genetic relations with individuals on the left side of the pedigree.

Whole-genome sequencing

<u>Pilot study</u>: two samples (individuals 501 and 612) were sequenced as a part of a pilot study to validate the performance of a sequencing library-preparation kit initially designed for cell-free DNA (DNA fragments) and that had not previously been used at the AIHG. The two samples were selected strictly based on having abundant high-quality genomic material available.

The samples were sequenced to evaluate sequencing quality with the library preparation kit. The two samples were included on an available lane of a flow cell as part of another sequencing project. Though low read coverage (~5X) was expected, quality could still be assessed.

Sample selection - Four MoDZT were selected for whole-genome sequencing. Individuals 608, 504, 405, and 305 were selected based on being the most distantly affected in the pedigree. Sequencing experiments were designed to obtain quality data of sufficient coverage (~30X) for rare variant identification.

Sample preparation – DNA was first fragmented via sonication to a 300 base-pair (bp) target size with a Covaris M220 Focused ultrasonicator. DNA fragment size was confirmed with an Agilent 2100 Bioanalyzer.

Library preparation – Sequencing libraries were generated from the fragmented DNA with ThruPLEX Plasma-seq chemistry (Rubicon Genomics). Positive and negative controls were included, in the form of a known reference genomic DNA sample and a non-template control (water in TE buffer), respectively. Index read sequencing primers of 6bp were included for multiplex sequencing. Compatibility of indices was determined with Illumina Experiment Manager. Libraries were purified with Agencourt AMPure XP beads. Prepared libraries were then assessed via Agilent 2100 Bioanalyzer to verify the addition of sequencing adaptors and indices (~140bp increase).

Library pooling – Libraries were pooled, purified, and quantified via quantitative polymerase chain reaction (PCR) with a KAPA Library Quantification Kit. Libraries were then denatured with 0.1 N NaOH and diluted to 15 pM for optimal cluster generation. Flow cell clustering was performed with an Illumina cBot 2 system.

Sequencing – Pooled libraries were sequenced on an Illumina Hi-Seq 2500 instrument using a high-output, 2x101 paired-end sequencing run with 1% PhiX spike-in serving as a control to aid in experiment troubleshooting.

Whole-genome sequence analysis

Sequence data were analyzed locally on a Linux workstation (OS: Ubuntu, Intel Core i7 6900 8 Core, 128 GB RAM) at AIHG in a stepwise manner. Initial quality was assessed with MultiQC [19] on raw FASTQ files. Pre-processing and variant discovery and identification of germline short variants (SNPs, insertions, deletions) were performed with the germline best practices workflow of the Genome Analysis Toolkit (GATK version 3.8) [15, 20, 21] (Figure 3.2). Reads were mapped to the reference human genome (GRCh37) with BWA-mem (v0.7.17). All reference variant databases were obtained from the GATK resource bundle. Duplicated alignments were marked with Picard tools (v2.14.1) (http://broadinstitute. github.io/picard/). Alignment scores were recalibrated with the *Base Quality Score Recalibration* (BQSR) module in GATK. Variants were called with *HaplotypeCaller* in GVCF mode for each sample and were then consolidated for joint calling to obtain raw variants. Variant quality scores were recalibrated with the *Variant Quality Score Recalibration* (VQSR) module.

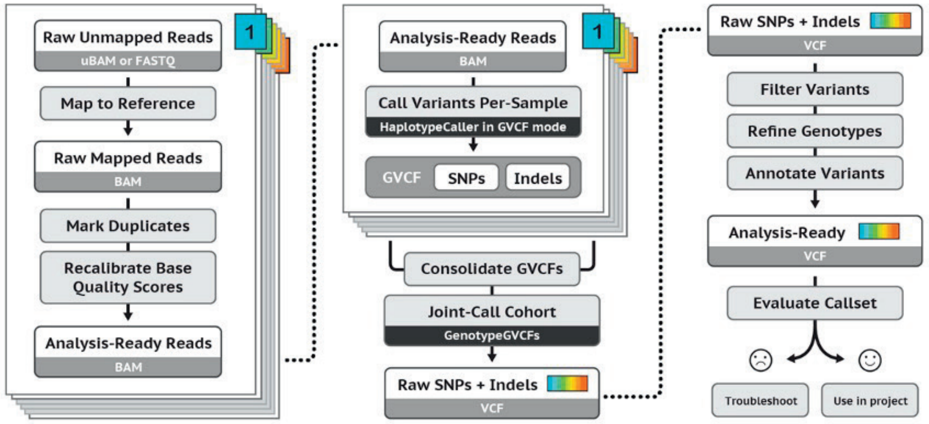


Figure 3.2 - Complete GATK best practices workflow (https://gatk.broadinstitute.org/hc/en-us).

Identification of shared genomic regions

Shared genomic regions were identified with Olorin [22], a Java package designed to combine within-family linkage analysis with sequence data (Figure 3.3). Olorin integrates patterns of gene flow estimated by Merlin to identify genomic regions shared by selected (i.e., affected) individuals in large pedigrees. This information can then be combined with whole-genome sequence data (single VCF file) to analyze variants within the shared regions. Variants can further be refined with filtering tools in Olorin. For example, a user may define the minimum number of individuals required to share a segment, enabling the search for variants of incomplete penetrance. We adjusted this option to search for variants possessed by two, three, or all four selected MoDZT. Additionally, Olorin supports the processing of 'consequence' strings in the information field of the VCF file for predicting variant effects. Consequence information can be obtained from Variant Effect Predictor (VEP).

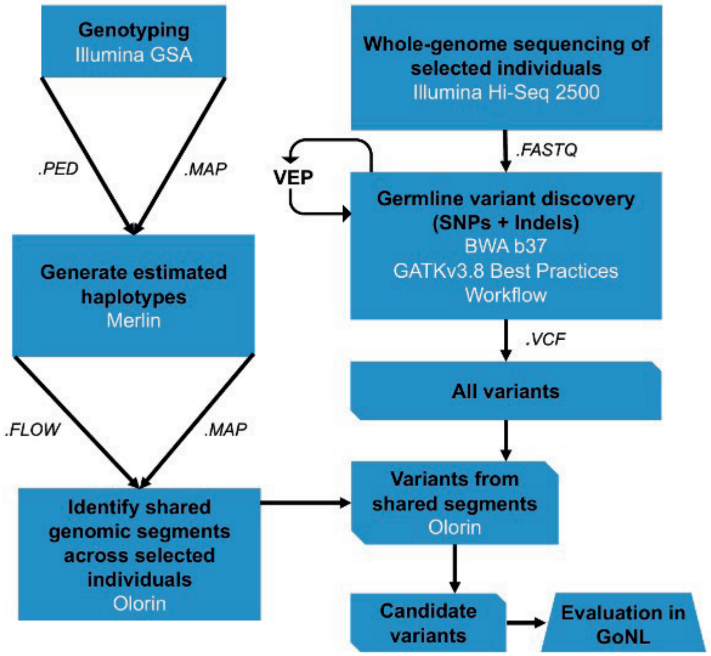

Genotyping
Illumina GSA
Whole-genome sequencing of selected individuals
Illumina Hi-Seq 2500
.PED
.MAP
VEP
.FASTQ
Germline variant discovery (SNPs + Indels)
BWA b37
GATKv3.8 Best Practices Workflow
Generate estimated haplotypes
Merlin
.VCF
All variants
.FLOW
.MAP
Variants from shared segments
Olorin
Identify shared genomic segments across selected individuals
Olorin
Candidate variants
Evaluation in GoNL

Figure 3.3 - Diagram of the Olorin workflow

RESULTS

Linkage

The DZ twinning pedigree (Figure 3.1) was previously analyzed by Dr. Hamdi Mbarek with custom identity-by-descent (IBD) mapping programs written in Wolfram Mathematica software. The pedigree was split into three clusters, two on the paternal side of the proband and one on the maternal side. Numerous large (>1Mb) shared regions were identified, depending on the selected individuals included in the analysis. A 1Mb region on chromosome 12 was shared by 11 MoDZT and one grandmother of DZ twins from the maternal and paternal sides of the proband. IBD regions shared by seven of the MoDZT on the paternal side of the proband were on chromosomes 15, 16, and 17.

Due to the complexity of the phenotype and uncertainty of the mode of inheritance, we employed non-parametric linkage analysis with an affected-only model to test for the co-segregation of chromosomal regions and being a MoDZT. Under the null hypothesis, the average Logarithm of Odds scores (LODs) should be zero in non-parametric linkage analysis. Negative non-parametric LODs imply less than expected allele sharing among the group of individuals and suggest that linkage is less likely. An excess of negative LODs indicates that the data contain genotyping errors and/or misspecification of familial relationships. Positive non-parametric LODs indicate excess allele sharing among affected individuals and favors the presence of linkage. By convention, LODs greater than 3 are considered strong evidence of linkage since they represent 1000 to 1 odds that a trait gene is linked to a genetic marker. LODs less than -2 are generally considered evidence to exclude linkage.

The goal of the non-parametric analysis was to identify large, shared regions indicated by broad peaks or plateaus in LODs plots. Per chromosome LODs are shown in Figure 3.4. The top hit was found on chromosome 5 (maximum LOD score=1.21, p=0.009). Maximum LOD scores were positive for all chromosomes, except for chromosome 21 (maximum LOD score=-0.01, p=0.6).

Aside from the complexity of the pedigree structure and the absence of genetic data for key individuals, issues related to impossible recombination were experienced in Merlin. Impossible recombination patterns resulted in one of the three family clusters being discarded. The origin of this issue is an obligate recombination event between two markers that are mapped to the same position, or very close to each other, or that have a recombination probability of zero given the genetic recombination map used. Another reason for this may be a possible point mutation or genotype error. A subset of markers (N=18,555) was excluded in the quality control step before analysis to resolve issues caused by the impossible recombination patterns.

Whole-genome sequence data quality and analysis

<u>Pilot study</u>: The initial library preparation and sequencing of samples 501 and 612 demonstrated that high-quality sequence data could be generated with the ThruPLEX Plasma-seq chemistry, originally designed for fragmented, cell-free DNA. The pilot study results confirmed the application of this library preparation kit for the whole-genome sequencing experiment on selected MoDZT.

Whole-genome sequencing was performed on four of the most distantly affected MoDZT in the pedigree. The sequence data from the four selected MoDZT were of high quality. The high output run yielded 343.63Gbp of data, with an average raw error rate of 0.301%. The average cluster density was ~790 K/mm^2 per flow cell lane (samples were pooled across all eight lanes). A vast majority of bases (94.21%) had Phred Quality Scores above Q30, indicating a base call accuracy of 99.9% for those bases. The mean GC content of reads for each sample was roughly normally distributed and was consistent with the mean value in the human genome of 41% [14].

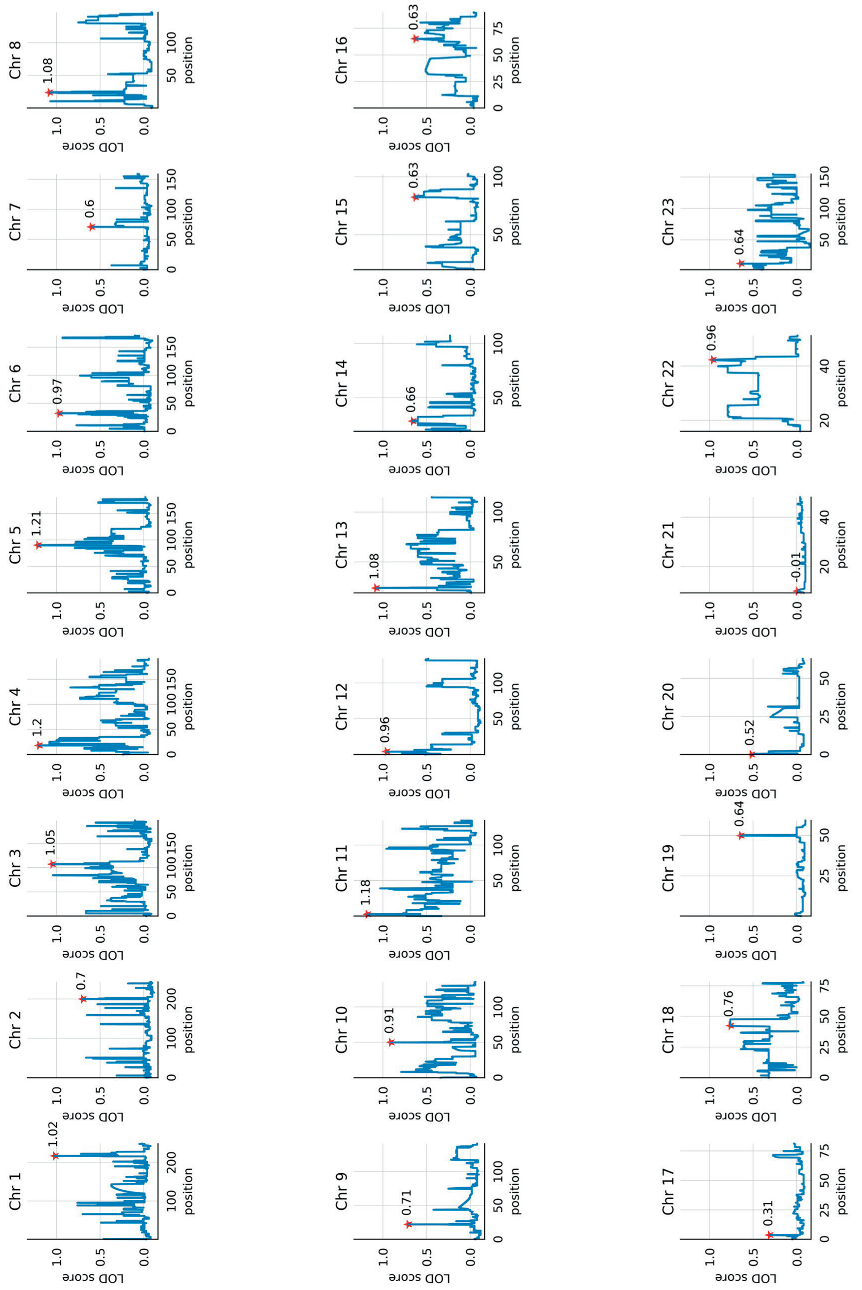


Figure 3.4 – Per chromosome LODs plots of the non-parametric linkage analysis for MoDZT

Although the sequence data were of high quality, the yield (343.63Gb) was less than projected for obtaining the desired ~30X coverage for confidently identifying rare variants. Sequencing instruments (identical instruments within the same lab or between labs) are known to vary in sequencing yield, provided a given input library concentration. Historical data from sequencing projects on the Illumina Hi-Seq 2500 at the AIHG suggested 15 pM as the optimal concentration for clustering and achieving the desired coverage. Lower data output can be due to over- or under-clustering. Over-clustering tends to result in poor image resolution, lower Q30 scores, and reduced data output. Alternatively, under-clustering usually maintains data quality but with lower overall data output. Given the robust quality results, the under-clustering scenario most likely reflects the lower-than-expected average coverage depth of 18.60X (range 17.37X to 19.51X) for the four sequenced samples following alignment and initial quality control.

A summary of the variants identified by GATKv3.8 is shown in Table 3.1. The results are shown for all four selected MoDZT. For a whole-genome sequencing experiment, roughly 4.4 million variants per individual are expected for human germline data (estimates from GATKv4). In total, nearly 7 million total variants were identified. We expected fewer total variants given the small number of sequenced individuals, the degree of relatedness amongst them, and strict variant filtering to avoid false positives. Our results were consistent with the known effects of sample size, filtering strictness, sample ethnicity, and state of the variant calling algorithm on variant discovery and identification with GATKv3.8.

The transition/transversion (Ti/Tv) ratio of 2.05 falls in line with a reported ratio of 2.0-2.2 for humans across the entire genome [15], indicating very few false positives and no bias due to artifactual variants. The bias avoidance is also supported by the insertion/deletion ratio of 0.83 (expected to be ~1) for common SNPs (https://gatk.broadinstitute.org/hc/en-us/articles/360035531572-Evaluating-the-quality-of-a-germline-short-variant-callset).

Table 3.1 - Summary of variants after filtering

Category	**dbSNP (b37)**	**Novel**	**Total**
Ti/Tv Ratio	2.10	1.88	**2.05**
SNPs:			
N (% total)	4,793,188 (80.07%)	1,192,935 (19.93%)	**5,986,123**
Insertions:			
N (% total)	238,787 (54.29%)	201,013 (45.71%)	**439,800**
Deletions:			
N (% total)	287,568 (52.05%)	264,927 (47.95%)	**552,495**
Insertion/Deletion Ratio	0.83	0.76	**0.80**

Ti/Tv is the transition to transversion ratio.

Identification of shared genomic regions

Estimated haplotypes were first generated in Merlin with SNP genotype data. The gene flow output was used to identify shared genomic regions of MoDZT. In the form of a VCF file, analyzed sequence data were then combined to identify variants within the shared segments. Individuals with sequence information were specified with the interactive features of Olorin and the information in the required pedigree file (Figure 3.5).

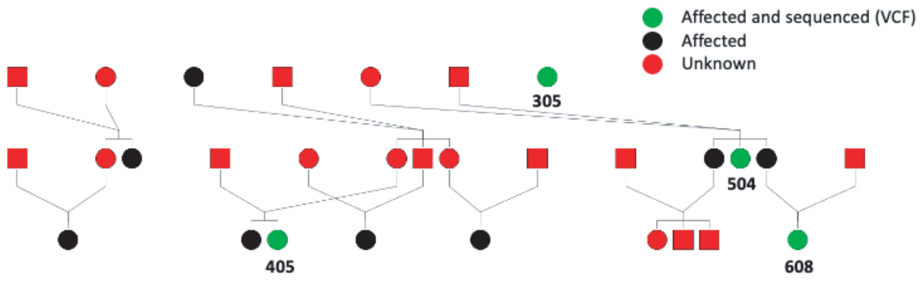


Figure 3.5 – Pedigree as defined in Olorin

Additional filtering was performed to identify shared genomic segments in two, three, or four sequenced MoDZT. The results are shown in the ideograms in Figure 3.6. Any two MoDZT shared large regions of all chromosomes. Regions shared by all four MoDZT were found on chromosomes 1-7, 10, 11, 16, 17 and were variable in size. The largest continuous regions were on chromosomes 11, 1, 3, and 6, respectively. Of the regions shared by all four MoDZT, none contained the previously identified and replicated SNPs near *FSHB* (rs11031006; Chr11; GRCh37 position 30,226,528) and within *SMAD3* (rs17293443; Chr15; GRCh37 position

67,437,863) [11]. The closest segment to rs11031006 was 18.2Mb upstream. No segments shared by all four MoDZT were found on chromosome 15. The shared regions did not overlap with an identified but not replicated intergenic SNP, rs12064669 (Chr1, GRCh37 position 230,688,643). The closest shared segment was 2.4Mb downstream. Assessment of the largest segments shared by three MoDZT revealed a region on chromosome 11 containing the *FSHB* associated SNP rs11031006. The region was rather large, spanning 30.9Mb (17,663,444 start; 48,581,765 end), and was shared by individuals 305, 608, and 504. Individual 305 possessed 14 of the 15 largest shared segments possessed by any 3 MoDZT.

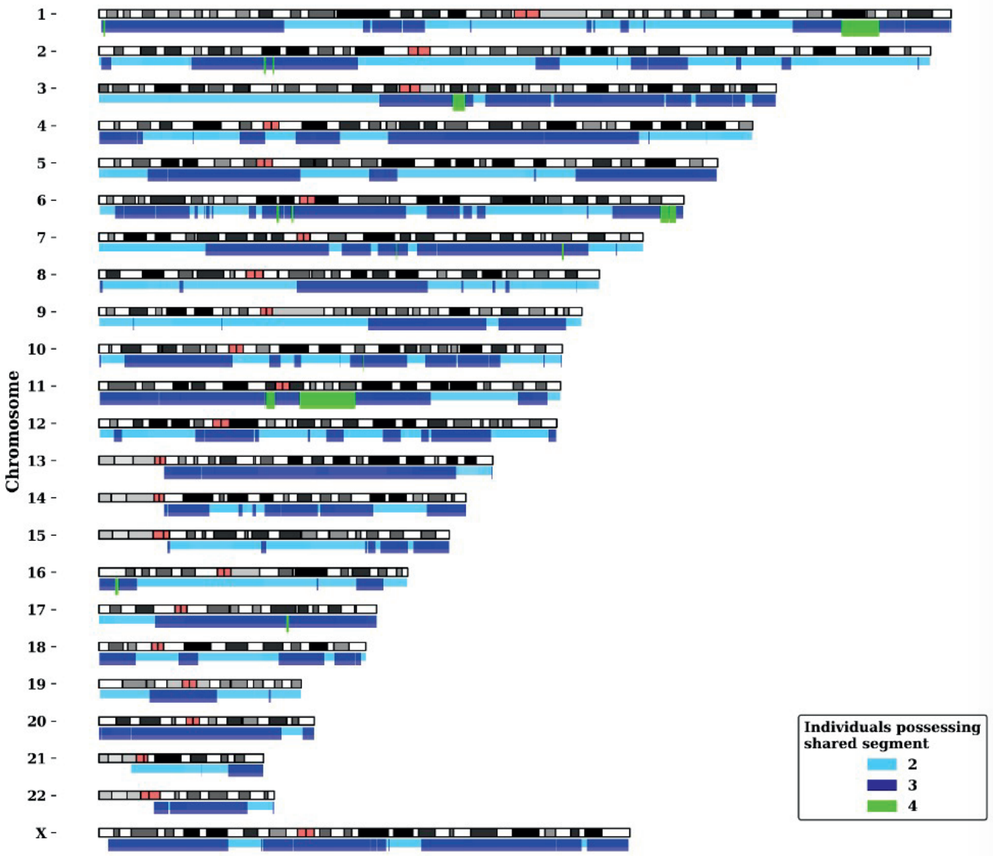


Figure 3.6 – Ideograms of shared genomic segments
Banding patterns of chromosomes are shown in grayscale, with the centromere colored in red. Note: segments smaller than ~50kb are extremely difficult to visualize due to their size relative to each chromosome. For example, on chromosome 4, the first segment shared by 3 mothers has a small gap, corresponding to a 4,677 base-pair region shared by 4 mothers.

CONCLUSIONS AND FUTURE DIRECTIONS

The enduring objective of this project is to use whole-genome sequencing as a follow-up approach to previous linkage and association studies of DZ twinning to identify new genetic biomarkers related to fertility measures and for the prediction of DZ twinning.

Based on previous work, variants affecting multiple ovulation rates (i.e., DZ twinning events) are most likely to occur in genes and pathways that control the synthesis and release of Follicle Stimulating Hormone (FSH), pathways in the ovary that control response to FSH, or pathways involved in growth and development of the dominant follicle. Results from whole-genome sequencing may implicate new pathways or novel routes for regulating known pathways, ultimately enabling new opportunities for treating infertility or fine-tuning assisted reproductive strategies.

We have identified genomic regions of interest with combined genotype and whole-genome sequence data, but considerable effort is still required to pinpoint specific (rare) variants with meaningful biological effects. In a first step, preliminary results of Olorin can be further analyzed to determine the functional consequence of particular variants, which will require analyzing the currently available VCF with Variant Effect Predictor (VEP) to obtain functional consequence information. The resulting VCF can then be reanalyzed in Olorin with subsequent filtering to help prioritize variants for further investigation.

Another useful strategy for evaluating the results is to screen the shared segments/variants possessed by the MoDZT from the pedigree against the genomes of 46 MoDZT from the Genome of the Netherlands project. This strategy will highlight and differentiate between variants possessed only by mothers in the pedigree and variants prevalent among all MoDZT from a population-matched cohort. Given global differences of DZ twinning [5, 6], it would be of further interest to investigate variants across MoDZT from diverse populations as sequence data become available.

Varying the unaffected status for specific individuals in the pedigree to an unknown status may also aid in identifying promising candidate regions in linkage analysis. However, these modifications and subsequent interpretation will need to be done with extreme and deliberate care.

Attention should also be given to the reference genome/resource bundle used for variant discovery and identification. The developers of GATK have recently transitioned all tool development and support to GRCh38 since retiring the GRCh37 resource bundle. The GRCh38 assembly is an improved version of the human genome reference [16], so it would be diligent to repeat all analyses employing this reference. This idea is supported by a recent study that found significant variant calling discrepancies due to the intrinsic differences between GRCh37 and GRCh38 [17]. Implementation of the GRCh38 reference would necessitate that genotype coordinates be converted (i.e., lifted over) to a consistent build for reanalysis with Olorin.

Overall, the overlapping portions represent regions of interest for identifying rare and highly penetrant variants or deleterious mutations. Rare variant identification for DZ twinning may elucidate novel genetic biomarkers for fertility and improve the ability to predict twinning events. Findings from this work have the potential to improve the outcomes of multiple gestation pregnancies and the reproductive capacity of infertile couples.

REFERENCES

1. Hoekstra, C., et al., *Familial twinning and fertility in Dutch mothers of twins.* Am J Med Genet A, 2008. **146A**(24): p. 3147-56.
2. Bulmer, M.G., *The Biology of Twinning in Man*. 1970: Clarendon. 206.
3. Hoekstra, C., et al., *Dizygotic twinning.* Hum Reprod Update, 2008. **14**(1): p. 37-47.
4. Hoekstra, C., et al., *Body composition, smoking, and spontaneous dizygotic twinning.* Fertil Steril, 2010. **93**(3): p. 885-93.
5. Monden, C., G. Pison, and J. Smits, *Twin Peaks: more twinning in humans than ever before.* Hum Reprod, 2021.
6. Smits, J. and C. Monden, *Twinning across the Developing World.* PLoS One, 2011. **6**(9): p. e25239.
7. Boomsma, D.I., *The Genetics of Human DZ Twinning.* Twin Res Hum Genet, 2020. **23**(2): p. 74-76.
8. Meulemans, W.J., et al., *Genetic modelling of dizygotic twinning in pedigrees of spontaneous dizygotic twins.* Am J Med Genet, 1996. **61**(3): p. 258-63.
9. Derom, C., et al., *Genome-wide linkage scan for spontaneous DZ twinning.* Eur J Hum Genet, 2006. **14**(1): p. 117-22.
10. Painter, J.N., et al., *A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins.* Hum Reprod, 2010. **25**(6): p. 1569-80.
11. Mbarek, H., et al., *Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility.* Am J Hum Genet, 2016. **98**(5): p. 898-908.
12. Mbarek, H., et al., *Biological insights into multiple birth: genetic findings from UK Biobank.* Eur J Hum Genet, 2019. **27**(6): p. 970-979.
13. Gajbhiye, R., J.N. Fung, and G.W. Montgomery, *Complex genetics of female fertility.* NPJ Genom Med, 2018. **3**: p. 29.
14. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
15. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
16. Schneider, V.A., et al., *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.* Genome Res, 2017. **27**(5): p. 849-864.
17. Li, H., et al., *Exome variant discrepancies due to reference-genome differences.* Am J Hum Genet, 2021. **108**(7): p. 1239-1250.
18. Abecasis, G.R., et al., *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees.* Nat Genet, 2002. **30**(1): p. 97-101.
19. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
20. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.

21. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11 10 1-33.

22. Morris, J.A. and J.C. Barrett, *Olorin: combining gene flow with exome sequencing in large family studies of complex disease.* Bioinformatics, 2012. **28**(24): p. 3320-1.

4

GENETIC SIMILARITY ASSESSMENT OF TWIN-FAMILY POPULATIONS BY CUSTOM-DESIGNED GENOTYPING ARRAY

Published as
Beck, J.J., Hottenga, J.J., Mbarek, H., Finnicum, C.T., Ehli, E.A., Hur, Y.M., Martin, N.G., de Geus, E.J.C., Boomsma, D.I. and Davies, G.E. (2019)
Genetic Similarity Assessment of Twin-Family Populations by Custom-Designed Genotyping Array. Twin Res Hum Genet, **22**, 210-219.

ABSTRACT

Twin registries often take part in large collaborative projects and are major contributors to genome-wide association (GWA) meta-analysis studies. In this article, we describe genotyping of twin-family populations from Australia, the Midwestern USA (Avera Twin Register), the Netherlands (Netherlands Twin Register), as well as a sample of mothers of twins from Nigeria to assess the extent, if any, of genetic differences between them. Genotyping in all cohorts was done using a custom-designed Illumina Global Screening Array (GSA), optimized to improve imputation quality for population-specific GWA studies. We investigated the degree of genetic similarity between the populations using several measures of population variation with genotype data generated from the GSA. Visualization of principal components analysis (PCA) revealed that Australian, Dutch, and Midwestern American populations exhibit negligible interpopulation stratification when compared to each other, to a reference European population, and to globally distant populations. Estimations of fixation indices (F_{ST} values) between the Australian, Midwestern American, and Netherlands populations suggest minimal genetic differentiation compared to the estimates between each population and a genetically distinct cohort (i.e., samples from Nigeria genotyped on GSA). Thus, results from this study demonstrate that genotype data from Australian, Dutch, and Midwestern American twin-family populations can be reasonably combined for joint-genetic analysis.

Keywords: Genetic similarity assessment, genotyping microarray, population genetics, population structure, principal component analysis, twin

INTRODUCTION

Scientific investigations aimed at disentangling the contribution of genetic factors underlying complex and polygenic traits have demonstrated the necessity of large sample sizes [1, 2]. Only when sample sizes are vast is it possible to estimate the contribution of each locus influencing a complex trait [3-5]. It is both difficult and financially challenging for a single site to accrue large enough sample sizes to achieve adequate statistical power. Therefore, one pragmatic approach for obtaining the large numbers of samples required is to aggregate samples collected by different groups, either through meta- or mega-analysis. Currently, twin registers from around the world routinely employ this strategy for genotypic and phenotypic data [6, 7]. This approach is powerful if genetic heterogeneity (e.g., as a result of dissimilar population ancestry and demographic histories) is not an issue or is appropriately accounted for. Here, we explore the degree of genetic similarity between multiple twin cohorts and indicate whether it is appropriate to combine data from these cohorts for joint-genetic analysis.

In 2006, a study by Sullivan *et al*. empirically showed that samples from Australian and Netherlands Twin Registers could be reasonably combined for joint-genetic analyses by estimating the proportion of total genetic variability attributable to the genetic difference between cohorts [8]. The calculation of the genetic variability attributable to genetic differences between cohorts, measured by Wright's fixation index (F_{ST} value), was estimated using analysis of molecular variance on 359 short tandem repeat polymorphism markers. The estimated F_{ST} between Australia (N=519) and the Netherlands (N=549) was found to be 0.30%, a value smaller than between many other European groups. The F_{ST} estimates suggested that it is reasonable to combine samples from Australian and Dutch cohorts but admittedly based on calculations in samples of modest size. Here we evaluate the genetic similarity in larger numbers of samples and augment the comparison by adding a third cohort of samples obtained from the Avera Twin Register (ATR), a representative population sampling of the Midwestern region of the United States. In this study, we test the genetic variation within and between three populations of interest - Australian, Dutch, and Midwestern American - by employing genomic data from a custom-designed genome-wide single nucleotide polymorphism (SNP) array. To further explore the genetic similarity across the cohorts under study, we incorporated genetic data from a globally and genetically distinct population - samples from Nigeria genotyped on the Global Screening Array (GSA) at the Avera Institute for Human Genetics (AIHG).

In collaboration with the Netherlands Twin Register [9-12], the AIHG (Sioux Falls, SD, USA) created the ATR in May 2016 [13]. The goal of the ATR is to study the genetic and environmental influences on health, disease, and complex traits by harnessing the power of longitudinal biological sample collection and survey correspondence. Participants have enrolled from across all regions of the USA, the great majority coming from Midwestern states, including South Dakota, North Dakota, Minnesota, Iowa, and Nebraska. In addition to serving as a prime research model for studying health and disease in a regional setting, another important role of the ATR is to contribute to consortia-driven large-scale genetic studies focusing on the genetic underpinnings of complex traits. Therefore, it is of interest to recognize the degree of genetic similarity between the Midwestern Americans comprising the ATR and the cohorts for which the genetic data are to be combined.

As long-established twin registers, the Netherlands and Australian Twin Registers have served as models for newly formed twin registers from around the world. As is the case for the ATR, the Netherlands and the Australian Twin Registers are population-based, with recruitment focused on the presence of twins or higher-order multiples in the family. Through this collaborative initiative, we included 100 saliva samples from Nigerian mothers of twins to use as a genetic contrast group to Australian, Dutch, and Midwestern American populations. The incorporation of genetically distinct samples further enhances the cross-ethnic comparisons that we describe here.

The AIHG recently joined the Illumina-initiated GSA consortium. Broadly, the goal of the consortium is to enable a variety of genotyping applications for biobanks, disease research, translational research, consumer genomics, and population genetic studies. Specifically, the GSA has been optimized for high-throughput population-scale studies at a lower cost than previous genotyping platforms. Participation of the AIHG in the GSA consortium has allowed for the unique opportunity to design a customized high-density SNP genotyping microarray. A similar strategy for designing population-specific arrays for genome-wide association (GWA) testing has already been described, albeit for a different genotyping platform [14].

Here, we report on the design and initial validation of the array, as assessed by the evaluation of concordance, coverage, and imputation quality of the core backbone against the Genome of the Netherlands (GoNL) reference set [15, 16]. Additionally, we provide evidence to suggest that the custom-selected content generally enhances imputation quality and provides robust genotype calls for population- and disease-relevant SNPs. Furthermore, we demonstrate that the GSA can be utilized to generate high-quality SNP data from multiple tissue sources, namely blood, buccal epithelial brushings, and saliva.

With high-density SNP genotype data obtained from the GSA run at AIHG, we assessed the level of genetic similarity across population cohorts of interest: Australian, Dutch, and Midwestern American. To facilitate the assessment of population genetic structure, we leveraged the power of state-of-the-art software capable of ingesting genome-wide SNP data obtained from the GSA. Population genetic variation was summarized by uncorrelated principal components (PCs) through principal components analysis (PCA) and estimations of F_{ST} values. Furthermore, we projected the PCs estimated from the samples onto data from the Human Genome Diversity Project (HGDP) [17]. Projection of calculated PCs onto the diverse populations comprising the HGDP fostered a global illustration of genetic relatedness between the populations of interest.

MATERIALS AND METHODS

Participants and Sample Collection

Australian subjects were from the QIMR Berghofer Medical Research Institute (N=1922) [18, 19]. Other subjects were registered participants of the Netherlands (NTR, N=10,226) [10] and Avera (ATR, Sioux Falls, SD, USA, N=602) [13] Twin Registers, and the Nigerian Twin and Sibling Registry (NTSR, N=100) [20] (see Table 4.1).

Table 4.1 - Characteristics of samples genotyped on GSA per cohort and tissue

Cohort	Country of Origin	Sample				
		N	Female (%)*	Composition	Unrelated Individuals	Tissue
Avera Twin Register	USA	602	66.4	MZ, DZ, parents, sibs	238	Buccal
NTR	Netherlands	1135	55.4	MZ, DZ, parents, sibs	6139	Blood
NTR	Netherlands	9091		MZ, DZ, parents, sibs		Buccal
Australian	Australia	1922	100	MODZT	1448	Blood
Nigerian	Nigeria	100	100	MOSDZ	96	Saliva
Total		**12,850**			**7921**	

*GSA = Global Screening Array; NTR = Netherland's Twin Register; N = number of samples; MZ = monozygotic twins; DZ = dizygotic twins; MODZT = mothers of DZ twins; MOSDZ = mothers of opposite-sex dizygotic twins; sibs = siblings; parents = parents of twins. *Percentage of female samples is reported from the unrelated set.*

Representative samples of the Midwestern American population were obtained from the ATR. Enrolled participants include twins, multiples, siblings, and their parents. Participants complete surveys and questionnaires and provide a cheek swab (buccal brushing) for zygosity testing and genotyping. The majority of the enrolled participants are located in the Midwestern region of the United States, with most being from South Dakota, Minnesota, and Iowa.

Samples from the QIMR Berghofer Medical Research Institute (Australia) are a combination of a number of different studies, conducted in many countries over many decades, focused on the genetics of dizygotic twinning [21]. Samples from mothers of dizygotic twins (MODZT) were collected from Australia and New Zealand and were shipped to the AIHG for genotyping on the GSA regardless of if they were ungenotyped or previously genotyped on an earlier SNP array. Included in the shipment were two small cohorts of special interest: (1) a Belgian sample of 40 MODZT from 14 multiplex families collected in the 1990s; (2) a sample of 10 MODZT from two multiplex families from the Utah Mormon Database collected in 1994. For the purposes of the study presented here, samples from Belgium and Utah were excluded from the Australian cohort.

The Nigerian sample in the present study was drawn from the NTSR, which included over 3000 adolescent monozygotic and dizygotic twins, their parents, and siblings collected mainly from public schools in Lagos State and Abuja, Federal Capital Territory in Nigeria. Participants of the NTSR completed questionnaires and provided saliva or buccal samples for genotyping. The sample used in the present study consisted of 100 mothers of opposite-sex twins attending public schools in Lagos State, collected for the purpose of a pilot study to understand the genetic underpinnings of dizygotic twinning. Lagos State is located in the southwestern geopolitical zone of Nigeria and is one of the most populous urban areas in Nigeria. Although residents of Lagos State are ethnically diverse, they are mainly members of the Yoruba group.

Participants from the NTR included twins, their parents, and other relatives (mainly siblings of twins). NTR participants take part in surveys and other research projects and provide blood or buccal samples for DNA isolation and genotyping.

DNA Extraction and Genotyping

DNA was isolated from whole blood, buccal epithelial cells [22], and saliva using standard protocols for downstream SNP genotyping. High-density SNP genotyping for all samples was done at the AIHG (Sioux Falls, SD) using a custom-designed Illumina GSA according to the manufacturer's protocol.

Design of a Customized Genotyping Array and Generation of Genotype Data

The GSA employed for this study was designed following a previously defined strategy for designing population-specific customized genotyping arrays [14]. Specifically, the GSA was custom-designed to contain a core imputation backbone (approximately 660,000 markers) based on commonly utilized reference panels, such as the GoNL [15] and the 1000 Genomes Project [23]. In addition to the core backbone, the array includes approximately 30,000 additional markers for fine mapping to further enhance imputation quality and 8000 markers of interest associated with a variety of conditions, disorders, and traits, including neuropsychiatric disorders, drug metabolism, fertility, and twinning (Table 4.2). In total, the GSA contains 697,486 markers.

Table 4.2 - Content and marker selection categories of the custom-designed Illumina GSA

Marker Type	Number of SNPs (N=697,486)
GSA Core Backbone	**Total ~660,000**
Sex Chromosomes	17,880 X; 1480 Y; 578 PAR
ADME Genes/Exons	6668; 2787
ClinVar	17,020
MHC	9797
Ancestry Informative	3212
Fine Mapping Content (candidate genes and additional markers for imputation)	**Total ~30,000**
Custom Markers	**Total ~8000**
Fertility and Twinning	
Body Stature (height, BMI) and Sports and Exercise Behavior	
Mental State and Health (happiness, depression, schizophrenia)	
Chromosome X (imputation)	
Educational attainment	
Pharmacogenomics	

GSA = Global Screening Array; ADME = Absorption, distribution, metabolism, excretion; BMI = body mass index; ClinVar = NCBI archive for interpretations of clinical significance of genetic variants; MHC = major histocompatibility complex; PAR = pseudoautosomal region

Prior to the design of the array, initial validation of the GSA content for imputation was assessed by checking concordance, coverage, and imputation quality using an extracted subset of markers resembling the GSA (683,937 markers). In brief, a dataset mimicking the content on the GSA was curated from 249 unrelated female individuals of the GoNL project. Males from the GoNL were excluded as the tools used for assessment of genotyping array coverage could not properly handle a homozygous X chromosome. GSA-mimicked markers were quality controlled and retained if minor allele frequency (MAF) was >0.01, missingness per individual was <10%, missingness per SNP was <5%, and if there was no statistically significant deviation from Hardy-Weinberg Equilibrium ($p > 10^{-5}$). Quality control and filtering reduced the number of markers to 617,340. Extracted and quality-controlled markers were then selected if they were present in the 1000 Genomes (1000G) reference panel [23] (616,961 markers). The extracted set was phased with *SHAPEIT* [24] and imputed against the 1000G reference panel phase 3 using *IMPUTE2* [25]. For the ~12.1 million overlapping markers, concordance was calculated in *PLINK* [26, 27] by comparing the 1000G best-guess genotypes to the original GoNL genotypes.

Genotype calls from the GSA were made using Illumina GenomeStudio2.0 and custom-curated cluster files. In short, cluster positions were defined using genotype data on 1254 samples run on GSA at AIHG by a variety of technicians and across many batches (i.e., reagent and bead-chip lots) to account for as much sample variation as possible. Initial assessment of sample-dependent and sample-independent controls, preliminary call rates, and percentile distributions of GenCall scores (a quality metric indicating the reliability of genotype calls) yielded a final sample set of 1199 samples for defining cluster positions. Samples were grouped into males and females so that Y chromosome (1480 markers) and X chromosome (17,880) clusters could be generated using subsamples of the appropriate sex. Due to the behavior of the GenomeStudio clustering algorithm, only male samples were used for defining Y chromosome clusters. Similarly, only female samples were used for generating X chromosome clusters since males are not expected to be heterozygotes for X-linked markers. Therefore, X and Y markers were clustered and evaluated, taking gender into account. All samples were used to cluster autosomal SNPs (670,744 markers), including XY and mitochondrial markers.

Following initial clustering, cluster positions were evaluated and edited based on a sequential assessment of several cluster metrics. Cluster positions were zeroed (resulting in no genotype calls for a locus) based on low cluster separation (≤0.27), low call frequency (<0.96), low mean normalized intensity values for the heterozygote genotypes (≤0.2), extreme mean normalized theta values of the heterozygote cluster (<0.2 or >0.825), Mendelian inconsistencies, ambiguous clusters, excessive numbers of reproducibility errors, and excessive heterozygote calls relative to expectations based on Hardy-Weinberg Equilibrium (>0.2). Markers on X and Y chromosomes were manually evaluated and edited on a per-marker-basis.

Data Management, Quality Control, and Relationship Inference

Individual samples were removed if they had a missing rate greater than 10% or excess genome-wide inbreeding levels/heterozygosity (as calculated in *PLINK*, F coefficient <-0.10 or >0.10). Reported sex was compared with inferred sex from the genotype data. Sex mismatches were investigated, resolved, and subsequently replaced in the dataset.

From each population sample, we selected the largest group of unrelated individuals (shown in Table 4.1). Unrelated individuals were identified with *KING* software [28] using the '*--unrelated*' option. In brief, related individuals (estimated kinship coefficient <0.088) were clustered into families. Within each connected group, individuals were ranked according to the count of unrelated family members, corresponding to an estimated kinship coefficient <0.022. A set of unrelated individuals was then made by selecting the individuals with the largest count of unrelated individuals within the respective family group. Additional unrelated individuals were obtained by taking the individuals with the next most unrelated family members, only if that individual was not related to any of the previously selected unrelated individuals. The final selection contained no pairs of individuals with a 1st or 2nd-degree relationship, reducing the sample size to 7921 subjects. Following quality control, the number of samples was reduced to 7782.

SNP Quality Control and 1000G Alignment

All autosomal SNPs that passed quality control and filtering were analyzed. *PLINK* was used to perform quality control. A selection of high-performing markers (N=564,020) was used for subsequent quality control and analyses. Specifically, SNPs were removed if they were not in the 1000G reference panel (phase 3 version 5) or if they were palindromic SNPs with an allele frequency of 0.40-0.60. Polymorphic SNPs with more than two alleles were also excluded. SNP marker names were adjusted for congruity with 1000G, and strand flip issues were resolved. SNPs were removed if their call rate was less than 95% and if they differed significantly from Hardy-Weinberg Equilibrium ($p < 10^{-5}$).

Principal Component Analysis

PCA was performed with *smartpca* of the *EIGENSOFT* package [29] with its default parameter settings. PCA was used to compute 10 PCs for the populations under study. Initial ancestry outliers were determined by merging each independent dataset with 1000G data to project ethnicity with *smartpca*.

Ancestry outliers, based on non-European ancestry, were visually identified and subsequently removed.

Cleaned and 1000G aligned data for each population were filtered to retain SNPs having a MAF >0.05, linkage disequilibrium (LD) pruned and filtered to exclude confounding SNPs in long-range LD, as previously described [30, 31]. Filtering and exclusion of long-range LD regions reduced the number of autosomal SNPs from 564,020 to 109,702 SNPs. This number of SNPs was used for comparisons between samples genotyped on GSA, namely those from Australian, Midwestern American, Dutch, and Nigerian cohorts.

We also calculated if there were statistically significant pairwise differences between the Australian, Dutch, and Midwestern American populations and representative European populations from the HGDP using *smartpca*. For each pair of populations, ANOVA statistics along each eigenvector were summed across all 10 eigenvectors.

HGDP Data Management and Projection

To establish genetic similarity on a global scale, PCs of the Australian, Dutch, Midwestern American, and Nigerian populations were projected onto samples obtained from the HGDP [17, 32]. The HGDP data comprises genotypes (660,918 SNPs) from 1,043 fully consenting individuals representing 54 global populations from sub-Saharan Africa, North Africa, Europe, the Middle East, Central and South Asia, East Asia, Oceania, and the Americas and provide a representative sampling of worldwide genetic variation (available at: https://www.hagsc.org/hgdp/files.html).

Raw genetic data from the HGDP (sample call rate > 98.5%) were reformatted for *PLINK* using command line tools. Markers with greater than 5% missingness were removed. Following the same procedures as previously described, unrelated individuals were identified in the HGDP dataset and retained using the program *KING*. Removal of related individuals reduced the sample size from 1,043 to 857. To be consistent with GSA, HGDP data were converted from Build 36.1 coordinates to Build 37/hg19 using the University of California, Santa Cruz (UCSC's) batch coordinate conversion tool, *liftOver* [33, 34]. Overlapping markers between HGDP and GSA (prior to MAF filter, LD pruning, and exclusion of long-range LD) were identified in the variant information files (.map) using *R* [35]. Of the 133,833 common markers between the datasets, there were 21,667 multi-allelic variants due to strand inconsistencies. Strand flips were resolved, and data from the HGDP were merged with cleaned and filtered GSA data using *PLINK*. The merged set was filtered to remove markers with a MAF < 0.05, pruned for LD, and excluded SNPs in long-range LD. Quality control and filtering reduced the final number of markers to 54,820.

Ten principal components were calculated using *smartpca* within *EIGENSOFT* with default parameters. All HGDP populations were specified as reference populations for the PC projection.

Case-Control GWA Study

We performed a case-control GWA study (GWAS) between Midwestern American, Australian, and Dutch populations to gain insight into the degree of genetic relatedness between them. To avoid false positives, we excluded variants with a MAF < 0.10 in the quality-controlled and filtered data on unrelated individuals. Simple association testing was done in *PLINK* with the '*--assoc*' command. Two GWASs were performed, both with the Midwestern American population defined as cases and with Dutch and Australian samples serving as controls. Manhattan plots and quantile-quantile (QQ) plots were created to visualize regions of the genome that appeared statistically significant.

Calculation of F_{ST} Estimates

To quantify measures of structure in populations, we estimated F_{ST} values between Midwest American, Australian, Dutch, and Nigerian cohorts. Weir and Cockerham [36] and Hudson [37] estimators were calculated using two different software programs: *popstats* [38] and *scikit.allel* [39], implemented in Python.

RESULTS

Validation of the GSA

Imputation quality metrics, quantified by R^2 values, are presented in Table 4.3. For all 1000G imputed autosomal SNPs, including those present in African and Asian populations, the median R^2 values for the GSA are 0.02 for MAF>0.000-0.001, 0.69 for MAF>0.001-0.01, 0.97 for MAF>0.01-0.05, and 0.99 for MAF>0.05. For the selection of autosomal SNPs that were present in both the GoNL and 1000G reference data, indicative of true genetic variants in the Dutch population, the results demonstrate improved imputation quality compared to all SNPs present in 1000G. The median R^2 values of the SNPs in the GoNL and 1000G reference set are for 0.04 MAF>0.000-0.001, 0.80 for MAF>0.001-0.01, 0.97 for MAF>0.01-0.05, and 0.99 for MAF>0.05. Here, the improved imputation quality, captured by both median and mean scores, is mainly the result of the exclusion of a large number of rare SNPs (i.e., SNPs in African and Asian populations - captured by the full 1000G set), which are likely absent from the Dutch population.

Table 4.3 - Imputation quality metrics per minor allele frequency bin for the GSA

Selected SNPs	Chr	MAF range	N SNPs	Median R^2	Mean R^2	SD
1000G All SNPs[a]	1-22	>0.000-0.001	21,373,838	0.02	0.05	0.08
		>0.001-0.01	6,853,643	0.69	0.64	0.28
		>0.01-0.05	2,863,052	0.97	0.91	0.13
		>0.05	6,974,825	0.99	0.96	0.08
GoNL and 1000G[b]	1-22	>0.000-0.001	1,003,022	0.04	0.08	0.10
		>0.001-0.01	2,736,096	0.80	0.74	0.24
		>0.01-0.05	2,461,024	0.97	0.92	0.12
		>0.05	5,874,328	0.99	0.97	0.07

GSA = Global Screening Array; Chr = chromosome; MAF = minor allele frequency; N SNPs = number of SNPs; SD = standard deviation; GoNL = Genome of the Netherlands.
[a] Denotes full 1000G imputation with Asian/African/other SNPs not present in the Dutch population
[b] Denotes overlapping SNPs between GoNL and 1000G.
All monomorphic SNPs were excluded, thus only polymorphic SNPs were selected for each comparison.

Concordance of the genotyped GoNL SNPs that were reimputed with a 1000G imputation reference panel was high for most SNPs in the genome, as can be seen in Table 4.4. In the imputed data, of the 12,074,470 polymorphic variants with a MAF>0, up to 62.2% can be reimputed with very high quality. At lower levels of quality (below 80% concordant), 1.95% of the genome is not well covered.

Table 4.4 - Genotype concordance for GSA-mimicked, genotyped GoNL SNPs that were reimputed with 1000G reference panel

Concordance (%)	N SNPs	Percent
> 99	7,506,660	62.17
> 95-99	3,470,030	28.74
> 80-95	861,745	7.14
> 50-80	221,831	1.84
≤ 50	14,204	0.11

Note: Total number of 1000G SNPs that were reimputed, polymorphic and present in GoNL = 12,074,470.
GSA = Global Screening Array; GoNL = Genome of the Netherlands; N SNPs =number of SNPs

Principal Component Analysis

We performed a fine-scale PCA of unrelated subjects from Australian, Dutch, and Midwestern American populations to investigate the degree of genetic relatedness of these populations independent of other global populations. The PCA utilized 109,702 autosomal SNPs after stringent quality control, filtering, pruning, and exclusion of long-range LD regions. As seen in Figure 4.1, results of the PCA suggest that the Midwest American, Australian, and Dutch populations are not genetically distinct from one another since the clusters moderately overlap. The Midwest American cluster partially superimposes both Australian and Dutch clusters, which themselves also show a small degree of overlap. Visualization of PCs from the PCA on Australian, Dutch, and Midwestern Americans demonstrates the commonality of population clusters, thereby suggesting a high degree of genetic similarity between these populations.

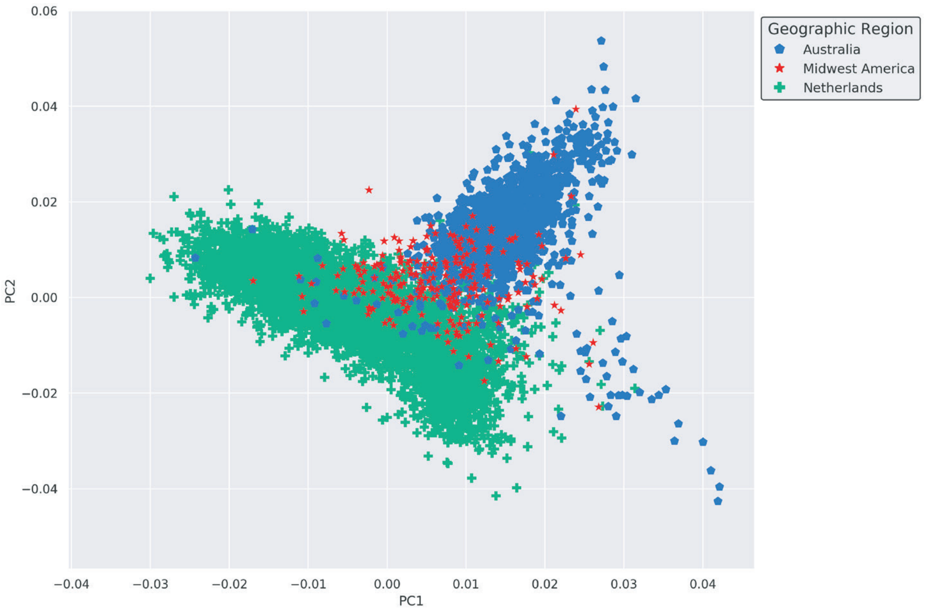


Figure 4.1 - Genetic ancestry of Midwestern American, Australian, and Dutch subjects. Shown are the results from PCA using autosomal genotyped SNPs after quality control, filtering, pruning, and exclusion of long-range LD (109,702 markers). Ancestry outliers were removed prior to performing PCA. PC1 and PC2 represent the first and second PCs and account for 18.864% and 11.919% of the variation, respectively.

Provided the unique opportunity to genotype Nigerian mothers of twins on the GSA, a PCA was performed with the inclusion of these globally distinct samples to serve as a genetic contrast group to the populations under study. Thus, to enhance the investigation of the genetic similarity of Australian, Dutch, and Midwestern American populations on a broader scale, we performed a PCA on all samples genotyped on GSA at AIHG, including those from Nigeria. The results of the PCA are depicted in Figure 4.2. The inclusion of a geographically and genetically distant population resulted in the distinct separation of European-ancestry-based populations and the Nigerian cohort, indicative of population stratification and genetic dissimilarities.

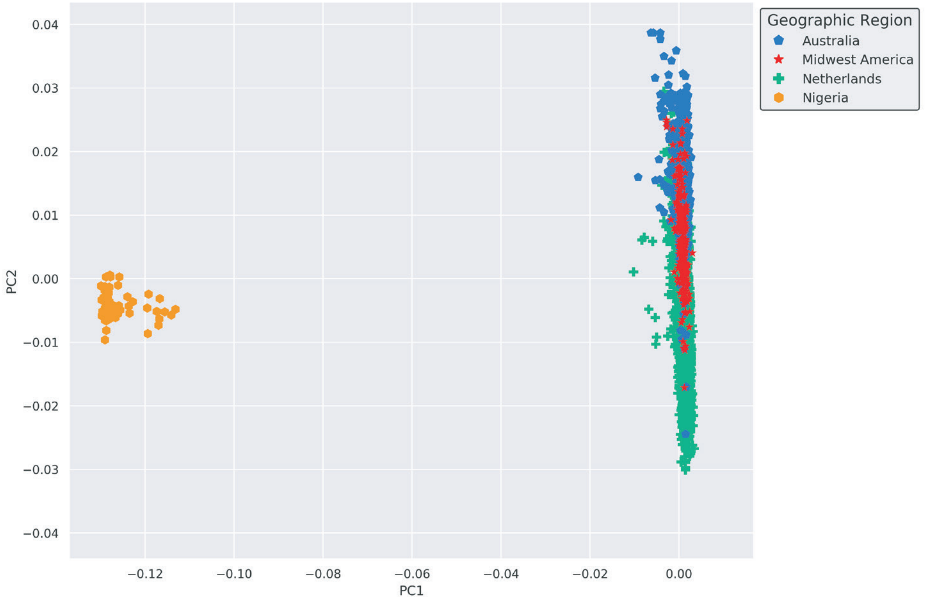


Figure 4.2 - Genetic ancestry of Midwestern American, Australian, Dutch, and Nigerian subjects.
Shown are the results from the PCA using all autosomal genotyped SNPs after quality control, filtering, pruning, and exclusion of long-range LD (109,702 markers). Ancestry outliers were removed prior to performing PCA. PC1 and PC2 represent the first and second PCs and account for 67.765% and 6.568 % of the variation, respectively.

Projection of PCs from all samples genotyped on GSA onto those from the HGDP allowed for visualization of populations on a globally diverse scale. Results of the PCA using the HGDP as reference populations showed clear layering of the Australian, Dutch, and Midwestern American populations on the representative European HGDP cohort (Figure 4.3). The HGDP European population is comprised of samples collected from France, Italy, Italy-Bergamo, Orkney Islands, Russia, and Russia-Caucasus. At the global level, the Australian, Dutch, and Midwestern American samples showed strong distinction from African and Asian populations. Alternatively, the PCs of the cross-ethnic comparison demonstrated strong overlap between the Nigerian samples and the representative African population from HGDP. The HGDP African population was made up of samples obtained from Angola, Botswana, Central African Republic, Congo, Kenya, Lesotho, Namibia, Nigeria, Senegal, South Africa, and Sudan. The results suggest that global genetic diversity can be observed by plotting PCs and that Australian, Dutch, and Midwestern American populations show nearest genetic relatedness to European populations.

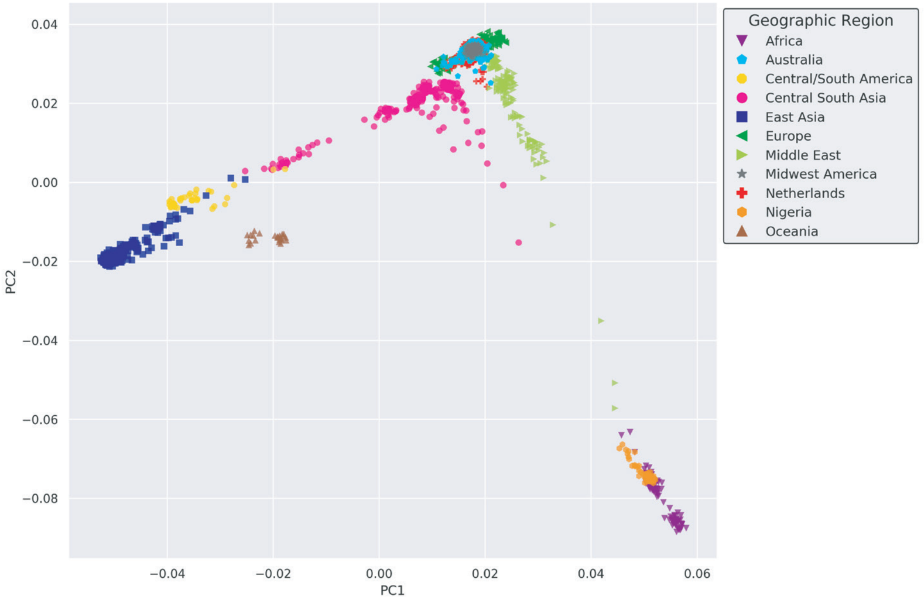


Figure 4.3 - Projection of PCs for Midwestern American, Australian, Dutch, and Nigerian subjects onto HGDP populations.
Shown are the results from the PCA using autosomal genotyped SNPs that were in common with HGDP after quality control, filtering, and exclusion of long-range LD (54,820 markers). Ancestry outliers were removed prior to performing the PCA. PC1 and PC2 represent the first and second PCs and account for 38.048% and 28.811% of the variation, respectively.

In order to provide a quantitative estimate of relationships between populations in a pairwise fashion, we used *smartpca* to sum ANOVA statistics across all eigenvectors (parameter default of 10 eigenvectors). The results of the pairwise comparisons of Australian, Dutch, and Midwestern Americans are shown in Table 4.5. Statistically significant differences were observed for each pairwise comparison, suggesting the existence of population stratification.

Table 4.5 - Statistical significance of differences between populations

Population 1	Population 2	Chi-square	P-value
Midwestern American	Netherlands	457.171	6.169×10^{-92}
Midwestern American	Australian	660.324	2.053×10^{-135}
Australian	Netherlands	7121.469	0

Note: For each pair of populations, ANOVA statistics along each eigenvector were summed across eigenvectors. Degrees of freedom are equal to 10, the default number of eigenvectors.

To put in context the genetic differences between Australian, Dutch, and Midwestern Americans, we tested for significant differences between them, the samples obtained from Nigeria, and all HGDP populations, including representative European populations (Supplementary Materials Tables 4.1 and 4.2). All comparisons between Australian, Dutch, and Midwestern American and HGDP European populations were statistically significant, with the comparison between Australian and the Orkney Islands populations being least significant. More generally, for each cohort, the statistical comparison with the Orkney Islands resulted in the least significant difference.

Case-Control GWAS

We performed two case-control GWAS between Midwestern American (cases) and Australian (controls) and Dutch (controls) populations using only common variants. A MAF filter (MAF>0.10) was employed to avoid false positives due to minor allele frequencies. The GWAS between Midwestern American (227 'cases') and Australian (1354 'controls') utilized 228,166 variants after MAF filtering. Results of the case-control GWAS between the two populations are visualized in a Manhattan plot (Figure 4.4c). Four chromosomal regions exhibited genome-wide significant differences ($p<5 \times 10^{-8}$). The dbSNP ID numbers for the significant SNPs are: rs6420020 (chromosome 5, $p=1.571 \times 10^{-15}$), rs10817415 (chromosome 9, $p=3.832 \times 10^{-15}$), rs11599284 (chromosome 10, $p=5.387 \times 10^{-14}$), rs78611721 (chromosome 20, $p=2.164 \times 10^{-14}$). The frequency of these SNPs was much greater in Australians than in American individuals. The QQ-plot (Figure 4.4d) of genome-wide p-values showed a modest deviation from the null hypothesis of no association. The overall GWAS genomic control statistic (λ) was 1.153, indicating slight inflation due to population structure, driven by a small number of polygenic variants, between the Midwestern American and Australian populations.

The second case-control GWAS was performed between Midwestern American (227 'cases') and Dutch (6139 'controls') using 228,025 common variants after MAF>0.10 filtering. The results of the case-control GWAS between the two populations are presented in the Manhattan plot in Figure 4.4a. No statistically significant variants exceeded the genome-wide significance threshold ($p<5 \times 10^{-8}$). The QQ-plot (Figure 4.4b) shows slight genomic inflation across the entire range of p-values. The GWAS genomic control statistic (λ) was 1.159, again indicating slight inflation due to population structure differences between the Midwestern American and Dutch populations.

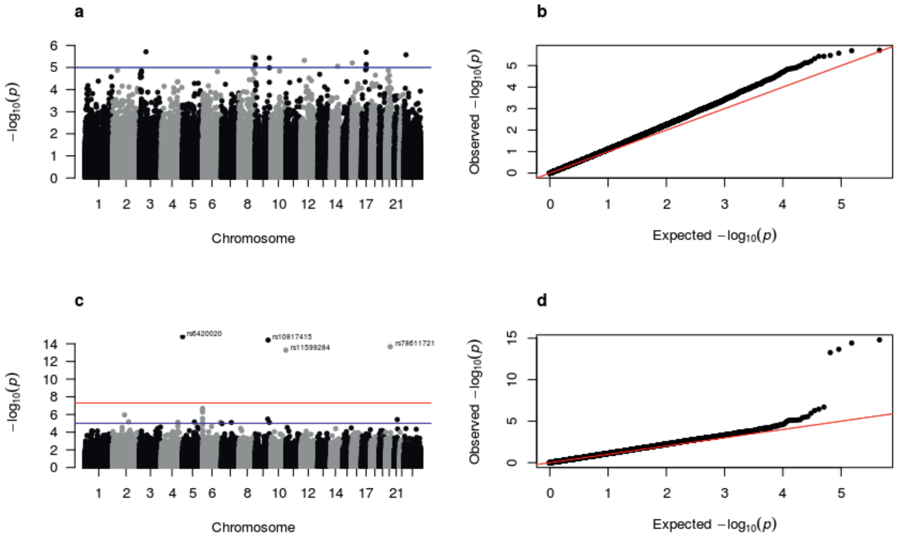


Figure 4.4 - Results of the case-control GWAS between Midwestern American (cases), Australian (controls), and Dutch (controls) populations.
(a) Manhattan plot of the case-control GWAS of Midwestern Americans (227 cases) and Dutch (6139 controls) using 228,025 variants after MAF>0.10 filter. (b) QQ plot of observed vs. expected p-values of the association results between Midwestern American and Dutch populations (λ= 1.159). (c) Manhattan plot of the case-control GWAS of Midwestern Americans (227 cases) and Australians (1581 controls) using 228,166 variants after MAF>0.10 filter. (d) QQ plot of observed vs. expected p-values of the association results between Midwestern American and Australian populations (λ= 1.153). Shown in each Manhattan plot is a blue line depicting a suggestive level of statistical significance ($p=1 \times 10^{-5}$). In panel (c), the red line represents a genome-wide level of statistical significance ($p = 5 \times 10^{-8}$). The rs numbers point to the chromosomal region that reached the genome-wide significance level. Variants with a MAF<0.10 were excluded. All related individuals and ancestry outliers were removed prior to performing the associations.

F_{ST} Estimates

F_{ST} values were calculated as a measure of genetic differentiation between populations. We generated F_{ST} values using two approaches, namely Weir and Cockerham [36] and Hudson [37] estimators. As demonstrated in 2013 by Bhatia et al., the Weir and Cockerham estimator is dependent on the ratio of sample sizes comprising each population [40]. Therefore, an alternative approach is to use the Hudson estimator, which can be implemented as a strategy independent of sample sizes, even when F_{ST} is not uniform across populations. The result produced by the Hudson estimator is a simple average of the population-specific estimators originally defined by Weir and Hill [41]. Ultimately the Hudson estimator was recommended by Bhatia et al. for estimating F_{ST} for pairs of populations with unequal sample sizes.

The F_{ST} estimates from both the Weir and Cockerham and Hudson estimators are shown in Table 4.6 and are relatively close to previously reported estimates from the HapMap consortium [42] and GoNL project (see Supplementary Table 5 of ref[16]). Regardless of the estimator, smaller F_{ST} estimates are observed between Australian, Dutch, and Midwestern American populations than between each population compared to the Nigerian cohort. Consistent with the work of Bhatia et al., it is important to note that the choice of the estimator made an impact on the resulting F_{ST} estimate.

DISCUSSION

To gain insight into the routine practice of aggregating genomic datasets from twin registers from around the world, we investigated interpopulation genetic variation with genome-wide data generated on GSA from Australian, Dutch, and Midwestern American populations. Here, we report on the inception and initial validation of a custom-designed Illumina GSA and its implementation in studying population genetic variation. Through quantitative measures and visualization of PCs, results of work presented here suggest a high level of genetic similarity between the Australian, Dutch, and Midwestern American populations, albeit with small yet statistically significant differences existing between them.

The custom-designed GSA provides a genotyping platform initially optimized for imputation containing a core imputation backbone supplemented with additional fine-mapping content to bolster imputation quality and genome-wide coverage. Also featured on the GSA are custom-selected markers specific to phenotypes of interest, notably for fertility and twinning.

Table 4.6 – F_{ST} between Midwestern American, Dutch, Australian and Nigerian populations

Comparison		F_{ST} Estimator							
		Weir and Cockerham[a]		Weir and Cockerham[b]		Hudson[a]		Hudson[b]	
Population 1	**Population 2**	**Est**	**Std. error**	**Est**	**Std. error**	**Est**	**Std. error**	**Est**	**Std. error**
Midwestern American	Netherlands	0.00017	7.40x10^{-6}	0.00017	7.35x10^{-6}	0.00017	7.45x10^{-6}	0.00018	8.19x10^{-6}
Midwestern American	Australian	0.00019	8.34x10^{-6}	0.00019	8.25x10^{-6}	0.00019	8.39x10^{-6}	0.00019	7.99x10^{-6}
Midwestern American	Nigerian	0.14362	0.00100	0.14556	0.00266	0.14424	0.00099	0.14614	0.00263
Netherlands	Australian	0.00045	8.73x10^{-6}	0.00045	2.43x10^{-5}	0.00045	8.62x10^{-6}	0.00045	2.34x10^{-5}
Netherlands	Nigerian	0.14382	0.00101	0.14585	0.00269	0.14470	0.00098	0.14683	0.00263
Australian	Nigerian	0.14347	0.00102	0.14545	0.00275	0.14463	0.00100	0.14639	0.00267

Note: The choice of the F_{ST} estimator impacts the resulting estimate. The number of SNPs used for each comparison was 564,020. Sample sizes: Midwestern American 238; Netherlands 6,139; Australian 1,448; and Nigerian sample 62.

[a]F_{ST} as estimated by popstats software. Block size for the jackknife estimator was default 5 Mb.

[b]F_{ST} as estimated by scikit-allel Python package. Block size for the jackknife estimator was 56,402 (equivalent to the number of total variants divided by 10).

With the use of the GoNL reference set and a selection of markers mimicking the GSA, we demonstrated that we can reimpute genotypes with a high degree of confidence. Exceptions are made for rare alleles (MAF<0.0001), which are never well imputed [43]. A limitation to the validation of the GSA is that we utilized the sequences of only 249 samples; therefore, the imputation and presence of alleles with a MAF<0.01 was likely less than ideal. However, by comparing the validation results of the Illumina GSA to other commercially available genotyping platforms, the imputation quality of the GSA is well in line with other genotyping products (refer to Table 1 and Table 2 in [14]). Additionally, the method of testing the coverage using two reference datasets to check concordance between genotyped and reimputed SNPs utilizes SNPs in union with both reference panels. Thus, we inherently assume that SNPs specific to a population – for example, those SNPs only appearing in GoNL – are covered and imputed in such a manner. Nevertheless, the GSA has been instrumental in generating high-quality genotype data from cohorts around the world for use in population genetic studies of complex traits.

PCs often show a remarkable correlation with geography, a manifestation of decreasing genetic similarity with increasing geographic distance. Thus, we performed PCA and visualized PCs to elucidate the degree of genetic resemblance between Australian, Dutch, and Midwestern American populations. Visualization of PCs for the three populations under study shows a high degree of overlap between PCs 1 and 2. The similarity observed between Midwestern American and Dutch populations is consistent with estimates of 4.1 million Americans (1.28% of the USA population in 2017) claiming total or partial Dutch heritage [44]. In large part, the majority of inhabitants of Midwest America have ancestral origins rooted in Northwestern Europe because of common migratory routes. Lending additional support to the Midwestern American and Dutch similarity is the fact that the majority of the Dutch Americans reside in Michigan, California, Montana, Minnesota, New York, Wisconsin, Idaho, Utah, Iowa, Ohio, West Virginia, and Pennsylvania. Together, it is apparent that there is strong Dutch influence and saturation in the Midwestern region, which is reflected in the genetic profiles of these populations.

Broad-scale comparison to diverse populations from around the world, such as those represented by the HGDP, further portrayed similarity between Australian, Dutch, and Midwestern American and with European populations more generally. The close resemblance of Australian and European populations is consistent with prior empirical results [45] and the fact that immigrants from Northern Europe colonized Australia (mainly from Britain and Ireland) and America. Incorporation of genotype data from a globally distinct population (i.e., Nigerian samples genotyped on GSA) facilitated the projection of the PCs onto the HGDP and recapitulated worldwide genetic diversity.

Quantitative measures of population similarity, as measured by summed ANOVA statistics over eigenvectors, revealed small yet statistically significant differences between Australian, Dutch, and Midwestern American populations. Additional comparisons of each population to individual HGDP European cohorts further demonstrated significant differences suggestive of population stratification. Likewise, patterns of F_{ST} estimations were consistent with the geographical clustering observed in PCA and with previous F_{ST} estimates of global population genetic differentiation. Altogether, it is likely that the observed population genetic dissimilarities are due to systematic allele frequency differences resulting from migration, adaptation, drift, and selection.

In general, large GWAS efforts aimed at discerning the genetic contributions to complex traits typically rely on meta-analyses of multiple cohorts of relatively homogeneous populations. Thus, to assess the level of homogeneity between Midwestern American, Australian, and Dutch cohorts, we performed case-control association testing between populations. Case-control GWAS of Midwestern American and Australian populations yielded results to suggest that only small genetic differences exist between the populations under study. We hesitate to interpret the few differences observed between the Australian and Midwestern American populations, although we note that it is possible that the genome-wide significant SNPs (rs6420020, rs10817415, rs78611721, and rs11599284) could be implicated in the dizygotic twinning phenotype given that the genotyped Australian group consisted of MODZT. Between Australian and Midwestern American populations, genomic inflation appeared to be primarily driven by a small number of highly polymorphic SNPs, while the remainder of the genome appears comparable. In a similar fashion, the results of the GWAS of Midwestern American and Dutch populations suggested moderate genetic differences between the two populations without genome-wide significant loci.

Twin families are major contributors of phenotype and genotype data to collaborative research initiatives. Twins are often motivated to take part because of information on their zygosity [46]. The results from work presented here are encouraging for ongoing collaborative projects, including the genetics of twinning. Collaborative efforts of the Australian, Netherlands, and other twin registers have contributed to many landmark genetic studies, including the identification of two genetic variants associated with dizygotic twinning [47]. Participation of twin cohorts and their families from geographically distinct regions such as Nigeria will undoubtedly help facilitate the elucidation of additional genetic variants underlying complex traits, including dizygotic twinning, due to the large regional differences in twinning rates [48-50].

ACKNOWLEDGEMENTS

We would like to acknowledge and thank the individuals and scientists who participated in the Genome of the Netherlands project and the Human Genome Diversity Project. We would also like to extend our gratitude to all of the enrolled participants of the twin registers represented in this study.

FINANCIAL SUPPORT

The Avera Twin Register is supported by Avera Health, Avera McKennan Hospital, and Avera Institute for Human Genetics. The collaboration between the Netherlands and the Avera Twin Register arose through NIMH Grant: 1RC2MH089995-01: Genomics of Developmental Trajectories in Twins.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICAL STANDARDS

All participants provided informed consent. This study was conducted under Institutional Review Board approval at all study sites. Netherlands Twin Register (Dutch cohort): The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance-FWA00017598; IRB/institute codes, NTR 03-180). Avera Twin Register (Midwestern American cohort): The study was approved by the Avera Institutional Review Board and the Avera Department of Human Subject's protection. Australia: The project was approved by the Queensland Institute of Medical Research Human Research Ethics Committee.

REFERENCES

1. Evangelou, E., et al., *Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits.* Nat Genet, 2018. **50**(10): p. 1412-1425.
2. Wray, N.R., et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nat Genet, 2018. **50**(5): p. 668-681.
3. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
4. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.
5. Tam, V., et al., *Benefits and limitations of genome-wide association studies.* Nat Rev Genet, 2019.
6. Silventoinen, K., et al., *Genetic and environmental effects on body mass index from infancy to the onset of adulthood: an individual-based pooled analysis of 45 twin cohorts participating in the COllaborative project of Development of Anthropometrical measures in Twins (CODATwins) study.* Am J Clin Nutr, 2016. **104**(2): p. 371-9.
7. Silventoinen, K., et al., *The CODATwins Project: The Cohort Description of Collaborative Project of Development of Anthropometrical Measures in Twins to Study Macro-Environmental Variation in Genetic and Environmental Effects on Anthropometric Traits.* Twin Res Hum Genet, 2015. **18**(4): p. 348-60.
8. Sullivan, P.F., et al., *Empirical evaluation of the genetic similarity of samples from twin registries in Australia and the Netherlands using 359 STRP markers.* Twin Res Hum Genet, 2006. **9**(4): p. 600-2.
9. Boomsma, D.I., et al., *Netherlands Twin Register: a focus on longitudinal research.* Twin Res, 2002. **5**(5): p. 401-6.
10. Boomsma, D.I., et al., *Netherlands Twin Register: from twins to twin families.* Twin Res Hum Genet, 2006. **9**(6): p. 849-57.
11. Willemsen, G., et al., *The Netherlands Twin Register biobank: a resource for genetic epidemiological studies.* Twin Res Hum Genet, 2010. **13**(3): p. 231-45.
12. Willemsen, G., et al., *The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection.* Twin Res Hum Genet, 2013. **16**(1): p. 271-81.
13. Kittelsrud, J., et al., *Establishment of the Avera Twin Register in the Midwest USA.* Twin Res Hum Genet, 2017. **20**(5): p. 414-418.
14. Ehli, E.A., et al., *A method to customize population-specific arrays for genome-wide association testing.* Eur J Hum Genet, 2017. **25**(2): p. 267-270.
15. Boomsma, D.I., et al., *The Genome of the Netherlands: design, and project goals.* Eur J Hum Genet, 2014. **22**(2): p. 221-7.
16. Genome of the Netherlands, C., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet, 2014. **46**(8): p. 818-25.
17. Cann, H.M., et al., *A human genome diversity cell line panel.* Science, 2002. **296**(5566): p. 261-2.
18. Hopper, J.L., et al., *Australian Twin Registry: a nationally funded resource for medical and scientific research, incorporating match and WATCH.* Twin Res Hum Genet, 2006. **9**(6): p. 707-11.
19. Hopper, J.L., *The Australian Twin Registry.* Twin Res, 2002. **5**(5): p. 329-36.

20. Hur, Y.M., et al., *The Nigerian Twin and Sibling Registry.* Twin Res Hum Genet, 2013. **16**(1): p. 282-4.
21. Painter, J.N., et al., *A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins.* Hum Reprod, 2010. **25**(6): p. 1569-80.
22. Min, J.L., et al., *High microsatellite and SNP genotyping success rates established in a large number of genomic DNA samples extracted from mouth swabs and genotypes.* Twin Res Hum Genet, 2006. **9**(4): p. 501-6.
23. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
24. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2011. **9**(2): p. 179-81.
25. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
26. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
27. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
28. Manichaikul, A., et al., *Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis.* PLoS Genet, 2012. **8**(4): p. e1002640.
29. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
30. Abdellaoui, A., et al., *Population structure, migration, and diversifying selection in the Netherlands.* European journal of human genetics : EJHG, 2013. **21**(11): p. 1277-85.
31. Price, A.L., et al., *Long-range LD can confound genome scans in admixed populations.* Am J Hum Genet, 2008. **83**(1): p. 132-5; author reply 135-9.
32. Rosenberg, N.A., *Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives.* Ann Hum Genet, 2006. **70**(Pt 6): p. 841-7.
33. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
34. Haeussler, M., et al., *The UCSC Genome Browser database: 2019 update.* Nucleic Acids Res, 2019. **47**(D1): p. D853-D858.
35. R Core Team, *R: A Language and Environment for Statistical Computing.* 2018.
36. Weir, B.S. and C.C. Cockerham, *Estimating F-Statistics for the Analysis of Population Structure.* Evolution, 1984. **38**(6): p. 1358-1370.
37. Hudson, R.R., M. Slatkin, and W.P. Maddison, *Estimation of levels of gene flow from DNA sequence data.* Genetics, 1992. **132**(2): p. 583-9.
38. Skoglund, P., et al., *Genetic evidence for two founding populations of the Americas.* Nature, 2015. **525**(7567): p. 104-8.
39. Miles, A., Harding, N., *scikit-allel - Explore and analyze genetic variation. 1.2.0 edn.* Github, 2018.
40. Bhatia, G., et al., *Estimating and interpreting FST: the impact of rare variants.* Genome Res, 2013. **23**(9): p. 1514-21.
41. Weir, B.S. and W.G. Hill, *Estimating F-statistics.* Annu Rev Genet, 2002. **36**: p. 721-50.
42. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.

43. Zheng, H.F., et al., *Performance of genotype imputation for low frequency and rare variants from the 1000 genomes.* PLoS One, 2015. **10**(1): p. e0116487.
44. Data Access and Dissemination Systems. *U.S. Census Bureau, 2013-2017 American Community Survey 5-Year Estimates*. 2017 [cited 2019; Available from: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_13_5YR_B04006&prodType=table.
45. Stankovich, J., et al., *On the utility of data from the International HapMap Project for Australian association studies.* Hum Genet, 2006. **119**(1-2): p. 220-2.
46. Odintsova, V.V., et al., *Establishing a Twin Register: An Invaluable Resource for (Behavior) Genetic, Epidemiological, Biomarker, and 'Omics' Studies.* Twin Res Hum Genet, 2018. **21**(3): p. 239-252.
47. Mbarek, H., et al., *Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility.* Am J Hum Genet, 2016. **98**(5): p. 898-908.
48. Hall, J.G., *Twinning.* Lancet, 2003. **362**(9385): p. 735-43.
49. Hoekstra, C., et al., *Dizygotic twinning.* Hum Reprod Update, 2008. **14**(1): p. 37-47.
50. Smits, J. and C. Monden, *Twinning across the Developing World.* PLoS One, 2011. **6**(9): p. e25239.

SUPPLEMENTARY MATERIALS

Supplementary Table 4.1 - Population codes and sampling information for HGDP and study populations

Population Code	Population Name	Sampling Location	Geographic Region Of Population
3	MidwestAmerican	MidwestAmerica	MIDWESTAMERICA
4	NTR	Netherlands	NETHERLANDS
5	Australian	Australia	AUSTRALIA
6	Nigerian	Nigeria	NIGERIA
20	Orcadian	OrkneyIslands	EUROPE
21	Adygei	Russia-Caucasus	EUROPE
22	Russian	Russia	EUROPE
24	Basque	France	EUROPE
25	French	France	EUROPE
27	Italian	Italy-Bergamo	EUROPE
28	Sardinian	Italy	EUROPE
29	Tuscan	Italy	EUROPE
34	Mozabite	Algeria-Mzab	MIDDLE_EAST
36	Bedouin	Israel-Negev	MIDDLE_EAST
37	Druze	Israel-Carmel	MIDDLE_EAST
38	Palestinian	Israel-Central	MIDDLE_EAST
50	Balochi	Pakistan	CENTRAL_SOUTH_ASIA
51	Brahui	Pakistan	CENTRAL_SOUTH_ASIA
52	Burusho	Pakistan	CENTRAL_SOUTH_ASIA
54	Hazara	Pakistan	CENTRAL_SOUTH_ASIA
56	Kalash	Pakistan	CENTRAL_SOUTH_ASIA
57	Makrani	Pakistan	CENTRAL_SOUTH_ASIA
58	Pathan	Pakistan	CENTRAL_SOUTH_ASIA
59	Sindhi	Pakistan	CENTRAL_SOUTH_ASIA
71	Melanesian	Bougainville	OCEANIA
75	Papuan	NewGuinea	OCEANIA
81	Colombian	Colombia	AMERICA
82	Karitiana	Brazil	AMERICA
83	Surui	Brazil	AMERICA
86	Maya	Mexico	AMERICA
87	Pima	Mexico	AMERICA

Supplementary Table 4.1 (continued)

Population Code	Population Name	Sampling Location	Geographic Region Of Population
430	BantuSouthAfrica	Angola	AFRICA
430	BantuSouthAfrica	BotswanaOrNamibia	AFRICA
430	BantuSouthAfrica	Lesotho	AFRICA
430	BantuSouthAfrica	SouthAfrica	AFRICA
441	BantuKenya	Kenya	AFRICA
457	Nilote	Sudan	AFRICA
464	Mandenka	Senegal	AFRICA
465	Yoruba	Nigeria	AFRICA
488	BiakaPygmy	CentralAfricanRepublic	AFRICA
489	MbutiPygmy	Congo	AFRICA
494	San	Namibia	AFRICA
601	Han	China	EAST_ASIA
602	Han-NChina	China	EAST_ASIA
606	Dai	China	EAST_ASIA
607	Daur	China	EAST_ASIA
608	Hezhen	China	EAST_ASIA
611	Lahu	China	EAST_ASIA
612	Miao	China	EAST_ASIA
613	Oroqen	China	EAST_ASIA
615	She	China	EAST_ASIA
616	Tujia	China	EAST_ASIA
617	Tu	China	EAST_ASIA
618	Xibo	China	EAST_ASIA
619	Yi	China	EAST_ASIA
622	Mongola	China	EAST_ASIA
625	Naxi	China	EAST_ASIA
629	Uygur	China	CENTRAL_SOUTH_ASIA
677	Cambodian	Cambodia	EAST_ASIA
684	Japanese	Japan	EAST_ASIA
699	Yakut	Siberia	EAST_ASIA

Supplementary Table 4.2 - Statistical significance of differences between populations. For each pair of populations, the ANOVA statistics are summed across eigenvectors. The result is approximately chisq with degrees of freedom equal to the number of eigen vectors.

pop1	pop2	chisq	p-value
3	4	749.816	1.26E-154
3	6	6156.44	0
5	3	429.224	5.62E-86
5	4	8857.501	0
5	6	22000.876	0
6	4	67639.754	0
20	3	63.894	6.59E-10
20	4	173.691	4.77E-32
20	5	20.734	0.0230255
20	6	1547.751	0
21	3	1852.742	0
21	4	6332.92	0
21	5	3420.913	0
21	6	2075.942	0
22	3	1460.048	1.07E-307
22	4	3904.679	0
22	5	2256.011	0
22	6	2150.141	0
24	3	1254.616	2.38E-263
24	4	2703.947	0
24	5	1662.729	0
24	6	2225.43	0
25	3	463.033	3.46E-93
25	4	1022.616	2.51E-213
25	5	438.795	5.12E-88
25	6	2132.711	0
27	3	876.529	7.16E-182
27	4	1597.523	0
27	5	1095.659	4.55E-229
27	6	1645.343	0
28	3	2907.991	0
28	4	9829.624	0
28	5	5965.776	0
28	6	2745.004	0
29	3	822.826	2.55E-170
29	4	1440.327	1.95E-303
29	5	983.13	8.05E-205
29	6	1455.056	1.28E-306
34	3	4139.525	0
34	4	36146.567	0
34	5	14510.285	0
34	6	2149.688	0
36	3	4010.109	0
36	4	33415.963	0
36	5	14381.805	0
36	6	2129.977	0
37	3	4597.353	0
37	4	20079.444	0
37	5	10717.317	0
37	6	2979.874	0
38	3	4779.207	0
38	4	24397.496	0
38	5	12236.738	0
38	6	2860.238	0
50	3	3860.641	0
50	4	22453.506	0
50	5	10992.391	0
50	6	2305.417	0
51	3	3975.99	0
51	4	23144.328	0
51	5	11360.598	0
51	6	2426.446	0
52	3	4479.567	0
52	4	26946.201	0
52	5	12402.601	0
52	6	2344.274	0
54	3	3076.783	0
54	4	25022.76	0
54	5	9306.791	0
54	6	1954.409	0
56	3	5192.329	0
56	4	31342.826	0
56	5	14830.511	0
56	6	2486.114	0

Supplementary Table 4.2 (continued)

Supplementary Table 4.2 (continued)

pop1	pop2	chisq	p-value
57	3	3335.635	0
57	4	20108.585	0
57	5	9648.131	0
57	6	2109.805	0
58	3	3950.296	0
58	4	20866.105	0
58	5	10379.845	0
58	6	2219.71	0
59	3	4167.409	0
59	4	24455.961	0
59	5	11762.972	0
59	6	2249.503	0
71	3	5364.411	0
71	4	44101.355	0
71	5	16727.68	0
71	6	2239.247	0
75	3	6307.217	0
75	4	68348.873	0
75	5	22851.259	0
75	6	2411.932	0
81	3	4503.517	0
81	4	38298.986	0
81	5	14528.519	0
81	6	1764.309	0
82	3	4622.311	0
82	4	36008.124	0
82	5	14049.373	0
82	6	1813.536	0
83	3	3780.704	0
83	4	24068.937	0
83	5	10410.769	0
83	6	1466.692	0.00E+00
86	3	5221.223	0
86	4	59233.304	0
86	5	19676.868	0
86	6	2246.835	0
87	3	5026.309	0
87	4	44523.395	0

Supplementary Table 4.2 (continued)

pop1	pop2	chisq	p-value
87	5	16322.743	0
87	6	1972.15	0
430	3	2964.017	0
430	4	33881.662	0
430	5	10520.31	0
430	6	302.222	5.27E-59
441	3	2899.047	0
441	4	33603.643	0
441	5	10584.149	0
441	6	335.485	4.78E-66
464	3	4456.269	0
464	4	46124.527	0
464	5	15217.264	0
464	6	73.328	1.01E-11
465	3	4810.955	0
465	4	49644.327	0
465	5	16405.907	0
465	6	22.846	0.0113302
488	3	5880.655	0
488	4	58992.567	0
488	5	20445.876	0
488	6	1311.773	1.10E-275
489	3	5613.946	0
489	4	54929.362	0
489	5	19602.729	0
489	6	1298.928	6.52E-273
494	3	4513.962	0
494	4	37823.504	0
494	5	14633.924	0
494	6	1037.472	1.58E-216
601	3	5536.788	0
601	4	56925.041	0
601	5	19025.951	0
601	6	2317.258	0
602	3	3685.594	0
602	4	33528.694	0
602	5	11589.288	0
602	6	1388.798	2.60E-292

Supplementary Table 4.2 (continued)

pop1	pop2	chisq	p-value
606	3	4229.327	0
606	4	37671.301	0
606	5	13638.214	0
606	6	1701.07	0
607	3	3925.066	0
607	4	32743.208	0
607	5	11813.288	0
607	6	1561.917	0
608	3	3448.338	0
608	4	29294.408	0
608	5	10473.989	0
608	6	1264.536	1.72E-265
611	3	3504.831	0
611	4	28452.012	0
611	5	10623.786	0
611	6	1354.066	8.19E-285
612	3	4160.859	0
612	4	36172.518	0
612	5	13036.093	0
612	6	1560.632	0
613	3	3785.704	0
613	4	34250.576	0
613	5	12199.557	0
613	6	1447.572	5.30E-305
615	3	4078.076	0
615	4	34889.792	0
615	5	12620.541	0
615	6	1486.517	2.05866e-313
616	3	4036.471	0
616	4	35419.931	0
616	5	12586.382	0
616	6	1491.838	1.46004e-314
617	3	3288.939	0
617	4	31307.943	0
617	5	10636.539	0
617	6	1243.201	6.90E-261
618	3	3232.679	0
618	4	30324.36	0

Supplementary Table 4.2 (continued)

pop1	pop2	chisq	p-value
618	5	10469.472	0
618	6	1330.962	7.95E-280
619	3	3582.878	0
619	4	33312.243	0
619	5	11452.352	0
619	6	1337.158	3.65E-281
622	3	3293.957	0
622	4	31949.858	0
622	5	10962.64	0
622	6	1311.818	1.08E-275
625	3	3257.196	0
625	4	28526.449	0
625	5	10033.39	0
625	6	1184.153	3.77E-248
629	3	2447.635	0
629	4	19025.266	0
629	5	7116.367	0
629	6	1571.626	0
677	3	3738.917	0
677	4	33689.003	0
677	5	12256.31	0
677	6	1595.084	0
684	3	4778.618	0
684	4	48812.379	0
684	5	15973.221	0
684	6	1998.414	0
699	3	4371.196	0
699	4	51285.367	0
699	5	16835.329	0
699	6	2275.401	0

5

GENETIC META-ANALYSIS OF TWIN BIRTH WEIGHT SHOWS HIGH GENETIC CORRELATION WITH SINGLETON BIRTH WEIGHT

Published as
Beck, J.J., Pool, R., van de Weijer, M., Chen, X., Krapohl, E., Gordon, S.D., Nygaard, M., Debrabant, B., Palviainen, T., van der Zee, M.D., Baselmans, B., Finnicum, C.T., Yi, L., Lundström, S., van Beijsterveldt, T., Christiansen, L., Heikkilä, K., Loukola, A., Ollikainen, M., Christensen, K., Martin, N.G., Plomin, R., Nivard, M., Bartels, M., Dolan, C., Willemsen, G., de Geus, E., Almqvist, C., Magnuson, P.K.E., Mbarek, H., Ehli, E.A., Boomsma, D.I., Hottenga, J.J. (2021)
Genetic Meta-Analysis of Twin Birth Weight Shows High Genetic Correlation with Singleton Birth Weight. Human Molecular Genetics, **30**, 1894-1905.

ABSTRACT

Birth weight (BW) is an important predictor of newborn survival and health and has associations with many adult health outcomes, including cardiometabolic disorders, autoimmune diseases, and mental health. On average, twins have a lower BW than singletons as a result of a different pattern of fetal growth and shorter gestational duration. Therefore, investigations into the genetics of BW often exclude data from twins, leading to a reduction in sample size and remaining ambiguities concerning the genetic contribution to BW in twins. In this study, we carried out a genome-wide association meta-analysis of BW in 42212 twin individuals and found a positive correlation of beta values (Pearson's *r* = 0.66, 95% confidence interval [CI]: 0.47–0.77) with 150 previously reported genome-wide significant variants for singleton BW. We identified strong positive genetic correlations between BW in twins and numerous anthropometric traits, most notably with BW in singletons (genetic correlation [r_g] = 0.92, 95% CI: 0.66–1.18). Genetic correlations of BW in twins with a series of health-related traits closely resembled those previously observed for BW in singletons. Polygenic scores constructed from a genome-wide association study on BW in the UK Biobank demonstrated strong predictive power in a target sample of Dutch twins and singletons. Together, our results indicate that a similar genetic architecture underlies BW in twins and singletons and that future genome-wide studies might benefit from including data from large twin registers.

Keywords: Birth weight, genome, twins, genetics, genome-wide association study, biobanks

INTRODUCTION

Birth weight (BW) is a powerful predictor of infant and newborn survival, with lower-weight infants being at higher risk of mortality [1-3]. BW is also associated with a wide array of health-related variables in later life [4], with varying effect sizes, including adult body mass index (BMI) [5, 6], cardiovascular disease [7, 8], type 2 diabetes [9], hypertension [10-12] and psychological distress [13]. Our knowledge of the biological pathways underlying BW is growing with the rapidly increasing number of genetic variants identified in genome-wide association (GWA) studies. Yet, these investigations mainly focus on BW in singletons and tend to exclude data from twins in the discovery analysis. Therefore, knowledge about the genetic overlap between BW in singletons and twins is limited, and it is not clear to what degree findings in singletons can be generalized to twins and to what extent data from twins can contribute to gene discovery for BW. This knowledge would be useful as a considerable genetic overlap would indicate that data from singletons and twins could be combined for attaining larger sample sizes.

BW is a complex and multifactorial trait [14, 15]. Maternal and fetal genomes conjointly determine fetal size, making estimations of the heritability of BW challenging as offspring and maternal genomes are not independent. In twins, BW is different from BW in singleton births because of their lower gestational age. The main factor explaining lower gestational age is uterine overdistension [16]. Still, twin and family studies suggest similar heritability estimates for BW, ranging from 10% to 40% [17-20], indicating a moderate contribution of genetic factors to BW variation. Of interest for our quest is a study from the Netherlands in which heritability was estimated from data on parents and their singleton offspring and from data on mono- and dizygotic twins [19]. The heritability estimates for BW and height were all around 0.3 and highly comparable in both groups.

The number of genetic variants identified for BW is growing based on findings from GWA studies (GWAS). In a 2010 study by Freathy et al. [21], two variants, in *ADCY5* and near *CCNL1*, were found to influence variation in BW in singletons. The number of associated variants increased to seven in 2013 with an expanded meta-analysis study of over 69000 European individuals [22]. In a multi-ancestry GWA meta-analysis (GWAMA) by Horikoshi and colleagues [23], BW and genotype data were collected for 153781 singletons. The result of this effort was the identification of 59 independent signals, capturing approximately 15% of the variance in BW. Beaumont and colleagues [24] also examined the contribution of fetal versus maternal genetic effects and identified ten maternal loci influencing offspring birthweight. Additional GWA efforts have been undertaken to ascertain the maternal and fetal genetic effects on BW and their relation to cardiometabolic risk, in which 190 independent associations were discovered [25]. To date, only one GWA study has been performed on BW in twins (4593 female twins from the UK), which identified one variant on chromosome 9, close to the *NTRK2* gene [26].

The Developmental Origins of Health and Disease (DOHaD) hypothesis is based on observations that adverse influences early in development, particularly in the intrauterine environment, result in permanent physiological and metabolomic changes leading to increased risk of disease in adulthood [27-29]. One hypothesis, postulated by Barker in the 1990s, proposed that intrauterine growth restriction, low BW, and premature birth have a causal relationship to hypertension, coronary heart disease, and non-insulin-dependent diabetes in later life. Barker and colleagues traced infant mortality rates in England during the early 1900s and found strong geographical relations between infant death and high rates of mortality resulting from coronary heart disease years later [27]. They postulated that the geographic associations of infant mortality and adult death rates 'reflects variations in nutrition in early life, which are expressed pathologically on exposure to later dietary influences' (p.1081). At the time, the typical certified cause of death in newborn babies was low BW. Thus, the hypothesis was that low BW babies surviving infancy suffered from fetal undernutrition, exhibiting non-communicable changes in metabolism and physiology, in turn increasing coronary heart disease risk in adulthood [30]. Low BW can serve as a proxy for a suboptimal intrauterine environment and is not only associated with cardiovascular disease [31] but also with respiratory disease [32], various psychiatric disorders [33], as well as mental health, cognitive and socioeconomic outcomes [34].

In general, the DOHaD and the Barker hypotheses are environmentally based. That is, the existence of an adverse intrauterine environment leads to decreased BW and long-term cardiometabolic sequelae in offspring. Alternatively, strong genetic correlations between low singleton BW and indicators of metabolic and cardiovascular health, as described in the meta-analysis by Horikoshi and colleagues [23], correspond more closely to the Fetal Insulin Hypothesis [35]. In this context, the correlations between BW and cardiometabolic disorders are driven by the transmission of maternal genes to the offspring. However, genetic correlations between BW and the cardiometabolic traits could be driven through the fetal and/or the maternal genome. The latter is broadly consistent with the DOHaD/Barker hypothesis since the maternal genome defines the intrauterine environment, whereas the former more likely reflects mechanisms of the Fetal Insulin Hypothesis [36]. Recent studies have investigated these differences in hopes of disentangling the relative contributions of fetal and maternal effects on BW and later life cardiometabolic disease [25, 37].

On average, twins have lower BW than singletons since twin pregnancy is characterized by a shorter gestational duration [16] and because fetal growth slows down after approximately 32 weeks of gestation [38-41]. Therefore, investigations into the genetic architecture of BW and other birth-related characteristics often exclude twins, even though this may lead to a significant decrease in sample size. Concerning the DOHaD hypothesis, there is no evidence that the relation between BW and later-life disease differs between twins and singletons as demonstrated for blood pressure or anti-hypertensive drug use [42-44] and diabetes [45, 46].

This study aimed to search for common genetic variants underlying BW in twins by carrying out a meta-analysis of genetic association studies in twins and comparing the results to those for BW in singletons. To this end, four approaches were employed: 1) A meta-analysis of combined GWA results from five European twin cohorts, UK Biobank, one Australian twin cohort, and one twin cohort from the Midwestern region of the United States of America. 2) An assessment of the genetic correlations between BW in twins and BW in singletons. 3) The evaluation of the genetic correlations between BW in twins and a range of traits and diseases in later life, including anthropometric and neuropsychiatric characteristics. 4) An assessment of the predictive performance of BW polygenic scores in twins and singletons.

RESULTS

Meta-analysis

We carried out a GWAMA for BW in 42212 twins. The meta-analysis QQ-plot, showing the expected distribution of genome-wide *P*-values compared to the observed values across SNPs, can be found in Supplementary Material, Figure 5.1. The Manhattan plot for the meta-analysis is shown in Figure 5.1. There were no genome-wide significant SNPs at the defined minimum *P*-value for lead SNPs ($P<5x10^{-8}$); however, two lead SNPs had an association signal of $P<5x10^{-7}$. These SNPs were located on chromosome 1 (rs10800682, hg19 position 1:200198946, $P=2.92x10^{-7}$) and chromosome 3 (rs3845913, hg19 position 3:123100606, $P=2.93x10^{-7}$). rs10800682 is independent (>12Mb, EUR $r^2<0.05$) of all genome-wide significant loci found by Horikoshi and colleagues [23]. rs3845913 is an intronic variant of *ADCY5* and is ~31kb downstream of rs11719201 (EUR r^2 0.154), one of 60 loci previously associated with BW [23].

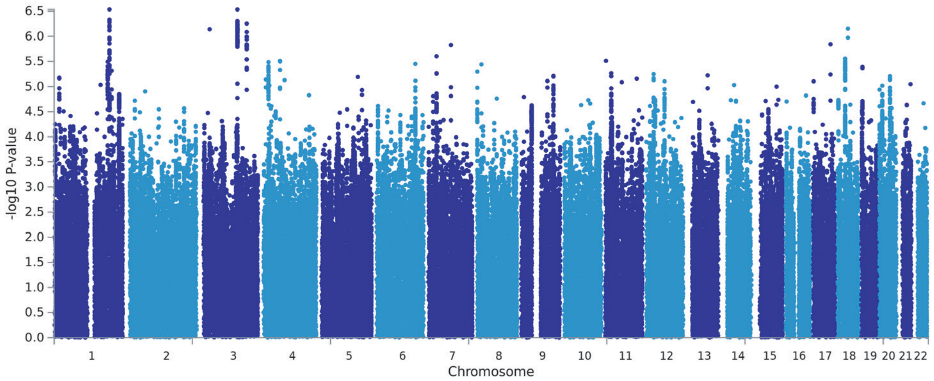


Figure 5.1 - Manhattan plot from the genome-wide association meta-analysis for BW. The association P-value (on -log$_{10}$ scale) for each of up to 7,692,335 SNPs (y-axis) is plotted against the genomic position according to NCBI Build 37 (x-axis). For plotting purposes, overlapping data points are not drawn for filtered SNPs with a P-value≥1x10^{-5}.

Replication of previous association results

Though no genome-wide significant SNPs were identified, we evaluated the performance of SNPs in the current study with the genome-wide significant SNPs signals (*P*<6.6x10^{-9}) recently identified by Warrington et al., 2019 [25] in a GWAS of own BW. Of the significant SNPs, 150 overlapped with the current study after retention of markers present in greater than 70% of all study participants. As shown in Figure 5.2, following the alignment of effect alleles, the beta estimates between overlapping markers are highly correlated (Pearson's r=0.66, 95% CI: 0.47-0.77). Summary statistics of the 150 overlapping variants are presented in Supplementary Material, Table 5.1. Overall, the positive linear relationship indicates that the previously reported significant variants behave in a similar fashion between singletons and twins.

Additionally, since gestational age was not available in all cohorts, we assessed heterogeneity of the overlapping SNPs mentioned above (i.e., 150) using METAL (implemented as Cochran's Q-test). No significant heterogeneity in allelic effects was observed after Bonferroni correction (*P*>0.00033). The smallest reported *P*-value of heterogeneity statistics in the current study was 0.002, which is in line with the smallest reported *P*-value of the genome-wide significant variants reported in Warrington et al., 2019 of 0.004 (Supplementary Material, Table 5.1).

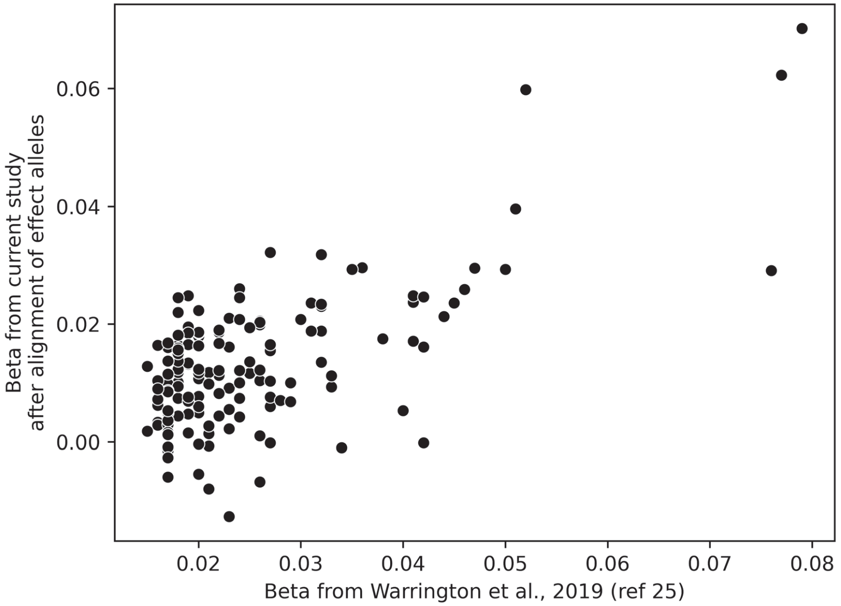


Figure 5.2 - Scatter plot of the beta estimates from the overlapping SNPs between the current study and those reported in Warrington et al., 2019 (25) for the GWAS on own BW (P<6.6x10^{-9}).
Of the significant SNPs, 150 overlapped with the current study.

Genetic correlations

The results from the genetic correlation analyses of BW in twins can be found in Figure 5.3 and Supplementary Material, Table 5.2. In general, the strongest genetic correlations were with anthropometric traits, specifically BW-related phenotypes. Previous studies have investigated and attempted to partition maternal and fetal genetic effects on BW, allowing for comparisons to individual and parental effects in this study.

The strongest genetic correlation was with 'child birth weight' (i.e., the individual's own genetic effect on their BW) (genetic correlation [r_g]=0.98, 95% confidence interval [CI]: 0.62-1.33) based on a discovery GWAS of 26836 European individuals [22]. Similarly, robust positive correlations were found with other phenotypes of the individuals own genetic effect on their BW, including UK Biobank birth weight (data field 20022) (r_g=0.95, 95% CI: 0.71-1.19), 'own birth weight' (r_g=0.92, 95% CI: 0.66-1.18) derived from an expanded GWAS of 286870 European individuals [25], and 'birth weight' (r_g= 0.91, 95% CI: 0.65-1.17) in 143677 European individuals [23]. It is important to note that genetic correlations referenced above are from three studies that are not entirely independent. Sequential studies (in chronological order, references [22, 23, 25]) used a core set of samples obtained by the Early Growth Genetics Consortium (EGG), which were expanded upon with new releases of the UK Biobank.

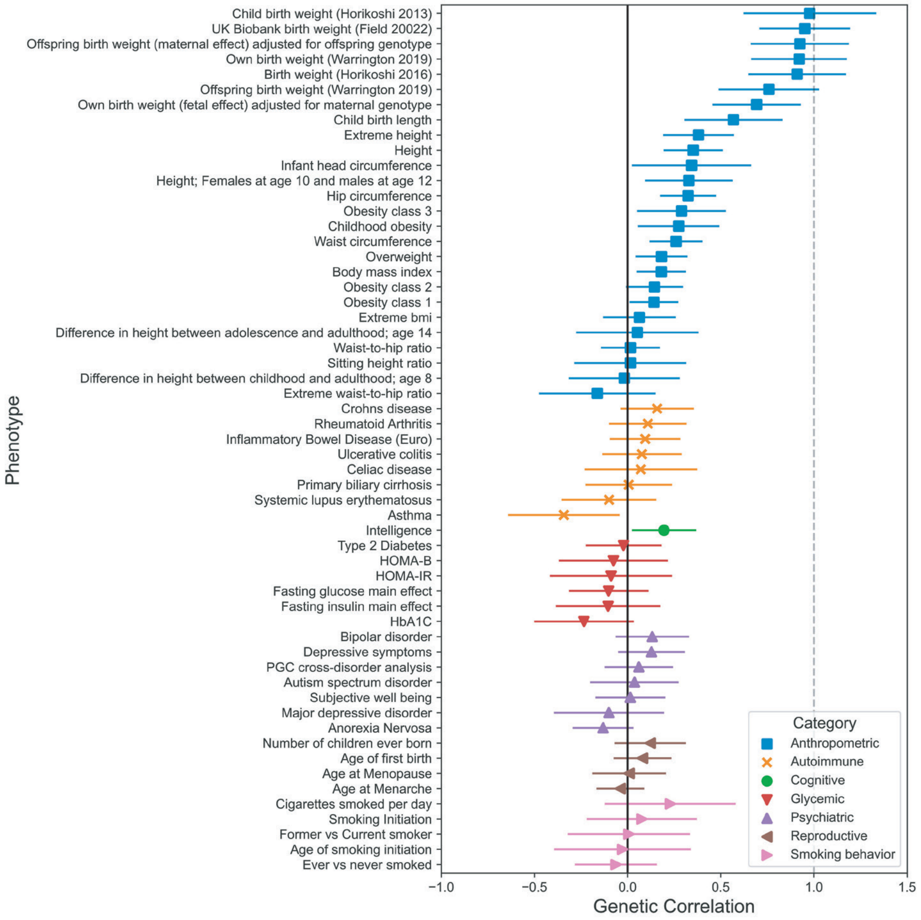


Figure 5.3 - Genetic relationships between BW in twins and 57 other phenotypes. SNP-based genetic correlations (r_g) between BW in twins and a range of other traits and diseases using LD Score regression. The bars represent 95% confidence intervals. The genetic correlation estimates are color-coded according to their respective category. HbA1C=hemoglobinA1C, HOMA-IR=homeostatic model assessment of insulin resistance, HOMA-B=homeostatic model assessment of beta cell function, PGC=Psychiatric Genetics Consortium, BMI=body mass index. PubMed reference numbers (PMID) for each trait are listed in Supplementary Material, Table 5.2.

A positive correlation was also observed with 'offspring birth weight' (i.e., the maternal genetic effect on offspring BW), as measured in 216611 mothers [25] (r_g=0.76, 95% CI: 0.49-1.03). Of the genetic correlations with other phenotypes, six additional anthropometric traits exhibited strong positive genetic correlations, including offspring birth weight (maternal genetic effect on offspring BW after adjusting for the correlated offspring's genotype) (r_g=0.92, 95% CI: 0.66-1.19), own birth weight (individuals own genetic effect on their own BW after adjusting for the correlated maternal genotype) (r_g=0.69, 95% CI: 0.45-0.93), child birth length (r_g=0.57, 95% CI: 0.30-0.83), extreme height (r_g=0.38, 95% CI: 0.19-0.57), height (r_g=0.35, 95% CI: 0.19-0.51), and hip circumference (r_g=0.32, 95% CI: 0.17-0.47).

Glycemic traits were all negatively associated with BW, whereas cognitive characteristics, measured by intelligence, correlated positively (r_g=0.20, 95% CI: 0.02-0.37). Genetic correlations of BW in twins with autoimmune disorders, psychiatric disorders, reproductive traits, and smoking behavior yielded mixed results.

The SNP heritability (h^2) was calculated using LD Score regression. The h^2 was estimated to be 0.0407 for BW in twins. For BW in singletons, the heritability estimates from three studies were h^2=0.1139, h^2=0.0985, and h^2=0.1016 for 'child birth weight' [22], 'own birth weight' [25], and 'birth weight' [23], respectively. The heritability estimate of UK Biobank birth weight was h^2=0.1006.

PolyGenic Score prediction

The PGS, based on summary statistics from GWA analyses of BW in UK Biobank, robustly predicted BW in NTR twins and singletons. The PGS, including the fraction of SNPs with a *P*-value selection threshold of 0.01, was the best predictor for BW in twins (β=68.19, p=2.10x10^{-51}, PGS R^2=0.02) and singletons (β=108.18, p=6.94x10^{-57}, PGS R^2=0.03), as shown in Table 5.1.

Table 5.1 - Results of the PGS prediction in NTR twins and singletons

	Twins (N=10487)				**Singletons (N=6892)**			
Prop	β_{PGS}	SE_{PGS}	P_{PGS}	PGS R^2	β_{PGS}	SE_{PGS}	P_{PGS}	PGS R^2
0.001	18.89	4.66	5.04E-05	0.00	34.28	6.87	6.09E-07	0.00
0.003	19.94	4.57	1.26E-05	0.00	38.67	6.72	8.77E-09	0.00
0.005	54.19	4.72	1.86E-30	0.01	75.01	6.75	1.13E-28	0.02
0.01	68.19	4.52	2.10E-51	0.02	108.18	6.81	6.94E-57	0.03
0.05	60.35	4.50	5.39E-41	0.01	101.71	6.88	1.73E-49	0.03
0.1	58.48	4.50	1.26E-38	0.01	99.71	6.88	1.52E-47	0.03
0.2	57.25	4.50	4.46E-37	0.01	98.45	6.89	2.24E-46	0.03
0.3	56.83	4.50	1.43E-36	0.01	98.10	6.89	4.96E-46	0.03
0.5	56.53	4.50	3.23E-36	0.01	97.77	6.89	1.01E-45	0.03
INF	55.39	4.53	2.09E-34	0.01	90.48	6.98	1.92E-38	0.02

Note: Prop (proportion) is the P-value threshold for SNP inclusion in the polygenic score (PGS), β is the regression coefficient for each term with standard error (SE) and P-value (P). PGS R^2 is the phenotypic variance explained by the PGS.

As shown in Figure 5.4A, a comb-like distribution of raw BW was observed in singletons, corresponding to even ~500g increments, reflecting the assessment of BW in this group.

BW category was also evaluated as the response variable (histograms in Figure 5.4B). The evaluation was done in all target samples (twins and singletons) by including twin status and interaction of PGS and twin status as predictors in the model (Table 5.2). As before, the PGS, including the proportion of SNPs with a *P*-value selection threshold of 0.01, represented the best predictor of BW category (β=0.18, p=1.68x10^{-49}, PGS R^2=0.02). Together, the results of PGS prediction analyses suggest that BW PGS constructed from a large representative discovery population predict BW similarly in a target population of twins and singletons.

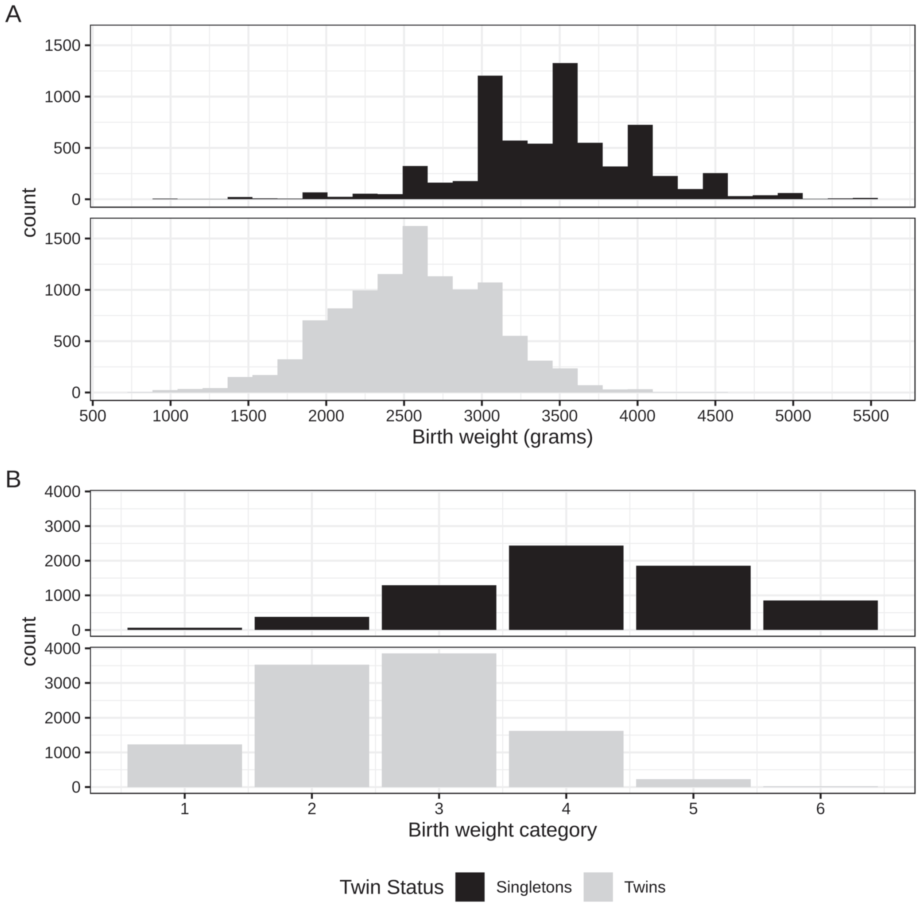


Figure 5.4 - Histograms of raw and categorical BW for NTR twins and singletons. Panel A shows histograms for raw BW in grams. Panel B portrays the distributions for BW categories 1-6 as described in the text. N=10487 twins; 6892 singletons. It is of note to point out the peaks corresponding to ~500g increments in the singletons in panel A, which simply may reflect the assessment of BW measures in this group.

Table 5.2 - Results of the PGS prediction of BW category for NTR twins and singletons (N=17379)

Prop	β_{PGS}	SE_{PGS}	P_{PGS}	β_{TS}	SE_{TS}	P_{TS}	β_{INT}	SE_{INT}	P_{INT}	PGS R^2
0.001	0.06	0.01	2.06E-06	-1.07	0.02	<0.001	-0.03	0.01	0.09	0.00
0.003	0.06	0.01	2.69E-07	-1.07	0.02	<0.001	-0.02	0.01	0.09	0.00
0.005	0.12	0.01	3.74E-23	-1.07	0.02	<0.001	-0.03	0.01	0.06	0.01
0.01	0.18	0.01	1.68E-49	-1.07	0.02	<0.001	-0.06	0.01	6.24E-05	0.02
0.05	0.17	0.01	1.48E-42	-1.07	0.02	<0.001	-0.06	0.01	3.23E-05	0.01
0.1	0.17	0.01	6.58E-41	-1.07	0.02	<0.001	-0.06	0.01	3.09E-05	0.01
0.2	0.17	0.01	7.29E-40	-1.07	0.02	<0.001	-0.06	0.01	2.98E-05	0.01
0.3	0.17	0.01	1.44E-39	-1.07	0.02	<0.001	-0.06	0.01	2.83E-05	0.01
0.5	0.17	0.01	2.68E-39	-1.07	0.02	<0.001	-0.06	0.01	2.83E-05	0.01
INF	0.15	0.01	3.21E-33	-1.07	0.02	<0.001	-0.05	0.01	<0.001	0.01

Note: Prop (proportion) is the P-value fraction for SNP inclusion in the polygenic score (PGS), β is the regression coefficient for each term with standard error (SE) and P-value (P) for the PGS, twin status (TS), and the interaction term (INT) of TS and PGS. PGS R^2 is the phenotypic variance explained by the PGS.

DISCUSSION

We performed a genome-wide meta-analysis of BW in twins and compared the genetic architecture of BW between twins and singletons. Our results, particularly the genetic correlation and PGS analyses, provide compelling evidence for considerable genetic overlap between BW in twins and singletons.

The genetic correlation between BW in twins and the most recent reported results in singletons was very strong (r_g=0.92, 95% CI: 0.66-1.18), indicating a large overlap in the genetic variants influencing BW in the two groups. The genetic associations with health-related traits, when comparing the size and direction from our genetic correlation analyses with the results from Horikoshi and colleagues [23], showed remarkably similar results. This similarity suggests that the differential pattern of fetal growth between twins and singletons does not affect the relation between BW and later-life disease.

We evaluated the predictive performance of PGS derived from a GWAS on BW from a large representative population from the UK Biobank in a large target sample of NTR twins and non-twins. The PGS calculated from the proportion of SNPs with a *P*-value selection threshold of 0.01 demonstrated robust prediction in both singletons ($p=6.94\times10^{-57}$) and twins ($p=2.10\times10^{-51}$). While the proportion of variation explained by the best predicting PGS was small for twins at 2% and non-twins at 3%, despite moderate heritability estimates, such PGS represents common genetic architecture underlying BW in twins and singletons even though there are clear differences in BW between the two groups. Smaller heritability estimates were also observed for BW in twins, potentially indicating a form of sibling competition. That is if one twin grows and occupies the growing space of the co-twin, the genes that increase the BW of the larger twin may also limit the growth of the co-twin. Consistent with our results, sibling competition would result in a dampened effect of the PGS and would be reflected in lower heritability estimates in twins.

The results of the GWAMA did not yield SNPs significantly associated with BW in twins. Two lead SNPs, rs10800682 and rs3845913, had association signals of $P<5\times10^{-7}$. rs10800682 was not near (>2Mb away) and was independent ($r^2<0.05$) of all genome-wide significant loci found by Horikoshi and colleagues [23], making it a potential candidate for future twin studies. rs3845913 is an intronic variant of *ADCY5*, which, along with *CCNL1*, were two of the first genes to be robustly associated with fetal growth and BW [21]. Additionally, rs3845913 is ~31kb downstream and is in LD (r^2=0.154) with rs11719201 (an intronic variant of *ADCY5*), one of 60 loci previously associated with BW [23]. To pinpoint exactly how and through which gene(s) rs10800682 and rs3845913 may exert an effect on BW, additional and functional follow-up studies are necessary. Previously associated alleles at *ADCY5* were found to be BW lowering and risk increasing for type 2 diabetes, consistent with the fetal insulin hypothesis [35].

The results from this study strongly suggest that BW data from twins and singletons may be meta-analyzed together in GWAMA, despite the limited sample size of the discovery GWAMA in twins (N=42212). Another limitation is that we corrected for birth order, gestational age, and maternal age at birth in a majority of cohorts but could not do so for all cohorts due to data availability. This information should ideally always be included when BW data are collected.

Additionally, we report genome-wide estimates of shared genetic effects based on common genetic variation (SNPs with MAF>0.01 per default settings in LDHub). Suppose the effects of rare variants are not shared similarly to the effects of common variants for each phenotype comparison. In that case, the genetic correlation estimates could be misleading. However, in terms of their shared influences on pairs of phenotypes, there is not a theoretical reason to expect systematic differences in the effects of rare and common variants. Rare variants with larger effects would not preclude carrying far more numerous common variants with smaller effects. Thus, the genetic correlations presented in this study may provide reasonable estimates based on common genetic variation; however, to validate these findings, rare variant studies are needed. Future studies may also expand upon our genetic correlation estimates by utilizing non-European populations, greater sample sizes (for discovery and trait-specific phenotypes as they become available), and increased density across the genome.

Concerning the results of the PGS prediction, we note that the *P*-value selection threshold of the most predictive PGS is a function of the effect size distribution, the statistical power of the discovery GWAMA and the NTR target data, the genetic architecture of BW, as well as the fraction of associated markers.

Follow-up research may aim to get a better understanding of BW as it is influenced by direct fetal and indirect maternal genetic influences through the intrauterine environment. The amount of variance in BW explained by the maternal genotype has been estimated as substantially smaller than the fetal genetic contribution [47]. Recent work suggests that fetal size measurements at birth are predominantly determined by the fetal genome, whereas the gestational duration is primarily dictated by the maternal genome [48]. A better understanding of the genetic architecture of BW and fetal growth, more generally, will aid in the elucidation of immediate health outcomes (e.g., preterm birth, fetal growth restriction) and reveal relationships with later-life health outcomes (e.g., cardiovascular disease, type-2 diabetes).

To conclude, we show that based on genetic correlation and PGS analyses, the genetic architecture of BW in twins and singletons is similar. Of course, it is known that mean differences in BW between twins and singletons exist; however, the findings of this work strongly suggest that the genetic causes of variation are the same. Bearing this in mind, the results of this work indicate that it is appropriate to meta-analyze twins and singletons for genetic studies of BW. However, careful consideration of analytical strategies will be needed since details specific to twins may not apply to full-term singletons. Small groups of twins might still need to be excluded, for example, the highly discordant BW pairs due to the possibility for twin-to-twin transfusion syndrome (TTTS). Also, in full-term singletons, a typical gestational age cut-off for exclusion (e.g., born before 37 weeks) is often applied, which will not be applicable with the inclusion of twins due to shorter gestational duration [16] and delayed fetal growth after 32 weeks [38-41]. One approach to address these issues would be to perform separate GWAS on standardized BW in each group with appropriate exclusion criteria and covariates specific to twins and non-twins with subsequent meta-analysis of *P*-values since beta estimates and intercepts will be affected by raw differences in BW.

MATERIALS AND METHODS

Samples

Eight population-based twin registers supplied data: the Netherlands Twin Register (NTR) [49, 50], Queensland Institute of Medical Research (QIMR – comprised of the Queensland Twin Registry [51] and the Australian Twin Registry [52, 53]), Danish Twin Registry (DTR) [54], Finnish Twin Cohort Study (FinnTwin) [55, 56], Twins Early Development Study (TEDS) [57], Child and Adolescent Twin Study in Sweden (CATSS) [58-60], Avera Twin Register (ATR) [61, 62], and the UK Biobank (UKB) [63]. In UKB, twins were identified as previously described [64]. A detailed description of cohort sample characteristics can be found in Table 5.3. Information on genotyping and quality control procedures for each cohort can be found in Supplementary Material, Table 5.3.

Study-level analyses

Birth weight (BW) measures were z-score transformed ($[BW_{value}-BW_{mean}]/BW_{standard\ deviation}$) before analysis. Each participating study group performed the association analyses between each SNP genotype and BW *z*-scores with the following covariates where available: sex, gestational age, year of birth, maternal age at birth, birth order, and relevant study-specific metrics (e.g., principal components (PCs) correcting for genomic ancestry). For all cohorts, except ATR, birth order was available. The analysis was performed without adjustment for maternal age at birth and gestational age in the DTR. Association analyses were performed in PLINK *v1.07* [65] with the Generalized Estimation Equation (GEE) package using the R-package plugin to correct for family relatedness or according to local best practices (details provided in Supplementary Material, Table 5.3). Sample exclusion criteria were phenotypic outliers (BW *z*-score greater than or less than five standard deviations from the mean), premature births (gestational age less than 33 weeks), monozygotic (MZ) twins with TTTS, including twin pairs with BW more than 35% discordant (a group likely including TTTS twins), triplets and higher-order multiple births and participants with non-European ancestry.

Table 5.3 – Number of individuals, birth weight, and associated measures per cohort

Cohort	Country	Sample Size (M/F)	Mean (SD) BW (grams)	Birth year range	Mean (SD) Maternal Age (years)	Mean (SD) Gestational Age (weeks)	Data Collection
AVERA	USA	279 (88/191)	2431.97(547.42)	1939-2018	29.09 (4.92)	36.75 (2.92)	Self-Report, Parent-Report
CATSS	Sweden	13595 (6706/6889)	2651.83 (564.34)	1985-2005	30.72 (4.62)	36.54 (2.64)	Medical birth registry
DTR	Denmark	1432 (687/745)	2688.80 (534.10)	1903-1952	NA	NA	Mid-Wife Records and Self-Report
FinnTwin	Finland	1778 (812/966)	2749 (448.73)	1974-1987	29.21 (4.63)	37.36 (1.81)	Parent-Report
NTR	The Netherlands	6951 (2942/4009)	2586.16 (467.62)	1922-2012	30.00 (4.33)	37.14 (2.04)	National Youth Health Services, Self-Report and Parent-Report
QIMR	Australia	5435 (2263/3172)	2626.53 (510.54)	1922-1999	29.34 (5.04)	37.90 (2.14)	For birthweight and gestational age: Self-report or parental report depending on study (for adults); maternal report (for adolescents). For gestational age: assumed 37 weeks if not available. For birth year and maternal age: derived from dates of birth.
TEDS	UK	6527 (3109/3418)	2522.25 (530.86)	1994-1996	31.01 (4.79)	36.47 (2.41)	Parent-Report
UKB	UK	6215 (2300/3915)	2431.64 (737.42)	1937-1970	NA	NA	Self-Report (UKB ID 20022)

CATSS=Child and Adolescent Twin Study in Sweden, DTR=Danish Twin Registry, NTR=Netherlands Twin Register, QIMR= Queensland Institute of Medical Research, TEDS=Twins Early Development Study, and UKB=UK Biobank. M/F are counts of male and female individuals, respectively. SD is standard deviation. NA represents unavailable information.

Meta-analysis

Summary statistics from each cohort GWA analysis underwent another round of standard quality control before meta-analysis. The R-package EasyQC [66] was used to perform quality control analyses. Insertions and deletions, SNPs with missing or invalid values, markers with Minor Allele Frequency (MAF)<0.01, and those with poor imputation quality (<0.30) were excluded. Resulting quality controlled summary statistics from each cohort were meta-analyzed using the inverse variance-based approach in METAL [67]. Genomic control was applied to adjust the statistics generated by each cohort [68]. In the meta-analysis, SNPs present in greater than 70% of all participants were retained.

Association tests

FUMA (FUnctional Annotation and Mapping *v1.3.6*) [69] was used to annotate GWAMA results and identify genomic risk loci. These loci were defined as independent lead SNPs exhibiting maximum distance between their linkage-disequilibrium (LD) block. For genome-wide significance in the meta-analysis, a *P*-value threshold of $5x10^{-8}$ was adopted. The minimum threshold for defining independent significant SNPs was $r^2 \geq 0.6$, which was used to determine the borders of the genomic risk loci. The minimum threshold for defining lead SNPs, used for clumping the independent significant SNPs, was $r^2 \geq 0.1$. Independent significant SNPs closer than 250kb were merged into one genomic risk locus. SNPs in LD with the independent significant SNPs were considered candidate SNPs and defined the borders of the genomic risk loci. We tested whether the signals from our analyses overlap with previously identified loci for BW in singletons. In agreement with Horikoshi et al. [23], if a lead SNP mapped >2Mb away from, and was statistically independent (LD $r^2<0.05$ based on European population reference set) of any of the 60 previously identified loci, it was considered novel. We calculated the r^2 between the signals with the web-based application LDmatrix contained within the LDlink (*v3.8*) [70] suite of tools.

Genetic correlations

To quantify the degree of shared genetic contribution between BW in twins and BW in singletons and to correlate BW in twins to other individual-level health-related traits and diseases, we employed LD Hub (*v1.9.3*) (http://ldsc.broadinstitute.org/ldhub/) [71]. LD Hub is a centralized database of summary-level GWA study results facilitating the calculation of genetic correlations [72] between user-supplied summary statistics and a variety of user-selected traits using LD score regression [73]. HapMap3 SNPs from summary statistics of the GWAS for each trait and pre-computed LD scores were used in the analyses (available on: https://github.com/bulik/ldsc). LD score regression requires large sample sizes and utilizes LD information from an ancestry-matched reference panel; therefore, genetic correlation analyses were constrained to European GWA study samples. SNPs with a MAF ≤0.01 were excluded.

For the comparisons with previous genome-wide genetic correlation analyses in singletons (7), we selected the following categories of traits: anthropometric traits, reproductive traits, glycemic traits, autoimmune disorders, cognitive abilities, psychiatric diseases, and smoking behavior. In total, we tested for association with 57 traits.

SNP heritability (h^2) was calculated in LD Hub with LD score regression to evaluate how much of the variation in BW could be ascribed to common additive genetic variation.

PolyGenic Score prediction

GWAS results on BW from the UK Biobank (data field 20022) (http://www.nealelab.is/uk-biobank/) served as the discovery set for calculating polygenic scores (PGS) in the NTR target dataset. For the PGS prediction of BW in the NTR, participants with complete BW data and maximum information on covariates (genomic PCs, sex, year of birth, gestational age, twin status, and genotyping platform) were included. When not available, gestational age was imputed with the mean gestational age separately for twins (mean=37.38 weeks) and singletons (mean=39.89 weeks). Genotyping platform and ten genomic PCs were included in the model to account for batch effects (i.e., non-random selection of samples genotyped on specific arrays) and residual population stratification. The target sample consisted of 17,379 individuals, comprising 10,487 twins and 6,892 singletons. Summary statistics from the UK Biobank GWAS on BW were adjusted for the effects of LD with LDpred [74] using the LD structure of European populations in the 1000 Genomes references set [75]. Recalculated effect size estimates representing ten thresholds of *P*-value significance (0.001, 0.003, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, INF (infinitesimal)) were used for allelic scoring in PLINK [65].

We used the PGS to predict BW in NTR twins and singletons using GEE methods in R [76], taking into account familial relationships. We also evaluated the predictive performance of the PGS on categorical BW in the entire target sample of twins and singletons by including twin status and an interaction term of PGS and twin status in the regression model. Six categories were constructed, representing the following BW ranges: <2000 grams, 2000-2500 grams, 2501-3000 grams, 3001-3500 grams, 3501-4000 grams, >4000 grams. Complete regression equations can be found in the Supplementary Material - Methods. The phenotypic variance explained, captured by R^2, was used to evaluate the predictive performance of each PGS. Our main interest was to determine how well PGS derived from a large discovery population, reflecting general population numbers of twins, could predict BW in a separate target population of twins and singletons.

DATA ACCESS

Summary statistics for the GWAMA of BW in twins can be downloaded from the GWAS catalog website: https://www.ebi.ac.uk/gwas/.

ACKNOWLEDGMENTS

We are extremely grateful to the participants, families, and teams of investigators who contributed to this work. The research has been conducted using data from UK Biobank, a major biomedical database, under Application Number 25472. For additional study-specific acknowledgments, please refer to Supplementary Material, Table 5.4.

CONFLICT OF INTEREST STATEMENT

None declared.

FUNDING

Wesley W. Parke Research Award Endowment. The Avera Twin Register is supported by Avera Health, Avera McKennan Hospital, and Avera Institute for Human Genetics. The collaboration between the Netherlands and the Avera Twin Register arose through NIHM Grant: 1RC2MH089995-01: Genomics of Developmental Trajectories in Twins. CATSS is part of the Swedish Twin Registry, which is managed by Karolinska Institutet and receives funding through the Swedish Research Council under the grant no 2017-00641. The DTR data collection is supported by grants from The National Program for Research Infrastructure 2007 from the Danish Agency for Science, Technology and Innovation and the US National Institutes of Health (P01 AG08761). Genotyping was supported by NIH R01 AG037985 (Pedersen). Phenotyping and genotyping of the Finnish twin cohorts was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (grants 213506, 129680), the Academy of Finland (grants 10049, 205585, 118555, 141054, 265240, 263278 and 264146 to J. Kaprio), National Institute of Alcohol Abuse and Alcoholism (grants AA-12502, AA-00145, and AA-09203 to R. J. Rose and AA15416 and K02AA018755 to D. M. Dick), and the Wellcome Trust Sanger Institute, UK. For the NTR, funding was provided by ZonMw (Grant Nos. 904-61-090, 985-10-002, 912-10-020, 904-61-193, 480-04-004, 463-06-001, 451-04-034, 400-05-717, 016-115-035, 481-08-011 and 056-32-010), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Grant Nos. Addiction-31160008, NWO-Middelgroot-911-09-032, OCW_NWO Gravity program -024.001.003, NWO-Groot 480-15-001/674, NWO-56-464-14192), Centre for Medical Systems Biology (CSMB, NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (Grant Nos. 184.021.007, 184.033.111), Koninklijke Nederlandse Akademie van Wetenschappen (NL) (Grant No. PAH/6635), European Science Foundation (Grant No. EU/QLRT-2001-01254), FP7 Health (Grant Nos. 01413: ENGAGE, 602768: ACTION), H2020 European Research Council (Grant Nos. ERC AG 230374, ERC SG 284167, ERC CG 771057), National Institutes of Health (Grant No. NIH R01 DK092127-04) and Avera Institute for Human Genetics. Bart Baselmans: NWO/ZonMw: Rubicon 45219101. For QIMR, funding for data collection and/or genotyping was provided by the Australian National Health and Medical Research Council (NHMRC); the Australian Research Council (ARC); the FP-5 GenomEUtwin Project; the US National Institutes of Health (NIH); and the Center for Inherited Disease Research (CIDR; Baltimore, MD, USA). TEDS is supported by a program grant to RP from the UK Medical Research Council (MR/M021475/1 and previously G0901245), with additional support from the US National Institutes of Health (AG046938). The research leading to these results has also received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007–2013)/ grant agreement nº 602768 and ERC grant agreement nº 295366. RP is supported by a Medical Research Council Professorship award (G19/2). High performance computing facilities were funded with capital equipment grants from the GSTT Charity (TR130505) and Maudsley Charity (980).

REFERENCES

1. Wilcox, A.J., *On the importance--and the unimportance--of birthweight.* Int J Epidemiol, 2001. **30**(6): p. 1233-41.
2. Wilcox, A.J., *Birth weight and perinatal mortality: the effect of maternal smoking.* Am J Epidemiol, 1993. **137**(10): p. 1098-104.
3. Wilcox, A.J. and I.T. Russell, *Birthweight and perinatal mortality: II. On weight-specific mortality.* Int J Epidemiol, 1983. **12**(3): p. 319-25.
4. Barker, D.J. and P.M. Clark, *Fetal undernutrition and disease in later life.* Rev Reprod, 1997. **2**(2): p. 105-12.
5. Johansson, M. and F. Rasmussen, *Birthweight and body mass index in young adulthood: the Swedish young male twins study.* Twin Res, 2001. **4**(5): p. 400-5.
6. Sorensen, H.T., et al., *Relation between weight and length at birth and body mass index in young adulthood: cohort study.* BMJ, 1997. **315**(7116): p. 1137.
7. Eriksson, M., et al., *Birth weight and cardiovascular risk factors in a cohort followed until 80 years of age: the study of men born in 1913.* J Intern Med, 2004. **255**(2): p. 236-46.
8. Wang, S.F., et al., *Birth weight and risk of coronary heart disease in adults: a meta-analysis of prospective cohort studies.* J Dev Orig Health Dis, 2014. **5**(6): p. 408-19.
9. Whincup, P.H., et al., *Birth weight and risk of type 2 diabetes: a systematic review.* JAMA, 2008. **300**(24): p. 2886-97.
10. Gamborg, M., et al., *Birth weight and systolic blood pressure in adolescence and adulthood: meta-regression analysis of sex- and age-specific results from 20 Nordic studies.* Am J Epidemiol, 2007. **166**(6): p. 634-45.
11. RG, I.J., C.D. Stehouwer, and D.I. Boomsma, *Evidence for genetic factors explaining the birth weight-blood pressure relation. Analysis in twins.* Hypertension, 2000. **36**(6): p. 1008-12.
12. Law, C.M. and A.W. Shiell, *Is blood pressure inversely related to birth weight? The strength of evidence from a systematic review of the literature.* J Hypertens, 1996. **14**(8): p. 935-41.
13. Cheung, Y.B., et al., *Birthweight and psychological distress in adult twins: a longitudinal study.* Acta Paediatr, 2004. **93**(7): p. 965-8.
14. Kramer, M.S., *Determinants of low birth weight: methodological assessment and meta-analysis.* Bull World Health Organ, 1987. **65**(5): p. 663-737.
15. Jarvelin, M.R., et al., *Ecological and individual predictors of birthweight in a northern Finland birth cohort 1986.* Paediatr Perinat Epidemiol, 1997. **11**(3): p. 298-312.
16. Gielen, M., et al., *Secular trends in gestational age and birthweight in twins.* Hum Reprod, 2010. **25**(9): p. 2346-53.
17. Hur, Y.M., et al., *A comparison of twin birthweight data from Australia, the Netherlands, the United States, Japan, and South Korea: are genetic and environmental variations in birthweight similar in Caucasians and East Asians?* Twin Res Hum Genet, 2005. **8**(6): p. 638-48.
18. Clausson, B., P. Lichtenstein, and S. Cnattingius, *Genetic influence on birthweight and gestational length determined by studies in offspring of twins.* BJOG, 2000. **107**(3): p. 375-81.

19. Mook-Kanamori, D.O., et al., *Heritability estimates of body size in fetal life and early childhood.* PLoS One, 2012. **7**(7): p. e39901.

20. van Dommelen, P., et al., *Genetic study of the height and weight process during infancy.* Twin Res, 2004. **7**(6): p. 607-16.

21. Freathy, R.M., et al., *Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight.* Nat Genet, 2010. **42**(5): p. 430-5.

22. Horikoshi, M., et al., *New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism.* Nat Genet, 2013. **45**(1): p. 76-82.

23. Horikoshi, M., et al., *Genome-wide associations for birth weight and correlations with adult disease.* Nature, 2016. **538**(7624): p. 248-252.

24. Beaumont, R.N., et al., *Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics.* Hum Mol Genet, 2018. **27**(4): p. 742-756.

25. Warrington, N.M., et al., *Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors.* Nat Genet, 2019. **51**(5): p. 804-814.

26. Metrustry, S.J., et al., *Variants close to NTRK2 gene are associated with birth weight in female twins.* Twin Res Hum Genet, 2014. **17**(4): p. 254-61.

27. Barker, D.J. and C. Osmond, *Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales.* Lancet, 1986. **1**(8489): p. 1077-81.

28. Barker, D.J., et al., *Weight in infancy and death from ischaemic heart disease.* Lancet, 1989. **2**(8663): p. 577-80.

29. Barker, D.J., et al., *Fetal nutrition and cardiovascular disease in adult life.* Lancet, 1993. **341**(8850): p. 938-41.

30. de Boo, H.A. and J.E. Harding, *The developmental origins of adult disease (Barker) hypothesis.* Aust N Z J Obstet Gynaecol, 2006. **46**(1): p. 4-14.

31. Geelhoed, J.J. and V.W. Jaddoe, *Early influences on cardiovascular and renal development.* Eur J Epidemiol, 2010. **25**(10): p. 677-92.

32. Xu, X.F., et al., *Effect of low birth weight on childhood asthma: a meta-analysis.* BMC Pediatr, 2014. **14**: p. 275.

33. O'Donnell, K.J. and M.J. Meaney, *Fetal Origins of Mental Health: The Developmental Origins of Health and Disease Hypothesis.* Am J Psychiatry, 2017. **174**(4): p. 319-328.

34. Orri, M., et al., *Contribution of birth weight to mental health, cognitive and socioeconomic outcomes: two-sample Mendelian randomisation.* Br J Psychiatry, 2021: p. 1-8.

35. Hattersley, A.T. and J.E. Tooke, *The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease.* Lancet, 1999. **353**(9166): p. 1789-92.

36. Evans, D.M., et al., *Elucidating the role of maternal environmental exposures on offspring health and disease using two-sample Mendelian randomization.* Int J Epidemiol, 2019. **48**(3): p. 861-875.

37. Moen, G.H., et al., *Mendelian randomization study of maternal influences on birthweight and future cardiometabolic risk in the HUNT cohort.* Nat Commun, 2020. **11**(1): p. 5404.

38. Loos, R.J., et al., *Determinants of birthweight and intrauterine growth in liveborn twins.* Paediatr Perinat Epidemiol, 2005. **19 Suppl 1**: p. 15-22.

39. Bleker, O.P., W. Breur, and B.L. Huidekoper, *A study of birth weight, placental weight and mortality of twins as compared to singletons.* Br J Obstet Gynaecol, 1979. **86**(2): p. 111-8.

40. Kingdom, J.C., O. Nevo, and K.E. Murphy, *Discordant growth in twins.* Prenat Diagn, 2005. **25**(9): p. 759-65.

41. Senoo, M., et al., *Growth pattern of twins of different chorionicity evaluated by sonographic biometry.* Obstet Gynecol, 2000. **95**(5): p. 656-61.

42. de Geus, E.J., et al., *Comparing blood pressure of twins and their singleton siblings: being a twin does not affect adult blood pressure.* Twin Res, 2001. **4**(5): p. 385-91.

43. McNeill, G., et al., *Blood pressure in relation to birth weight in twins and singleton controls matched for gestational age.* Am J Epidemiol, 2003. **158**(2): p. 150-5.

44. Andrew, T., et al., *Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women.* Twin Res, 2001. **4**(6): p. 464-77.

45. Petersen, I., et al., *No evidence of a higher 10 year period prevalence of diabetes among 77,885 twins compared with 215,264 singletons from the Danish birth cohorts 1910-1989.* Diabetologia, 2011. **54**(8): p. 2016-24.

46. Ijzerman, R.G., D.I. Boomsma, and C.D. Stehouwer, *Intrauterine environmental and genetic influences on the association between birthweight and cardiovascular risk factors: studies in twins as a means of testing the fetal origins hypothesis.* Paediatr Perinat Epidemiol, 2005. **19 Suppl 1**: p. 10-4.

47. Magnus, P., *Causes of variation in birth weight: a study of offspring of twins.* Clin Genet, 1984. **25**(1): p. 15-24.

48. Srivastava, A.K., et al., *Haplotype-based heritability estimations reveal gestational duration as a maternal trait and fetal size measurements at birth as fetal traits in human pregnancy.* bioRxiv, 2020: p. 2020.05.12.079863.

49. van Beijsterveldt, C.E., et al., *The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children.* Twin Res Hum Genet, 2013. **16**(1): p. 252-67.

50. Willemsen, G., et al., *The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection.* Twin Res Hum Genet, 2013. **16**(1): p. 271-81.

51. Wright, M.J., Martin, N., *The Brisbane Adolescent Twin Study: Outline of study methods and research projects.* The Australian Journal of Psychology, 2004. **56**: p. 56-78.

52. Slutske, W.S., et al., *The Australian Twin Study of Gambling (OZ-GAM): rationale, sample description, predictors of participation, and a first look at sources of individual differences in gambling involvement.* Twin Res Hum Genet, 2009. **12**(1): p. 63-78.

53. Medland, S.E., et al., *Common variants in the trichohyalin gene are associated with straight hair in Europeans.* Am J Hum Genet, 2009. **85**(5): p. 750-5.

54. Pedersen, D.A., et al., *The Danish Twin Registry: An Updated Overview.* Twin Res Hum Genet, 2019. **22**(6): p. 499-507.

55. Kaprio, J., *The Finnish Twin Cohort Study: an update.* Twin Res Hum Genet, 2013. **16**(1): p. 157-62.

56. Kaprio, J., L. Pulkkinen, and R.J. Rose, *Genetic and environmental factors in health-related behaviors: studies on Finnish twins and twin families.* Twin Res, 2002. **5**(5): p. 366-71.

57. Haworth, C.M., O.S. Davis, and R. Plomin, *Twins Early Development Study (TEDS): a genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood.* Twin Res Hum Genet, 2013. **16**(1): p. 117-25.

58. Anckarsater, H., et al., *The Child and Adolescent Twin Study in Sweden (CATSS).* Twin Res Hum Genet, 2011. **14**(6): p. 495-508.

59. Magnusson, P.K., et al., *The Swedish Twin Registry: establishment of a biobank and other recent developments.* Twin Res Hum Genet, 2013. **16**(1): p. 317-29.

60. Ortqvist, A.K., et al., *Familial factors do not confound the association between birth weight and childhood asthma.* Pediatrics, 2009. **124**(4): p. e737-43.

61. Kittelsrud, J., et al., *Establishment of the Avera Twin Register in the Midwest USA.* Twin Res Hum Genet, 2017. **20**(5): p. 414-418.

62. Kittelsrud, J.M., et al., *Avera Twin Register Growing Through Online Consenting and Survey Collection.* Twin Res Hum Genet, 2019. **22**(6): p. 686-690.

63. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.

64. Mbarek, H., et al., *Biological insights into multiple birth: genetic findings from UK Biobank.* Eur J Hum Genet, 2019. **27**(6): p. 970-979.

65. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

66. Winkler, T.W., et al., *Quality control and conduct of genome-wide association meta-analyses.* Nat Protoc, 2014. **9**(5): p. 1192-212.

67. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.

68. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.

69. Watanabe, K., et al., *Functional mapping and annotation of genetic associations with FUMA.* Nat Commun, 2017. **8**(1): p. 1826.

70. Machiela, M.J. and S.J. Chanock, *LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants.* Bioinformatics, 2015. **31**(21): p. 3555-7.

71. Zheng, J., et al., *LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis.* Bioinformatics, 2017. **33**(2): p. 272-279.

72. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-41.

73. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-5.

74. Vilhjalmsson, B.J., et al., *Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.* Am J Hum Genet, 2015. **97**(4): p. 576-92.

75. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

76. Minica, C.C., et al., *Sandwich corrected standard errors in family-based genome-wide association studies.* Eur J Hum Genet, 2015. **23**(3): p. 388-94.

SUPPLEMENTARY MATERIALS

Supplementary Methods

Formulas for PGS prediction

PGS were calculated from summary statistics of a UK Biobank GWAS on BW (http://www.nealelab.is/uk-biobank/) and were used to predict BW in NTR twins and singletons.

The formula below was used to evaluate the prediction in twins and singletons separately:

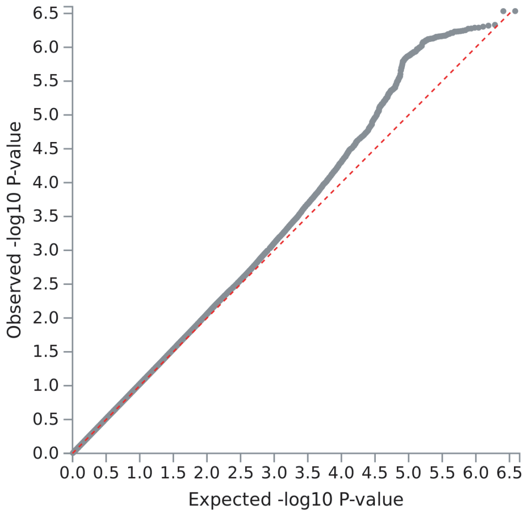
$$\mathbf{BW}_{\text{raw}} \sim \beta_{\text{Genomic PCs}} + \beta_{\text{Sex}} + \beta_{\text{Gestational age}} + \beta_{\text{Year of birth}} + \beta_{\text{PGS}} + \beta_{\text{Genotyping platform}}$$

The formula below was used to evaluate the prediction in the entire target sample (twins and singletons) using BW category as the response. In this model, we included a main effect of twin status and an interaction term between twin status and the PGS.

For prediction in the full target sample using BW category:

$$\mathbf{BW}_{\text{category}} \sim \beta_{\text{Genomic PCs}} + \beta_{\text{Sex}} + \beta_{\text{Gestational age}} + \beta_{\text{Year of birth}} + \beta_{\text{PGS}} + \beta_{\text{Genotyping platform}} + \beta_{\text{Twin status}} + \beta_{\text{PGS}} * \beta_{\text{Twin status}}$$

Supplementary Figure 5.1



Supplementary Table 5.1 - Summary statistics of the 150 overlapping SNPs between the genome-wide significant loci identified by Warrington et al., 2019 (https://pubmed.ncbi.nlm.nih.gov/31043758/) for own birth weight and the current study. Those highlighted in bold indicate SNPs with opposing effects between the two studies following alignment of effect alleles.

				Summary statistics from Warrington et al., 2019: GWAMA of own birth weight. Shown are the genome-wide significant SNPs ($P<6.6x10^{-9}$)											Summary statistics from the current study: GWAMA results of twin birth weight									
SNP	Chromosome	Position (hg19)	Band	Signal Name (fet_nearest/mat_nearest/additional)[a]	Signal Number[b]	Effect Allele	Other Allele	EAF[c]	Beta	Se[c]	P-value	Sample size	LD score regression adjusted P-value	Heterogeneity P-value	Allele1[d]	Allele2[e]	Freq1[f]	FreqSE[f]	Effect[g]	StdErr[g]	Meta-analysis P-value	totalN	AdjEffect[h]	HetPVal[i]
rs12401656	1	43456767	1p34.2	FLJ32224/SLC2A1	2	G	A	0.865	0.025	0.004	3.40E-11	292712	1.70E-10	0.255	A	G	0.1245	0.0129	-0.0116	0.0088	0.1879	42212	0.0116	0.3805
rs80278614	1	119412317	1p12	TBX15	3	A	G	0.054	0.04	0.006	6.50E-12	292074	3.70E-11	0.676	A	G	0.0607	0.0135	0.0053	0.0127	0.6778	41367	0.0053	0.1083
rs905938	1	154991389	1q21.3	DCST2/KCNN3	4	C	T	0.262	0.026	0.003	2.80E-19	298135	5.40E-18	0.149	T	C	0.7374	0.0098	-0.0199	0.0069	0.004259	40780	0.0199	0.3715
rs670523	1	155878732	1q22	RIT1/LMNA	5	G	A	0.669	0.019	0.003	7.60E-12	291451	4.30E-11	0.075	A	G	0.3442	0.0124	-0.0195	0.0069	0.004477	35261	0.0195	0.2806
rs72480273	**1**	**161644871**	**1q23.3**	**FCGR2B/FCGR2C/HSPA6**	**6**	**C**	**A**	**0.182**	**0.023**	**0.003**	**4.00E-11**	**291667**	**2.00E-10**	**0.319**	**A**	**C**	**0.8338**	**0.0193**	**0.0127**	**0.0087**	**0.1478**	**39280**	**-0.0127**	**0.323**
rs10913200	1	176521655	1q25.2	PAPPA2	7	G	A	0.972	0.051	0.008	2.00E-10	287089	9.10E-10	0.2	A	G	0.0251	0.0044	-0.0396	0.0234	0.09092	38723	0.0396	0.1875
rs61830764	1	212289976	1q32.3	DTL	8	A	G	0.377	0.017	0.003	1.10E-09	291445	4.50E-09	0.004	A	G	0.3875	0.0096	0.0139	0.0061	0.02285	41646	0.0139	0.6545
rs3806315	1	214724668	1q41	PTPN14	9	A	G	0.591	0.018	0.003	2.80E-11	289070	1.40E-10	0.439	A	G	0.5919	0.0112	0.0101	0.0061	0.09586	42212	0.0101	0.2742
rs708122	1	228216997	1q42.13	WNT3A	10	C	A	0.681	0.017	0.003	2.50E-09	292718	9.20E-09	0.543	A	C	0.3208	0.0061	-0.0098	0.0062	0.1171	42212	0.0098	0.593
rs10495563	2	9662210	2p25.1	ADAM17	11	A	G	0.664	0.022	0.003	2.10E-16	298133	2.50E-15	0.253	A	G	0.6508	0.0234	0.0188	0.0061	0.001967	42212	0.0188	0.1907
rs11893688	2	9695282	2p25.1	ADAM17	11	T	C	0.661	0.022	0.003	1.30E-15	292716	1.40E-14	0.452	T	C	0.6432	0.0242	0.0186	0.0063	0.003171	42212	0.0186	0.1468
rs2551347	2	23912401	2p24.1	KLHL29	12	T	C	0.749	0.024	0.003	1.90E-16	292714	2.30E-15	0.621	T	C	0.7578	0.018	0.0074	0.0069	0.2815	42212	0.0074	0.1905
rs754868	2	43185532	2p21	HAAO	14	G	A	0.419	0.016	0.003	6.70E-10	298139	2.80E-09	0.998	A	G	0.5716	0.0059	-0.0033	0.0058	0.5731	42212	0.0033	0.1542
rs17034876	2	46484310	2p21	EPAS1	16	T	C	0.7	0.042	0.003	3.10E-47	287749	6.60E-44	0.044	T	C	0.7013	0.0159	0.0161	0.0069	0.0198	40780	0.0161	0.912
rs4953353	2	46567276	2p21	EPAS1	17	G	T	0.632	0.018	0.003	3.50E-11	292721	1.80E-10	0.33	T	G	0.3705	0.0141	-0.0073	0.0064	0.2536	40780	0.0073	0.8827
rs2280235	2	191843830	2q32.2	STAT1	21	G	A	0.259	0.018	0.003	6.90E-10	292718	2.90E-09	0.979	A	G	0.7502	0.0117	-0.022	0.0069	0.001452	42212	0.022	0.4564
rs10181515	2	227019461	2q36.3	LOC646736/COL4A4/IRS1	22	T	C	0.225	0.021	0.003	2.10E-12	298138	1.30E-11	0.742	T	C	0.233	0.0076	0.0097	0.0069	0.1639	42212	0.0097	0.3316
rs2168443	**3**	**46947087**	**3p21.31**	**PTH1R**	**24**	**T**	**A**	**0.379**	**0.017**	**0.003**	**3.90E-10**	**292713**	**1.70E-09**	**0.617**	**A**	**T**	**0.6307**	**0.0146**	**0.0015**	**0.0061**	**0.8105**	**42212**	**-0.0015**	**0.566**
rs11708067	3	123065778	3q21.1	ADCY5	25	G	A	0.238	0.041	0.003	1.60E-42	298128	1.60E-39	0.029	A	G	0.7676	0.0176	-0.0237	0.0068	0.0005198	42212	0.0237	0.002629
rs9851257	3	123125711	3q21.1	ADCY5	26	T	A	0.733	0.02	0.003	2.40E-12	298130	1.50E-11	0.249	A	T	0.2655	0.0135	-0.0179	0.0066	0.006922	42212	0.0179	0.7552
rs2306700	3	142123841	3q23	XRN1	28	T	C	0.136	0.023	0.004	1.80E-09	290416	7.00E-09	0.644	T	C	0.1424	0.0061	0.021	0.0084	0.01267	42212	0.021	0.3651
rs10935733	3	148622968	3q24	CPA3/AGTR1	29	T	C	0.399	0.019	0.003	2.30E-13	292713	1.70E-12	0.848	T	C	0.3983	0.0117	0.0132	0.006	0.02699	42212	0.0132	0.2502
rs1482852	3	156798294	3q25.31	LOC339894/CCNL1	31	A	G	0.599	0.05	0.003	1.60E-82	298130	1.40E-76	0.007	A	G	0.5859	0.0182	0.0293	0.0061	1.74E-06	40780	0.0293	0.363
rs11711420	3	183349010	3q27.1	KLHL24	32	T	G	0.747	0.019	0.003	3.20E-10	292710	1.40E-09	0.983	T	G	0.7583	0.0111	0.0134	0.007	0.05478	42212	0.0134	0.8188
rs4144829	4	17903654	4p15.31	LCORL/DCAF16	33	C	T	0.267	0.036	0.003	4.30E-34	292713	1.00E-31	0.407	T	C	0.723	0.0133	-0.0296	0.0066	7.44E-06	42212	0.0296	0.6675
rs2174633	4	17917781	4p15.31	LCORL/DCAF16	33	A	C	0.27	0.035	0.003	7.10E-33	292712	1.30E-30	0.324	A	C	0.28	0.0132	0.0293	0.0065	7.43E-06	42212	0.0293	0.7244
rs6533183	4	106133184	4q24	TET2	34	C	T	0.352	0.022	0.003	6.80E-16	292715	7.50E-15	0.947	T	C	0.6372	0.0171	-0.0113	0.006	0.0606	42212	0.0113	0.5824
rs116807401	4	135121721	4q28.3	PABPC4L	35	C	T	0.018	0.077	0.01	2.20E-13	265314	1.60E-12	0.949	T	C	0.9811	0.0035	-0.0623	0.0233	0.007508	38723	0.0623	0.3317
rs6845999	4	145565826	4q31.21	LOC646576/HHIP	36	T	C	0.431	0.026	0.003	1.50E-24	298140	6.90E-23	0.168	T	C	0.4372	0.0089	0.0104	0.0059	0.0775	42212	0.0104	0.3263
rs2131354	4	145599908	4q31.21	LOC646576/HHIP	36	A	G	0.527	0.026	0.003	3.50E-24	292719	1.50E-22	0.438	A	G	0.5319	0.0085	0.0122	0.0059	0.03751	42212	0.0122	0.5018
rs1818782	5	39424628	5p13.1	DAB2	38	C	A	0.637	0.016	0.003	4.20E-09	313072	1.50E-08	0.714	A	C	0.344	0.0111	-0.0061	0.0062	0.3298	42212	0.0061	0.09383
rs351930	5	52003397	5q11.2	PELO	39	T	A	0.801	0.019	0.003	2.90E-09	292714	1.10E-08	0.674	A	T	0.2017	0.0107	-0.0015	0.0073	0.8411	42212	0.0015	0.7152
rs854037	5	57091783	5q11.2	ACTBL2	40	A	G	0.814	0.027	0.003	9.40E-16	292718	1.00E-14	0.032	A	G	0.8074	0.011	0.006	0.0075	0.4204	42212	0.006	0.8522
rs28365970	5	67585723	5q13.1	PIK3R1	41	C	A	0.741	0.02	0.003	1.70E-11	292712	8.90E-11	0.682	A	C	0.2607	0.0081	-0.0049	0.0067	0.4644	42212	0.0049	0.5081
rs6871635	5	133830395	5q31.1	PHF15	42	G	A	0.566	0.016	0.003	3.00E-09	292716	1.10E-08	0.554	A	G	0.4232	0.0202	-0.0028	0.006	0.6477	41646	0.0028	0.2942
rs1981627	5	133838180	5q31.1	PHF15	42	G	A	0.585	0.017	0.003	8.40E-11	292716	4.00E-10	0.599	A	G	0.4011	0.021	-0.0029	0.006	0.6311	42212	0.0029	0.4521
rs2946179	**5**	**157886627**	**5q33.3**	**EBF1**	**43**	**C**	**T**	**0.734**	**0.02**	**0.003**	**1.10E-11**	**298129**	**6.30E-11**	**0.901**	**T**	**C**	**0.2573**	**0.0108**	**0.0055**	**0.0067**	**0.4098**	**42212**	**-0.0055**	**0.7596**
rs35261542	6	20675792	6p22.3	CDKAL1	46	C	A	0.733	0.041	0.003	2.80E-45	298124	4.20E-42	0.074	A	C	0.2708	0.0152	-0.0171	0.0065	0.008888	42212	0.0171	0.0217
rs9379832	6	26186200	6p22.2	HIST1H2BE/HIST1H2BH	47	A	G	0.73	0.022	0.003	1.10E-13	291448	8.20E-13	0.156	A	G	0.6843	0.0353	0.0044	0.0065	0.5007	42212	0.0044	0.4044
rs9366778	6	31269173	6p21.33	HLA-C	48	G	A	0.627	0.018	0.003	2.90E-11	282578	1.50E-10	0.744	A	G	0.4137	0.0335	-0.0121	0.0062	0.05298	35997	0.0121	0.103
rs9267812	6	32128394	6p21.32	PPT2	50	T	C	0.133	0.023	0.004	3.10E-09	280156	1.20E-08	0.13	T	C	0.1281	0.0116	0.0091	0.0089	0.3106	42212	0.0091	0.07448
rs1547669	6	33775641	6p21.31	MLN	51	G	A	0.497	0.018	0.003	6.20E-12	289000	3.60E-11	0.908	A	G	0.5097	0.0138	-0.0118	0.0058	0.04183	42212	0.0118	0.3461
rs75104038	6	34190104	6p21.31	HMGA1	52	A	G	0.06	0.045	0.006	4.30E-16	289515	4.90E-15	0.228	A	G	0.0566	0.0038	0.0236	0.0133	0.07581	40155	0.0236	0.2913
rs75034466	6	34199815	6p21.31	HMGA1	52	T	C	0.046	0.046	0.006	1.80E-13	289010	1.30E-12	0.532	T	C	0.0423	0.0027	0.0259	0.0152	0.08885	38723	0.0259	0.4599
rs6911621	6	35529025	6p21.31	FKBP5/MAPK13/TEAD3	53	T	C	0.344	0.018	0.003	1.60E-11	292722	8.80E-11	0.749	T	C	0.3614	0.0208	0.0123	0.0061	0.04317	42212	0.0123	0.1844
rs9348981	6	35687249	6p21.31	FKBP5/MAPK13/TEAD3	53	T	G	0.71	0.021	0.003	2.20E-13	292710	1.60E-12	0.905	T	G	0.718	0.0127	0.0118	0.0066	0.07107	42212	0.0118	0.8254
rs7744700	6	53349401	6p12.1	GCLC	54	T	A	0.711	0.02	0.003	1.60E-11	291448	8.80E-11	0.128	A	T	0.2986	0.0079	-0.0185	0.0068	0.006776	40780	0.0185	0.4149

rs76094073	6	109288036	6q21	ARMC2/SESN1	55	G	C	0.121	0.027	0.004	1.60E-11	292719	8.80E-11	0.235	C	G	0.8793	0.0076	-0.0155	0.0088	0.0784	42212	0.0155	0.551
rs6925689	6	126865884	6q22.32	CENPW	56	T	C	0.494	0.015	0.003	6.40E-09	292716	2.30E-08	0.322	T	C	0.4927	0.0094	0.0128	0.0058	0.02721	42212	0.0128	0.9354
rs6569647	6	130337266	6q23.1	L3MBTL3	57	T	C	0.802	0.02	0.003	6.30E-10	292720	2.60E-09	0.715	T	C	0.8029	0.0072	0.0077	0.0074	0.2981	42212	0.0077	0.7147
rs1415701	6	130345835	6q23.1	L3MBTL3	57	G	A	0.736	0.018	0.003	1.40E-09	298129	5.30E-09	0.138	A	G	0.2625	0.0091	-0.0074	0.007	0.2862	40780	0.0074	0.5545
rs6930558	6	141878920	6q24.1	NMBR	58	T	G	0.747	0.022	0.003	3.40E-13	292714	2.50E-12	0.744	T	G	0.7476	0.0041	0.019	0.0067	0.004635	42212	0.019	0.1557
rs962554	**6**	**142734204**	**6q24.1**	**GPR126**	**59**	**T**	**C**	**0.715**	**0.017**	**0.003**	**3.80E-09**	**292717**	**1.40E-08**	**0.679**	**T**	**C**	**0.7173**	**0.009**	**-0.006**	**0.0064**	**0.3532**	**42212**	**-0.006**	**0.4938**
rs10872678	6	152039964	6q25.1	ESR1	60	T	C	0.724	0.032	0.003	9.80E-29	298136	9.10E-27	0.022	T	C	0.7125	0.0191	0.0188	0.0064	0.003357	41646	0.0188	0.7059
rs7772579	6	152042502	6q25.1	ESR1	60	A	C	0.721	0.031	0.003	6.40E-28	292718	5.30E-26	0.049	A	C	0.7108	0.0187	0.0188	0.0064	0.003207	41646	0.0188	0.7609
rs2934844	**6**	**166142456**	**6q27**	**PDE10A**	**61**	**T**	**A**	**0.672**	**0.021**	**0.003**	**1.80E-13**	**292253**	**1.30E-12**	**0.363**	**A**	**T**	**0.3151**	**0.0113**	**0.008**	**0.0066**	**0.2256**	**40780**	**-0.008**	**0.7102**
rs4719648	7	2756832	7p22.3	AMZ1/GNA12	62	C	T	0.577	0.019	0.003	2.60E-13	292711	1.90E-12	0.75	T	C	0.4212	0.0149	-0.0047	0.0059	0.4291	42212	0.0047	0.1451
rs59084784	7	22739562	7p15.3	IL6	63	A	C	0.323	0.017	0.003	2.40E-09	292716	9.10E-09	0.878	A	C	0.3138	0.0202	0.0016	0.0063	0.7938	42212	0.0016	0.4198
rs34776209	7	23513093	7p15.3	IGF2BP3	65	C	T	0.755	0.023	0.003	8.50E-15	292718	7.90E-14	0.419	T	C	0.2326	0.012	-0.0022	0.0069	0.7492	42212	0.0022	0.658
rs11983722	7	46298647	7p12.3	IGFBP3	70	A	T	0.938	0.032	0.005	3.10E-09	290622	1.20E-08	0.653	A	T	0.9375	0.0073	0.0135	0.0126	0.2835	40155	0.0135	0.1959
rs10265057	7	47275737	7p12.3	TNS3	71	G	A	0.092	0.027	0.004	1.30E-09	292446	5.10E-09	0.143	A	G	0.913	0.0047	-0.0076	0.0105	0.4712	41380	0.0076	0.8432
rs2237467	7	50733316	7p12.1	GRB10	72	A	G	0.221	0.018	0.003	5.30E-09	292710	1.90E-08	0.762	A	G	0.2336	0.015	0.0138	0.0069	0.04624	42212	0.0138	0.8739
rs112139215	7	73034559	7q11.23	MLXIPL	73	A	C	0.068	0.047	0.005	2.80E-20	295398	6.50E-19	0.144	A	C	0.0697	0.0044	0.0295	0.0132	0.02589	33204	0.0295	0.1779
rs2282978	7	92264410	7q21.2	CDK6	74	C	T	0.326	0.018	0.003	1.70E-11	298140	9.00E-11	0.498	T	C	0.6687	0.0107	-0.0094	0.0062	0.1271	42212	0.0094	0.4995
rs13231367	7	127509070	7q32.1	SND1	76	G	A	0.714	0.017	0.003	4.40E-09	292714	1.60E-08	0.339	A	G	0.2711	0.0156	-0.0166	0.0064	0.009717	42212	0.0166	0.08128
rs6467157	7	127660763	7q32.1	SND1	76	T	C	0.713	0.02	0.003	1.50E-11	292717	8.00E-11	0.426	T	C	0.7326	0.0213	0.0186	0.0066	0.004808	42212	0.0186	0.4907
rs62496903	8	6446938	8p23.1	MCPH1	78	T	C	0.083	0.033	0.005	6.70E-12	290687	3.90E-11	0.861	T	C	0.065	0.0111	0.0093	0.0138	0.5005	38723	0.0093	0.6399
rs732563	8	23345526	8p21.2	ENTPD4/NKX3-1	79	C	T	0.504	0.017	0.003	1.30E-11	292723	7.10E-11	0.288	T	C	0.4944	0.0159	-0.0087	0.0059	0.1401	42212	0.0087	0.05899
rs34036147	8	38366249	8p11.22	C8orf86/FGFR1	81	T	C	0.688	0.018	0.003	8.40E-11	292711	4.00E-10	0.239	T	C	0.6797	0.0095	0.0137	0.0064	0.03155	40780	0.0137	0.07201
rs13266210	**8**	**41533514**	**8p11.21**	**ANK1**	**82**	**A**	**G**	**0.786**	**0.027**	**0.003**	**1.50E-17**	**292718**	**2.30E-16**	**0.617**	**A**	**G**	**0.7994**	**0.0163**	**-0.0002**	**0.0073**	**0.9808**	**42212**	**-0.0002**	**0.7407**
rs72656010	8	57122215	8q12.1	PLAG1	83	T	C	0.868	0.028	0.004	1.40E-13	292713	1.00E-12	0.006	T	C	0.8703	0.0049	0.007	0.0089	0.4297	42212	0.007	0.74
rs7819593	8	106115172	8q22.3	ZFPM2	85	C	T	0.243	0.022	0.003	6.20E-13	292718	4.30E-12	0.427	T	C	0.7397	0.0373	-0.0082	0.0068	0.226	42212	0.0082	0.7987
rs10283100	**8**	**120596023**	**8q24.12**	**ENPP2**	**86**	**G**	**A**	**0.946**	**0.042**	**0.006**	**7.00E-13**	**291951**	**4.70E-12**	**0.704**	**A**	**G**	**0.0538**	**0.0042**	**0.0002**	**0.0141**	**0.9861**	**39948**	**-0.0002**	**0.6389**
rs13271368	8	126506140	8q24.13	TRIB1	87	C	T	0.761	0.02	0.003	2.30E-11	296867	1.20E-10	0.256	T	C	0.2421	0.0136	-0.0125	0.0068	0.06603	42212	0.0125	0.7174
rs13257363	8	142252580	8q24.3	SLC45A4	88	G	A	0.591	0.018	0.003	2.00E-11	292711	1.10E-10	0.278	A	G	0.4097	0.0091	-0.0164	0.006	0.006217	42212	0.0164	0.8867
rs7854962	9	96900505	9q22.32	PTPDC1	90	C	G	0.785	0.022	0.003	1.00E-11	292711	5.70E-11	0.808	C	G	0.7947	0.0114	0.0121	0.0075	0.1051	41646	0.0121	0.8577
rs28457693	9	98217348	9q22.32	PTCH1/FANCC	91	G	A	0.109	0.044	0.004	9.90E-26	288037	5.70E-24	0.431	A	G	0.8712	0.017	-0.0213	0.0093	0.02145	42212	0.0213	0.1364
rs1411424	9	113892963	9q31.3	LPAR1	92	A	G	0.523	0.02	0.003	1.50E-14	292717	1.40E-13	0.213	A	G	0.5205	0.0081	0.0107	0.0058	0.06509	42212	0.0107	0.8088
rs2418135	9	113901309	9q31.3	LPAR1	92	A	G	0.522	0.02	0.003	1.50E-14	292715	1.40E-13	0.229	A	G	0.5192	0.0073	0.0117	0.0058	0.04402	42212	0.0117	0.7831
rs1323438	9	119115531	9q33.1	PAPPA	94	C	T	0.718	0.019	0.003	5.60E-11	292712	2.80E-10	0.567	T	C	0.2892	0.0133	-0.0185	0.0065	0.004387	42212	0.0185	0.583
rs3933326	9	123633948	9q33.2	PHF19	95	G	A	0.676	0.021	0.003	2.30E-14	292715	2.00E-13	0.416	A	G	0.3401	0.0208	-0.0014	0.0062	0.8148	42212	0.0014	0.9397
rs10985827	9	125701608	9q33.2	RABGAP1/GPR21	96	G	T	0.141	0.03	0.004	6.10E-16	292715	6.80E-15	0.739	T	G	0.8471	0.0088	-0.0208	0.0086	0.01522	42212	0.0208	0.5464
rs4350272	10	25056118	10p12.1	ARHGAP21	98	A	G	0.269	0.017	0.003	3.60E-09	298133	1.30E-08	0.021	A	G	0.2621	0.0167	0.0139	0.0068	0.03989	41646	0.0139	0.8502
rs5030938	10	70975916	10q22.1	HKDC1/HK1	99	T	C	0.686	0.024	0.003	1.20E-17	292718	1.80E-16	0.918	T	C	0.702	0.0234	0.0121	0.0064	0.05681	42212	0.0121	0.459
rs9645500	10	70986723	10q22.1	HKDC1/HK1	99	G	T	0.694	0.024	0.003	1.80E-18	298136	3.00E-17	0.943	T	G	0.2911	0.0231	-0.01	0.0065	0.1245	42212	0.01	0.4566
rs1112718	10	94479107	10q23.33	HHEX/IDE	100	G	A	0.404	0.026	0.003	3.80E-23	298134	1.40E-21	0.4	A	G	0.5955	0.0118	-0.0205	0.0059	0.0005016	42212	0.0205	0.01178
rs3740360	**10**	**96025491**	**10q23.33**	**PLCE1**	**102**	**C**	**A**	**0.109**	**0.026**	**0.004**	**4.00E-10**	**292719**	**1.70E-09**	**0.902**	**A**	**C**	**0.9**	**0.0113**	**0.0068**	**0.0099**	**0.4914**	**41659**	**-0.0068**	**0.3157**
rs2274224	**10**	**96039597**	**10q23.33**	**PLCE1**	**103**	**C**	**G**	**0.434**	**0.021**	**0.003**	**9.80E-17**	**298132**	**1.30E-15**	**0.376**	**C**	**G**	**0.4146**	**0.0195**	**-0.0007**	**0.0059**	**0.9019**	**42212**	**-0.0007**	**0.9764**
rs10883846	**10**	**104958244**	**10q24.33**	**NT5C2/CYP17A1**	**105**	**C**	**T**	**0.615**	**0.017**	**0.003**	**1.30E-10**	**298138**	**6.00E-10**	**0.184**	**T**	**C**	**0.3741**	**0.0225**	**0.0009**	**0.006**	**0.8871**	**42212**	**-0.0009**	**0.1791**
rs7076938	10	115789375	10q25.3	ADRB1	107	T	C	0.735	0.032	0.003	2.10E-28	298136	1.90E-26	0.005	T	C	0.7349	0.0073	0.023	0.0066	0.0005276	42212	0.023	0.2864
rs1801253	10	115805056	10q25.3	ADRB1	107	C	G	0.727	0.031	0.003	1.40E-25	297700	7.60E-24	0.117	C	G	0.7362	0.0086	0.0236	0.0067	0.0004468	42212	0.0236	0.3227
rs71486610	10	124134803	10q26.13	PLEKHA1	108	C	G	0.477	0.02	0.003	3.20E-15	292714	3.20E-14	0.517	C	G	0.4898	0.0289	0.0123	0.0058	0.03337	42212	0.0123	0.01124
rs11042596	11	2118860	11p15.5	INS-IGF2	109	T	G	0.336	0.027	0.003	4.30E-22	292715	1.30E-20	0.944	T	G	0.331	0.0101	0.0165	0.0065	0.01097	40780	0.0165	0.3428
rs234864	11	2857297	11p15.4	KCNQ1	110	A	G	0.547	0.016	0.003	1.70E-09	296865	6.40E-09	0.837	A	G	0.5441	0.0211	0.0062	0.0059	0.2952	42212	0.0062	0.3422
rs4444073	**11**	**10331664**	**11p15.4**	**ADM**	**112**	**A**	**C**	**0.52**	**0.02**	**0.003**	**2.70E-15**	**298137**	**2.70E-14**	**0.708**	**A**	**C**	**0.5296**	**0.0122**	**-0.0004**	**0.0059**	**0.9474**	**42212**	**-0.0004**	**0.377**
rs5030317	**11**	**32410337**	**11p13**	**WT1**	**113**	**C**	**G**	**0.733**	**0.017**	**0.003**	**2.70E-09**	**292715**	**1.00E-08**	**0.339**	**C**	**G**	**0.7386**	**0.0125**	**-0.0027**	**0.0066**	**0.6809**	**42212**	**-0.0027**	**0.3249**
rs667515	11	69449076	11q13.3	CCND1	116	G	C	0.618	0.018	0.003	9.30E-12	292266	5.20E-11	0.287	C	G	0.3705	0.0123	-0.0044	0.0061	0.477	41646	0.0044	0.4453
rs61885091	11	69791952	11q13.3	ANO1/FGF4	117	A	G	0.169	0.023	0.004	4.80E-10	277677	2.00E-09	0.718	A	G	0.1562	0.0305	0.0055	0.0097	0.5676	40227	0.0055	0.5875
rs10830963*	11	92708710	11q14.3	MTNR1B	118	G	C	0.277	0.019	0.003	2.80E-11	298126	1.40E-10	0.045	C	G	0.723	0.0145	-0.0165	0.007	0.01775	40780	0.0165	0.427
rs76895963	12	4384844	12p13.32	CCND2	120	G	T	0.021	0.076	0.01	2.50E-13	278956	1.80E-12	0.719	T	G	0.982	0.0035	-0.0291	0.0278	0.2948	38723	0.0291	0.7545
rs11055030	12	12878349	12p13.1	APOLD1	121	G	C	0.718	0.02	0.003	3.90E-12	292715	2.30E-11	0.113	C	G	0.2718	0.0147	-0.006	0.0066	0.359	42212	0.006	0.9092
rs2306547	12	26877885	12p11.23	ITPR2	122	C	T	0.534	0.019	0.003	4.40E-13	292721	3.00E-12	0.874	T	C	0.4663	0.0089	-0.0248	0.0058	2.04E-05	42212	0.0248	0.678
rs6582623	12	46613394	12q13.11	SLC38A1	124	C	T	0.869	0.024	0.004	1.10E-09	292715	4.30E-09	0.75	T	C	0.129	0.0117	-0.026	0.009	0.003815	41659	0.026	0.4713
rs8756	12	66359752	12q14.3	HMGA2	126	C	A	0.487	0.041	0.003	2.40E-59	298139	3.90E-55	0.077	A	C	0.4949	0.0179	-0.0248	0.0058	1.66E-05	42212	0.0248	0.09682
rs7968682	12	66371880	12q14.3	HMGA2	126	G	T	0.486	0.042	0.003	4.20E-60	298092	7.80E-56	0.083	T	G	0.4956	0.0194	-0.0246	0.0058	2.03E-05	42212	0.0246	0.112
rs1480470	12	66412130	12q14.3	HMGA2	127	G	A	0.631	0.024	0.003	1.40E-19	292712	2.80E-18	0.712	A	G	0.3705	0.0094	-0.0208	0.0061	0.0007226	42212	0.0208	0.1547
rs2647873	12	103081192	12q23.2	LINC00485/IGF1	129	A	G	0.52	0.018	0.003	2.90E-12	292715	1.80E-11	0.228	A	G	0.5115	0.0165	0.0159	0.0059	0.007086	42212	0.0159	0.5116
rs3184504	12	111884608	12q24.12	SH2B3	131	C	T	0.521	0.023	0.003	2.60E-19	296867	5.00E-18	0.011	T	C	0.466	0.0158	-0.0161	0.0058	0.005808	42212	0.0161	0.6028
rs9549046	13	40647206	13q14.11	LINC00332	132	A	G	0.118	0.029	0.004	8.00E-13	291448	5.30E-12	0.584	A	G	0.1286	0.0188	0.01	0.0089	0.2645	42212	0.01	0.9669
rs34217484	13	48854550	13q14.2	LINC00441/RB1	133	A	T	0.264	0.019	0.003	6.80E-11	287438	3.30E-10	0.477	A	T	0.2759	0.0123	0.0069	0.0066	0.2931	42212	0.0069	0.5238

rs9318511	13	78601413	13q22.3	LINC00446	134	C	A	0.873	0.027	0.004	6.00E-12	292266	3.50E-11	0.058	A	C	0.1142	0.0091	-0.0103	0.0092	0.2632	42212	0.0103	0.002236
rs6575803	14	101257755	14q32.2	MIR2392/DLK1	136	C	T	0.895	0.032	0.004	1.30E-12	284076	8.40E-12	0.317	T	C	0.1031	0.0111	-0.0318	0.0103	0.00198	39948	0.0318	0.8334
rs75844534	15	38667117	15q14	SPRED1	137	A	C	0.124	0.026	0.004	4.90E-11	292715	2.40E-10	0.649	A	C	0.1178	0.0084	0.001	0.0094	0.9123	41646	0.001	0.1405
rs339969	15	60883281	15q22.2	RORA	139	A	C	0.619	0.017	0.003	2.20E-10	292719	9.80E-10	0.689	A	C	0.6278	0.0195	0.016	0.006	0.007721	42212	0.016	0.5874
rs7183988	15	91428589	15q26.1	FES/FURIN	143	G	T	0.529	0.018	0.003	1.70E-12	294939	1.10E-11	0.676	T	G	0.4659	0.0131	-0.0181	0.0059	0.002199	42212	0.0181	0.9017
rs4932373	15	91429287	15q26.1	FES/FURIN	143	A	C	0.68	0.02	0.003	3.00E-13	295749	2.20E-12	0.851	A	C	0.6809	0.0151	0.0223	0.0064	0.0004698	42212	0.0223	0.7967
rs55958435	15	96852638	15q26.2	NR2F2	144	A	G	0.748	0.025	0.003	1.60E-16	292710	1.90E-15	0.693	A	G	0.7342	0.0197	0.0136	0.0068	0.04625	42212	0.0136	0.6375
rs7402983	15	99193276	15q26.3	IGF1R	145	A	C	0.405	0.024	0.003	2.60E-19	292717	5.20E-18	0.986	A	C	0.4101	0.015	0.0042	0.0062	0.5022	40780	0.0042	0.09049
rs2045457	16	20046115	16p12.3	GPR139/GPRC5B	147	G	A	0.311	0.016	0.003	6.30E-09	292716	2.20E-08	0.051	A	G	0.6692	0.021	-0.0164	0.0063	0.009368	42212	0.0164	0.9936
rs40434	16	55699525	16q12.2	SLC6A2	148	G	A	0.391	0.017	0.003	3.00E-10	292714	1.30E-09	0.798	A	G	0.5855	0.0251	-0.0085	0.0059	0.1531	42212	0.0085	0.6969
rs222857	17	7164563	17p13.1	CLDN7/SLC2A4	151	T	C	0.575	0.026	0.003	1.10E-24	298132	5.10E-23	0.131	T	C	0.5811	0.0125	0.0203	0.006	0.0006773	42212	0.0203	0.2358
rs2428362,Ã†	17	7180274	17p13.1	CLDN7/SLC2A4	151	T	C	0.576	0.025	0.003	1.80E-22	292709	6.40E-21	0.152	T	C	0.5846	0.0126	0.0194	0.0063	0.002066	35997	0.0194	0.6122
rs4511593	17	7455536	17p13.1	TNFSF12-TNFSF13	152	T	C	0.65	0.017	0.003	1.10E-10	292717	5.10E-10	0.847	T	C	0.6547	0.0086	0.0115	0.0061	0.05824	42212	0.0115	0.2133
rs78378222	17	7571752	17p13.1	TP53	153	G	T	0.013	0.079	0.012	1.80E-11	287415	9.80E-11	0.095	T	G	0.9836	0.0031	-0.0702	0.0254	0.005718	38723	0.0702	0.8742
rs9909342	17	25652275	17q11.1	WSB1	154	A	G	0.381	0.018	0.003	2.20E-11	292713	1.10E-10	0.334	A	G	0.3835	0.0172	0.015	0.006	0.01281	42212	0.015	0.4503
rs7223535	17	29211667	17q11.2	ATAD5	155	G	A	0.732	0.021	0.003	2.10E-13	292715	1.60E-12	0.984	A	G	0.2584	0.0104	-0.0098	0.0066	0.1391	42212	0.0098	0.1548
rs11867479	17	68090207	17q24.3	KCNJ16	156	T	C	0.353	0.017	0.003	1.10E-10	298138	5.10E-10	0.2	T	C	0.351	0.0146	0.0137	0.006	0.02358	42212	0.0137	0.5333
rs10221267	17	68464662	17q24.3	KCNJ2	157	T	C	0.512	0.017	0.003	6.50E-11	296641	3.20E-10	0.314	T	C	0.4883	0.0388	0.0049	0.0059	0.4015	42212	0.0049	0.9113
rs11082304	18	20720973	18q11.2	CABLES1	160	T	G	0.508	0.016	0.003	4.20E-10	296792	1.80E-09	0.035	T	G	0.5042	0.0173	0.0104	0.0058	0.07352	42212	0.0104	0.3975
rs2779165	19	4915447	19p13.3	UHRF1	161	G	C	0.184	0.022	0.003	7.60E-11	291447	3.70E-10	0.34	C	G	0.8318	0.0084	-0.0167	0.0082	0.04272	40780	0.0167	0.8018
rs8106042	19	7161849	19p13.2	INSR	162	G	C	0.281	0.02	0.003	2.20E-12	291451	1.40E-11	0.691	C	G	0.721	0.0148	-0.0163	0.007	0.02011	40780	0.0163	0.9251
rs41355649	**19**	**33790556**	**19q13.11**	**CEBPA**	**164**	**G**	**A**	**0.934**	**0.034**	**0.005**	**1.20E-10**	**291155**	**5.50E-10**	**0.911**	**A**	**G**	**0.0585**	**0.01**	**0.001**	**0.0133**	**0.9411**	**38723**	**-0.001**	**0.01458**
rs1129156	19	40719076	19q13.2	MAP3K10/AKT2	165	T	C	0.268	0.017	0.003	2.50E-09	292719	9.50E-09	0.188	T	C	0.2714	0.0092	0.0036	0.0067	0.5911	42212	0.0036	0.4687
rs516246	19	49206172	19q13.33	FUT2	167	C	T	0.506	0.018	0.003	9.30E-12	295749	5.20E-11	0.285	T	C	0.4691	0.0321	-0.0156	0.0059	0.007668	42212	0.0156	0.8071
rs255773	19	54723546	19q13.42	LILRB3/RPS9	168	C	T	0.536	0.018	0.003	1.30E-11	288702	7.30E-11	0.945	T	C	0.4467	0.0305	-0.0245	0.0066	0.000223	40780	0.0245	0.4291
rs147110934	19	55993436	19q13.42	ZNF628	169	G	T	0.975	0.052	0.009	1.60E-09	276061	6.30E-09	0.299	T	G	0.0184	0.004	-0.0598	0.0247	0.01545	31772	0.0598	0.3105
rs6040076	20	10658882	20p12.2	JAG1	172	C	G	0.5	0.019	0.003	4.40E-13	292711	3.00E-12	0.407	C	G	0.5094	0.0173	0.0076	0.0062	0.2225	40780	0.0076	0.7974
rs6033062	20	11207419	20p12.2	LOC339593	173	A	T	0.46	0.016	0.003	5.20E-10	292717	2.20E-09	0.859	A	T	0.4541	0.0107	0.0072	0.0059	0.2206	42212	0.0072	0.1537
rs1203876	20	22540915	20p11.21	LINC00261/FOXA2	174	C	A	0.046	0.038	0.006	9.40E-10	291539	3.80E-09	0.843	A	C	0.95	0.008	-0.0175	0.0149	0.2414	38723	0.0175	0.8145
rs11698914	20	31327144	20q11.21	COMMD7	175	C	G	0.233	0.032	0.003	1.20E-24	292713	5.80E-23	0.91	C	G	0.2463	0.0182	0.0234	0.0069	0.0007618	41646	0.0234	0.2195
rs2889874	20	33715777	20q11.22	EDEM2/MYH7B	177	G	T	0.452	0.016	0.003	9.40E-10	292712	3.80E-09	0.731	T	G	0.5653	0.0165	-0.009	0.006	0.1332	41646	0.009	0.06341
rs1012167	20	39159119	20q12	MAFB	178	C	T	0.401	0.024	0.003	1.20E-19	292373	2.50E-18	0.541	T	C	0.6172	0.0234	-0.0245	0.006	4.70E-05	42212	0.0245	0.2864
rs753381	20	39797465	20q12	PLCG1	179	T	C	0.451	0.015	0.003	3.40E-09	297797	1.30E-08	0.031	T	C	0.4521	0.01	0.0018	0.0059	0.7543	42212	0.0018	0.508
rs6026449	20	57272617	20q13.32	STX16-NPEPL1/GNAS	180	C	T	0.627	0.017	0.003	2.50E-10	292375	1.10E-09	0.387	T	C	0.383	0.0155	-0.0012	0.0061	0.8392	42212	0.0012	0.6036
rs73143584	20	62445702	20q13.33	ZBTB46	181	A	G	0.11	0.029	0.004	1.80E-11	286584	9.60E-11	0.719	A	G	0.0971	0.0136	0.0068	0.0109	0.5324	40814	0.0068	0.7021
rs2229742	21	16339172	21q11.2	NRIP1	182	G	C	0.881	0.027	0.004	7.40E-11	297794	3.60E-10	0.095	C	G	0.1016	0.009	-0.0322	0.01	0.001319	42212	0.0322	0.8739
rs220193	21	43581308	21q22.3	UMODL1	183	A	G	0.225	0.021	0.003	4.10E-11	292712	2.10E-10	0.94	A	G	0.2377	0.0239	0.0027	0.007	0.6997	42212	0.0027	0.6395
rs134594	22	29468456	22q12.1	KREMEN1	184	C	T	0.351	0.017	0.003	5.80E-10	290627	2.40E-09	0.227	T	C	0.6498	0.0051	-0.0053	0.0062	0.3908	42212	0.0053	0.6698
rs41311445	22	42070374	22q13.2	NHP2L1/SREBF2	185	A	C	0.903	0.033	0.004	3.30E-13	289016	2.40E-12	0.024	A	C	0.9064	0.0089	0.0112	0.0118	0.3387	35718	0.0112	0.4134
rs7285579	22	46441980	22q13.31	LOC100271722	186	C	T	0.698	0.017	0.003	2.70E-09	290177	1.00E-08	0.23	T	C	0.2815	0.0205	-0.0168	0.0081	0.03737	35345	0.0168	0.2395

[a] *Name of nearest gene followed by any additional genes that had functional evidence for potentially being the causal gene. As defined in Supplemental Table 5a of Warrington et al., 2019, from the GWAS of own birth weight*

[b] *Signal number of each of the 209 SNPs. As defined in Supplemental Table 5a of Warrington et al., 2019, from the GWAS of own birth weight.*

[c] *EAF = effect allele frequency; SE = standard error*

[d] *Allele1 = the first allele for this marker in the first file where it occurs*

[e] *Allele2 = the second allele for this marker in the first file where it occurs*

[f] *Freq1 = weighted average of frequency for allele 1 across all studies; FreqSE = corresponding standard error for allele frequency estimate*

[g] *Effect = overall estimated effect size for allele1; StdErr = overall standard error for effect size estimate*

[h] *AdjEffect = estimated effect size after alignment to effect allele in Warrington et al., 2019. If alleles were swapped, Effect * -1 was used to calculated AdjEffect.*

[i] *HetPVal = P-value for heterogeneity statistic*

Supplementary Table 5.2 - Genetic correlations with BW in twins. Traits are sorted by category and descending r_g.

Trait	PMID[a]	Category	r_g[b]	SE[c]	z[d]	p[e]
Child birth weight (Horikoshi 2013)	23202124	Anthropometric	0.98	0.18	5.38	7.34E-08
UK Biobank birth weight (Field 20022)		Anthropometric	0.95	0.12	7.64	2.19E-14
Offspring birth weight (maternal effect) adjusted for offspring genotype	31043758	Anthropometric	0.92	0.13	6.88	6.15E-12
Own birth weight (Warrington 2019)	31043758	Anthropometric	0.92	0.13	7.03	2.07E-12
Birth weight (Horikoshi 2016)	27680694	Anthropometric	0.91	0.13	6.79	1.12E-11
Offspring birth weight (Warrington 2019)	31043758	Anthropometric	0.76	0.14	5.51	3.62E-08
Own birth weight (fetal effect) adjusted for maternal genotype	31043758	Anthropometric	0.69	0.12	5.71	1.11E-08
Child birth length	25281659	Anthropometric	0.57	0.13	4.21	2.52E-05
Extreme height	23563607	Anthropometric	0.38	0.1	3.92	8.92E-05
Height	20881960	Anthropometric	0.35	0.08	4.34	1.43E-05
Infant head circumference	22504419	Anthropometric	0.34	0.16	2.1	0.0358
Height; Females at age 10 and males at age 12	23449627	Anthropometric	0.33	0.12	2.73	0.0062
Hip circumference	25673412	Anthropometric	0.32	0.08	4.23	2.34E-05
Obesity class 3	23563607	Anthropometric	0.29	0.12	2.38	0.0173
Childhood obesity	22484627	Anthropometric	0.27	0.11	2.46	0.0141
Waist circumference	25673412	Anthropometric	0.26	0.07	3.56	0.0004
Overweight	23563607	Anthropometric	0.18	0.07	2.54	0.0112
Body mass index	20935630	Anthropometric	0.18	0.07	2.68	0.0074
Obesity class 2	23563607	Anthropometric	0.14	0.08	1.84	0.0652
Obesity class 1	23563607	Anthropometric	0.14	0.07	2.11	0.0352
Extreme bmi	23563607	Anthropometric	0.06	0.1	0.63	0.529
Difference in height between adolescence and adulthood; age 14	23449627	Anthropometric	0.05	0.17	0.31	0.756
Waist-to-hip ratio	25673412	Anthropometric	0.01	0.08	0.18	0.854
Sitting height ratio	25865494	Anthropometric	0.01	0.15	0.09	0.925
Difference in height between childhood and adulthood; age 8	23449627	Anthropometric	-0.02	0.15	-0.12	0.905
Extreme waist-to-hip ratio	23563607	Anthropometric	-0.16	0.16	-1.02	0.309
Crohns disease	26192919	Autoimmune	0.16	0.1	1.58	0.114
Rheumatoid Arthritis	24390342	Autoimmune	0.11	0.11	1.02	0.307
Inflammatory Bowel Disease (Euro)	26192919	Autoimmune	0.09	0.1	0.97	0.332
Ulcerative colitis	26192919	Autoimmune	0.08	0.11	0.7	0.483

Supplementary Table 5.2 - Genetic correlations with BW in twins. Traits are sorted by category and descending r_g. (continued)

Trait	PMID[a]	Category	r_g[b]	SE[c]	z[d]	p[e]
Celiac disease	20190752	Autoimmune	0.07	0.15	0.46	0.646
Primary biliary cirrhosis	26394269	Autoimmune	0.01	0.12	0.05	0.961
Systemic lupus erythematosus	26502338	Autoimmune	-0.1	0.13	-0.77	0.44
Asthma	17611496	Autoimmune	-0.34	0.15	-2.25	0.0247
Intelligence	28530673	Cognitive	0.2	0.09	2.21	0.0268
Type 2 Diabetes	22885922	Glycemic	-0.02	0.1	-0.21	0.833
HOMA-B[f]	20081858	Glycemic	-0.08	0.15	-0.51	0.609
HOMA-IR[g]	20081858	Glycemic	-0.09	0.17	-0.53	0.594
Fasting glucose main effect	22581228	Glycemic	-0.1	0.11	-0.94	0.349
Fasting insulin main effect	22581228	Glycemic	-0.11	0.14	-0.73	0.463
HbA1C[h]	20858683	Glycemic	-0.23	0.14	-1.72	0.0846
Bipolar disorder	21926972	Psychiatric	0.13	0.1	1.31	0.189
Depressive symptoms	27089181	Psychiatric	0.13	0.09	1.39	0.163
PGC[i] cross-disorder analysis	23453885	Psychiatric	0.06	0.09	0.64	0.521
Autism spectrum disorder	0	Psychiatric	0.04	0.12	0.3	0.764
Subjective well being	27089181	Psychiatric	0.01	0.1	0.15	0.882
Major depressive disorder	22472876	Psychiatric	-0.1	0.15	-0.67	0.505
Anorexia Nervosa	24514567	Psychiatric	-0.13	0.08	-1.59	0.113
Number of children ever born	27798627	Reproductive	0.12	0.1	1.24	0.214
Age of first birth	27798627	Reproductive	0.08	0.08	1.01	0.313
Age at Menopause	26414677	Reproductive	0.01	0.1	0.08	0.936
Age at Menarche	25231870	Reproductive	-0.04	0.07	-0.59	0.552
Cigarettes smoked per day	20418890	Smoking behavior	0.23	0.18	1.28	0.202
Smoking Initiation	30617275	Smoking behavior	0.08	0.15	0.5	0.616
Former vs Current smoker	20418890	Smoking behavior	0.01	0.17	0.04	0.969
Age of smoking initiation	20418890	Smoking behavior	-0.03	0.19	-0.15	0.88
Ever vs never smoked	20418890	Smoking behavior	-0.06	0.11	-0.56	0.576

[a] *PubMed reference number*
[b] *Genetic correlation*
[c] *Standard error of r_g*
[d] *Z-score*
[e] *P-value*
[f] *Homeostatic model assessment of beta cell function*
[g] *Homeostatic model assessment of insulin resistance*
[h] *HemoglobinA1C*
[i] *Pyschiatric Genetics Consortium*

Supplementary Table 5.3 - Genotyping information per cohort

		Sample QC[a]		SNP QC[a]				Phasing and Imputation	
Cohort	**Genotyping Arrays**	**Call Rate**	**Additional Filters**	**Call Rate**	**HWE[b]**	**MAF[c]**	**Additional Filters**	**Reference Panel(s)**	**Imputation Software**
AVERA	Illumina GSA	95%	Gender mismatches, Relatedness, Heterozygosity Plink -0.10 <= F >= 0.10	95%	$p<10^{-4}$	<0.10	NA	1000GP3	MINIMAC3
CATSS	Illumina Infinium PsychArray-24 BeadChip	98%	Heterozygosity F> +/-0.2, >6sd excessive relatedness, sex violations, non-European (>6sd from mean of the first two principal components in 1000GP3)	98%	$p<10^{-6}$	<0.01	Markers showing discordances among 37 cross-batch duplicate samples, markers with more than one discordant genotype among 84 pairs of MZ twins, y chromosome and mitochondrial markers with poor variant calling, markers associated with genotyping batch p<5x10-8.	1000GP3	MINIMAC3
DTR	Illumina Infinium PsychArray	99%	Gender mismatch, relatedness	98%	$p<10^{-6}$	<0.05	NA	1000GP3	IMPUTE2 version 2.3.2

Supplementary Table 5.3 - Genotyping information per cohort (continued)

		Sample QC[a]		SNP QC[a]				Phasing and Imputation	
FinnTwin	Illumina Human670-QuadCustom v1.0 A, Human610-Quad v1.0B, HumanCoreExome-12 v1.0B, v1.1A, HumanCoreExome-24 v1.0A	95%	Gender mismatches, Heterozygosity PLINK -0.03 < F < 0.05	95%	$p<10^{-6}$	<0.01	MDS PCA outlier exclusions	1000GP3	MINIMAC3
NTR	Perlegen Affymetrix, Affymetrix 6.0, Affymetrix Axiom, Illumina 660K, Illumina 1M, Illumina GSA, GoNL Sequence platform	90%	Gender mismatches, Relatedness, Heterozygosity Plink -0.10 <= F >= 0.10	95%	$p<10^{-5}$	<0.01	Mendel error => 20, Palindromic SNPs with MAF .40-.50	1000GP3	MINIMAC3
QIMR	Illumina 317K, 370K, 610K, 660K, CoreExome, PyschArray, Omni2.5	97%	Gender mismatches, sample mixups (if not able to be corrected)	95%	$p<10^{-6}$	<0.01	Mean GC score < 0.7, one distinct genotype	1000GP3	MINIMAC3
TEDS	AffymetrixGeneChip 6.0, HumanOmniExpressExome-8v1.2	99%	Gender mismatches, relatedness, heterozygosity, non-European ancestry	98%	$p<10^{-5}$	<0.01	Association with platform, batch, or plate if $p<10^{-3}$	HRC r1	MINIMAC3
UKB	UK BiLEVE Array by Affymetrix, the Applied Biosystems UK Biobank Axiom Array	95%	Relatedness, non-UK ancestry, gender mismatches	95%	$p<10^{-6}$	<0.005	Association with platform, batch, or plate if $p<10^{-12}$, Difference of >= .1 MAF compared to MAF in UK10K	1000GP3 + UK10K	IMPUTE2

[a] *Quality control*
[b] *Hardy-Weinberg equilibrium*
[c] *Minor allele frequency*

Supplementary Table 5.4 - Ethics statements, funding, and acknowledgements.

Cohort	Ethics Statement	Funding	Acknowledgments
AVERA	The study was approved by the Avera Institutional Review Board and the Avera Department of Human Subject's protection.	The Avera Twin Register is supported by Avera Health, Avera McKennan Hospital and Avera Institute for Human Genetics. The collaboration between the Netherlands and the Avera Twin Regsiter arose through NIHM Grant: 1RC2MH089995-01: Genomics of Developmental Trajectories in Twins.	We would like to extend our gratitude to all of the enrolled participants of the Avera Twin Register.
CATSS	The study is approved by the Regional Ethical Review Board in Stockholm, Sweden and all participants gave informed consent.	CATSS is part of the Swedish Twin Registry which is managed by Karolinska Institutet and receives funding through the Swedish Research Council under the grant no 2017-00641.	We wish to thank the Biobank at Karolinska Institutet for professional biobank service. The local computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppmax partially funded by the Swedish Research Council through grant agreement no. 2018-05973
DTR	Written informed consents were obtained from all participants. Collection and use of biological material and survey information were approved by the Regional Scientific Ethical Committees for Southern Denmark, and the study was approved by the Danish Data Protection Agency.	The DTR data collection is supported by grants from The National Program for Research Infrastructure 2007 from the Danish Agency for Science, Technology and Innovation and the US National Institutes of Health (P01 AG08761). Genotyping was supported by NIH R01 AG037985 (Pedersen).	Genotyping was conducted by the SNP&SEQ Technology Platform, Science for Life Laboratory, Uppsala, Sweden (http://snpseq.medsci.uu.se/genotyping/snp-services/).
FinnTwin	The FTC data collection has been approved by the ethics committees of the University of Helsinki (113/E3/01 and 346/E0/05) and Helsinki University Central Hospital (270/13/03/01/2008 and 154/13/03/00/2011). Written informed consent was provided by the participants before the sample collection.	Phenotyping and genotyping of the Finnish twin cohorts was supported by the Academy of Finland Center of Excellence in Complex Disease Genetics (grants 213506, 129680), the Academy of Finland (grants 10049, 205585, 118555, 141054, 265240, 263278 and 264146 to J. Kaprio), National Institute of Alcohol Abuse and Alcoholism (grants AA-12502, AA-00145, and AA-09203 to R. J. Rose and AA15416 and K02AA018755 to D. M. Dick), and the Wellcome Trust Sanger Institute, UK.	We warmly thank the participating twin pairs for their contribution. Anja Happola is acknowledged for her valuable contribution in recruitment and data collection.

Supplementary Table 5.4 - Ethics statements, funding, and acknowledgements. (continued)

Cohort	Ethics Statement	Funding	Acknowledgments
NTR	The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the US Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance - FWA00017598; IRB/institute codes, NTR 03-0180).	Funding was provided by ZonMw (Grant Nos. 904-61-090, 985-10-002, 912-10-020, 904-61-193, 480-04-004, 463-06-001, 451-04-034, 400-05-717, 016-115-035, 481-08-011 and 056-32-010), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Grant Nos. Addiction-31160008, NWO-Middelgroot-911-09-032, OCW_NWO Gravity program –024.001.003, NWO-Groot 480-15-001/674, NWO-56-464-14192), Centre for Medical Systems Biology (CSMB, NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (Grant Nos. 184.021.007, 184.033.111), Koninklijke Nederlandse Akademie van Wetenschappen (NL) (Grant No. PAH/6635), European Science Foundation (Grant No. EU/QLRT-2001-01254), FP7 Health (Grant Nos. 01413: ENGAGE, 602768: ACTION), H2020 European Research Council (Grant Nos. ERC AG 230374, ERC SG 284167, ERC CG 771057), National Institutes of Health (Grant No. NIH R01 DK092127-04) and Avera Institute for Human Genetics. Bart Baselmans: NWO/ZonMw: Rubicon 45219101.	We would like to thank all members of twin families registered with the Netherlands Twin Register for their continued support of scientific research.
QIMR	The study was approved by the QIMR Human Research Ethics Committee.	Funding for data collection and/or genotyping was provided by the Australian National Health and Medical Research Council (NHMRC); the Australian Research Council (ARC); the FP-5 GenomEUtwin Project; the US National Institutes of Health (NIH); and the Center for Inherited Disease Research (CIDR; Baltimore, MD, USA).	We thank the twins for their cooperation.

Supplementary Table 5.4 - Ethics statements, funding, and acknowledgements. (continued)

Cohort	Ethics Statement	Funding	Acknowledgments
TEDS	The study was approved by King's College London Ethics Committee (PNM/09/10-104 Twins Early Development Study).	TEDS is supported by a program grant to RP from the UK Medical Research Council (MR/M021475/1 and previously G0901245), with additional support from the US National Institutes of Health (AG046938). The research leading to these results has also received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ grant agreement n° 602768 and ERC grant agreement n° 295366. RP is supported by a Medical Research Council Professorship award (G19/2). High performance computing facilities were funded with capital equipment grants from the GSTT Charity (TR130505) and Maudsley Charity (980).	We gratefully acknowledge the ongoing contribution of the participants in the Twins Early Development Study (TEDS) and their families.
UKB			This research has been conducted using the UK Biobank Resource under Application Number 25472.

6

INFERENCE OF GENETIC ANCESTRY: EVALUATION WITHIN FAMILIES AND ACROSS GENOTYPING ARRAYS

Based on
Beck, J.J., Ahmed, T., Finnicum, C.T., Zwinderman, K., Ehli, E.A., Boomsma, D.I., Hottenga, J.J. (2021)
Inference of genetic ancestry: evaluation within families and across genotyping arrays. Submitted for publication.

ABSTRACT

Inference of genetic ancestry is an essential aspect of population-based association studies to account for population heterogeneity and structure. A key question is how ancestry estimates compare when family members participate in a study and when genetic data are sourced from multiple genotyping arrays. In this paper, we analyze genome-wide SNP data to compare genetic ancestry estimates between pairs of family members across the spectrum of relatedness, from independently genotyped identical twins through to unrelated parent pairs, and between individuals genotyped on multiple arrays. Genetic ancestry estimates were obtained utilizing two conventionally performed tests, principal component analysis (PCA) and a model-based approach exemplified by the software ADMIXTURE. We discover that Euclidean distances of genetic ancestry estimates between pairs of family members are inversely related to the degree of genetic relatedness between them irrespective of estimation method and genotyping array, confirming that ancestry estimates are more similar in closely related individuals. Ancestry estimates of the same individuals genotyped across arrays were nearly indistinguishable, and we attribute the slight differences to the array-dependent variation in SNPs used for calculation. We also explore if non-identical twin offspring of ancestrally diverse parents exhibit more appreciable differences in ancestry than those with ancestrally similar parents. We uncover that ancestry estimates in offspring of more diverse parents are not considerably different than those with ancestry-similar parents. This study demonstrates the utility and robustness of current tools used to infer genetic ancestry, PCA and ADMIXTURE, even when considering the confounders of relatedness and genotyping array.

Keywords: Within-family analysis, genetic ancestry estimation, population structure, principal components analysis (PCA), ADMIXTURE

INTRODUCTION

Genetic association studies have become an effective research tool for identifying genetic loci related to complex phenotypes and diseases [1]. A fundamental step of performing genetic association studies is the detection of and correction for population structure. In this paper, we focus on population structure created by ancestry divergence and its detection based on genotype data. In general, strategies for estimating global ancestry can be categorized into two broad groups: algorithmic and model-based approaches. Commonly employed, each method has been shown to provide reliable inferences of genetic ancestry in unrelated individuals and to elucidate population structure from genome-wide data [2].

Algorithmic methods are exemplified by cluster analysis and principal component analysis (PCA). Generally, PCA is a method for obtaining low-dimensional summaries of high-dimensional data, increasing interpretability while minimizing information loss. In genetic datasets, PCA is performed to identify systematic variation amongst individuals' genotypes. In this context, a large set of variables (individuals' genotypes) are transformed into a smaller group of uncorrelated variables, called principal components (PCs), usually with the constraint that each PC successively captures less variation in the original data. PCA of genotypic data yields a series of scores per individual corresponding to the values of these PCs. Top PCs calculated from genetic data typically reflect population structure, allowing inferences of genetic ancestry. Over the years, PCA has demonstrated its utility for elucidating genetic ancestry from seemingly unrelated samples [2], correcting for confounding due population structure [2, 3], and understanding population ancestry composition and migration [4-6].

Best practices for implementing PCA have been suggested [7], but applying PCA in genetic analysis to capture population structure is not without challenges. Care must be taken to ensure that PCs are unbiased and reflect variation in ancestry and not some other form of systematic variation present within the data. Rather than capturing population structure, some PCs may reflect linkage disequilibrium (LD) structure. If PCs capturing LD are included as covariates in analyses, the power for detection in association studies is reduced [6, 8-10]. The degree of population structure captured by PCA may also be diminished by the presence of outlier samples reflecting batch effects or family structure. Therefore, commonly employed steps in PCA include determining unrelated individuals, pruning genetic markers in LD, and excluding outlier samples that may be indicative of poor genotyping quality.

Model-based approaches, such as those embodied by the programs STRUCTURE [11], fastSTRUCTURE [12], FRAPPE [13], and ADMIXTURE [14], present alternative methods for elucidation of population structure. These approaches provide relative proportions of ancestry, given that the genetic composition of an individual is a mosaic of the ancestral populations they represent. In general, model-based methods estimate global individual ancestry proportions based on parameterized statistical models. Commonly, these techniques take Bayesian or maximum likelihood estimation approaches to optimize the probability of observed genotypes by alternatively updating ancestry coefficient and population allele frequency matrices. The resulting individual ancestry proportions are more directly interpretable than PCs and can account for population composition in a study.

The two types of methods appear to have little in common at the surface due to underlying analytical differences. One involves the explicit definition of a model, while the other does not. A link between the approaches has been investigated, and strategies for identifying admixture proportions from PCs of PCA have been suggested [15-18]. In this context, ancestry proportions interpreted from PCA and the results of model-based approaches, like ADMIXTURE, are consistent [19, 20]. Given the congruent results, the main interest of the current paper was to compare ancestry estimates in realistic situations, where individuals in large studies have genome-wide data from different genotyping arrays. This situation arises in large cohort studies, where successive generations of genotyping arrays were applied across time. Our second interest involves study designs incorporating family members. Within families, siblings with the same biological parents necessarily should be assigned the same ancestry, even when genotyped across different arrays. Here, families with parents from two different ancestry backgrounds are of special interest. We compare algorithmic and model-based approaches to obtain ancestry estimates as a function of genomic relatedness within families, across pairs of individuals from monozygotic twins to nominally unrelated parent pairs, and within and across genotyping arrays.

One strategy for mitigating concerns of population structure in genetic association studies is to employ a family-based design [21, 22]. These designs have gained popularity with the increasing availability of large-scale family datasets [21-24]. With the inclusion of closely related family members, a new set of questions may arise. A key consideration involves the extent to which close relatives may have different ancestry distributions. For example, when two individuals from diverse populations mate, their offspring will be admixed and have ancestry distributions that differ from both parents. When a child's ancestry 'differs' from its biological parents, the child and at least one parent represent potential population outliers and will be excluded from the study. In this example, genetic ancestry estimates between sibling offspring of diverse parents may show variation in calculated ancestry due to the 'random' assortment of inherited alleles. We assess the conditions under which such situations can occur by examining ancestry estimates between family members and focusing on sibling offspring of more diverse parents to determine if they are more dissimilar to each other than those with ancestry-similar parents.

This study examines genetic ancestry estimates between pairs of family members across the spectrum of genetic relatedness, from MZ twins to nominally unrelated parent pairs, and across genotyping arrays. We leverage data from the 1000 Genomes Project (1000G) [25] and the Genome of the Netherlands (GoNL) [26, 27] reference panels as well as multiple large single nucleotide polymorphism (SNP) datasets from twin-family participants of the Netherlands Twin Register (NTR) [28, 29]. The NTR includes nuclear families, mainly two-generation, forming parent, parent-offspring, dizygotic twin and sibling, and monozygotic twin pairs, all independently genotyped. The NTR also includes SNP datasets of individuals who were genotyped on at least two separate genotyping arrays, allowing for assessment of potential platform effects on genetic ancestry estimates.

METHODS

An overview of the analytical strategies employed in this study is shown in Figure 6.1.

Sample selection and genotyping

All individuals in the study are participants of the Netherlands Twin Register (NTR) [28, 29]. The NTR recruits twins, higher-order multiples, and their family members, including parents, siblings, and spouses. DNA from NTR participants was isolated using standard protocols for obtaining high-quality DNA suitable for genome-wide SNP genotyping with high-density DNA microarrays [30]. Individuals with genotype data obtained from the Affymetrix 6.0 (AFFY6 $N_{\text{raw individuals}}$=12779, $N_{\text{raw variants}}$=905422), Affymetrix Axiom-NTR (AXIOM $N_{\text{raw individuals}}$=3606, $N_{\text{raw variants}}$=642716) [31], or Illumina GSA-NTR (ILLGSA $N_{\text{raw individuals}}$=14553, $N_{\text{raw variants}}$=669322) [32] platforms were selected. All genotyping was performed at the Avera Institute for Human Genetics (Sioux Falls, South Dakota) according to the manufacturer's protocol.

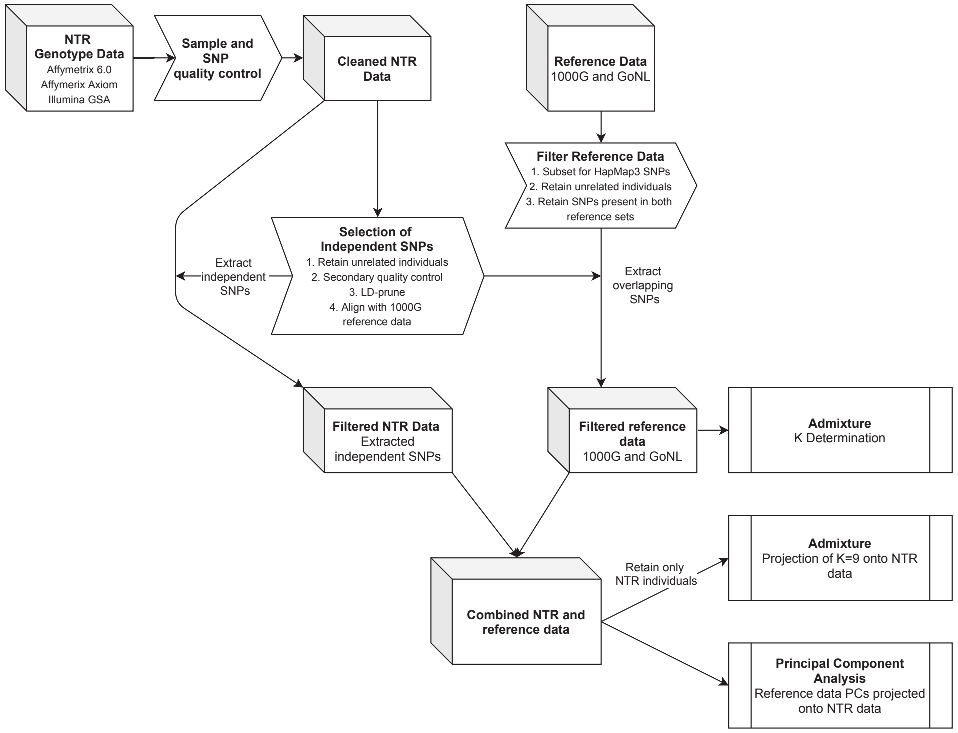

NTR Genotype Data
Affymetrix 6.0
Affymerix Axiom
Illumina GSA
Sample and SNP quality control
Cleaned NTR Data
Reference Data
1000G and GoNL
Filter Reference Data
1. Subset for HapMap3 SNPs
2. Retain unrelated individuals
3. Retain SNPs present in both reference sets
Selection of Independent SNPs
1. Retain unrelated individuals
2. Secondary quality control
3. LD-prune
4. Align with 1000G reference data
Extract independent SNPs
Extract overlapping SNPs
Filtered NTR Data
Extracted independent SNPs
Filtered reference data
1000G and GoNL
Admixture
K Determination
Combined NTR and reference data
Retain only NTR individuals
Admixture
Projection of K=9 onto NTR data
Principal Component Analysis
Reference data PCs projected onto NTR data

Figure 6.1 - Process flowchart of the analytical strategies employed in this study.

Dataset Curation

Three platform datasets were created with the backbone and custom content of each array (AFFY6, AXIOM, ILLGSA). Sample and SNP quality control was done on each dataset separately. In addition, a harmonized dataset (61,433 overlapping markers from all three platforms) was created from the cleaned platform datasets since family members could be genotyped on different arrays. The four datasets underwent the same analytical procedures.

Sample and SNP quality control

Samples were excluded if phenotypic sex did not match the genotypic sex (indicating potential sample swap) (N=271; 0.88%), if the Plink heterozygosity F value was <-0.10 or >0.10 (N=292; 0.94%), or if the sample call rate was less than 90% (N=16; 0.05%). Within families, pairwise identity-by-descent (IBD) was estimated with Plink v1.9 [33], and samples were removed if they did not match the expected familial relations (indicating potential sample swap) (N=312; 1.00%). NTR samples present in GoNL or related to GoNL participants were excluded (N=51; 0.16%).

Following sample quality control, SNPs were evaluated in each platform. They were excluded if they fit any of the following criteria: minor allele frequency (MAF)<0.005, Hardy-Weinberg Equilibrium (HWE) p-value<0.00001, SNP call rate<95%, or Mendelian error rate>2%. SNPs were also removed if they were palindromic, A/T or C/G alleles, with an allele frequency between 0.40-0.60. All platform data were aligned to build 37 of the Human Genome (hg19), and alleles were flipped to the plus strand if needed. SNPs were also removed if the allele frequencies differed more than 0.10 with the GoNL reference panel (AFFY6 N=81 of 640265 [0.01%], AXIOM N=52 of 589258 [0.01%], GSA N=233 of 497095 [0.05%]).

Final sample composition

Table 6.1 describes the final sample composition after quality control. In a first step, familial relationships were identified through IBD sharing. In the harmonized dataset, which enabled family relationships across genotyping platforms, there were 23,086 unique individuals representing 6,692 unique families. Included were 3,406 MZ twin pairs, 8,464 DZ twin or sibling pairs, 16,878 parent-offspring pairs, and 3,023 parent pairs (unrelated). In the per-platform datasets, only relationships where family members were genotyped on the same platform were considered. Across all per-platform data, the final sample consisted of 21,117 unique individuals belonging to 6,361 unique families. Of the total per-platform data (three datasets), there were 3,258 MZ twin pairs and 7,246 DZ twin or sibling pairs. In total, 13,437 parent-offspring and 2,691 parent pairs (unrelated) were identified.

Table 6.1 - Final NTR sample description after quality control and filtering

Genotyping Platform	Unique families	Unique individuals	MZ twin pairs	DZ twins/ sibling pairs	Parent-offspring pairs	Parent pairs
AFFY6	2800	7575	1279	2966	2849	438
AXIOM	734	2593	433	591	2222	448
ILLGSA	3562	11597	1546	3689	8366	1805
Across all platforms	6361	21177	3258	7246	13437	2691
Harmonized	6692	23086	3406	8464	16878	3023

As shown in Figure 6.2, 751 individuals were genotyped on at least two of the platforms, 35 of which were genotyped on all three genotyping arrays.

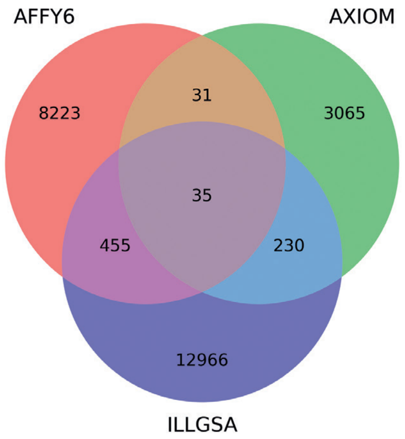


Figure 6.2 - Venn diagram of genotyped NTR individuals according to genotyping platform

Reference Dataset

Unrelated individuals in the 1000G (N=2,487) and GoNL (N=498) reference panels were determined with HapMap3 SNPs. Following the alignment of alleles between the reference panels, SNPs present in both datasets were identified (N=562,607). A final reference set was created with the overlapping markers and subsequent exclusion of SNPs with a call rate less than 98% (final marker count 562,447).

Principal Components Analysis

Following SNP quality control, the largest set of unrelated NTR participants was determined with KING v2.2.0 software [34] with options *--unrelated --degree 2*. These individuals had no 1st or 2nd-degree relationships with any other individuals. In each set, the unrelated individuals were further filtered to exclude samples with a call rate of less than 95% (a more stringent threshold than the first round of quality control).

The generation and selection of SNPs for PCA and ADMIXTURE from each platform and harmonized datasets was determined with the unrelated individual (including parent pairs) datasets. Autosomal SNPs were selected and filtered to exclude those with a call rate of less than 95%, MAF<0.01, and HWE<0.001. SNPs were pruned for Linkage Disequilibrium (LD) using Plink v1.9 by removing each SNP with an R^2 value greater than 0.5 with any other SNP within a 250-SNP sliding window (advanced by one SNP each iteration). Lastly, long-range LD regions were removed as previously described [8]. These steps resulted in a dataset-specific selection of high-quality, independent SNPs for PCA.

Four analysis datasets were generated, corresponding to each genotyping platform and the harmonized set. Related individuals were reincluded so that PCs could be calculated for all genotyped individuals, but without confounding the PC estimates since SNP selection was determined on unrelated individuals. The selected SNPs of each dataset were then merged with the final reference dataset and filtered to exclude SNPs with a call rate of less than 98% (final SNP count per dataset: AFFY6=193,840; AXIOM=215,848; ILLGSA=305,121; harmonized=50,030).

For each dataset, 10 PCs were calculated with SMARTPCA software [15], where 1000G and GoNL populations were denoted as reference populations. Related individuals were present in each dataset, but with a selection of SNPs determined from unrelated individuals. PCs were calculated for 27 total populations (26 global populations represented in 1000G plus the GoNL population) and subsequently projected onto all NTR individuals. PCs were compared between datasets and genetic relatedness groups using descriptive statistics, correlations, and Euclidean distances.

ADMIXTURE analysis

The cross-validation procedure implemented in ADMIXTURE v.1.3.0 [14, 19] revealed the optimal number of detectable populations in the merged 1000G and GoNL reference data. For this approach, the reference data were filtered for MAF<0.01 and pruned for LD (SNPs with R^2>0.5 were excluded using a 250 SNP window, advanced by one SNP each iteration). The resulting SNP set (N=394,918) was analyzed with the cross-validation procedure by increasing the number of populations from 3 to 27. The optimal number of distinct ancestral populations (denoted as K) was nine (see Results and Figure 6.3).

With K=9 specified as a sensible model parameter for ADMIXTURE, the nine populations were projected onto all NTR individuals, including related individuals. The projection analysis estimated ancestry proportions (Q1-Q9) for each individual in the NTR in each dataset. Admixture proportions were compared between genotyping platforms and genetic relatedness groups using descriptive statistics, correlations, and Euclidean distances.

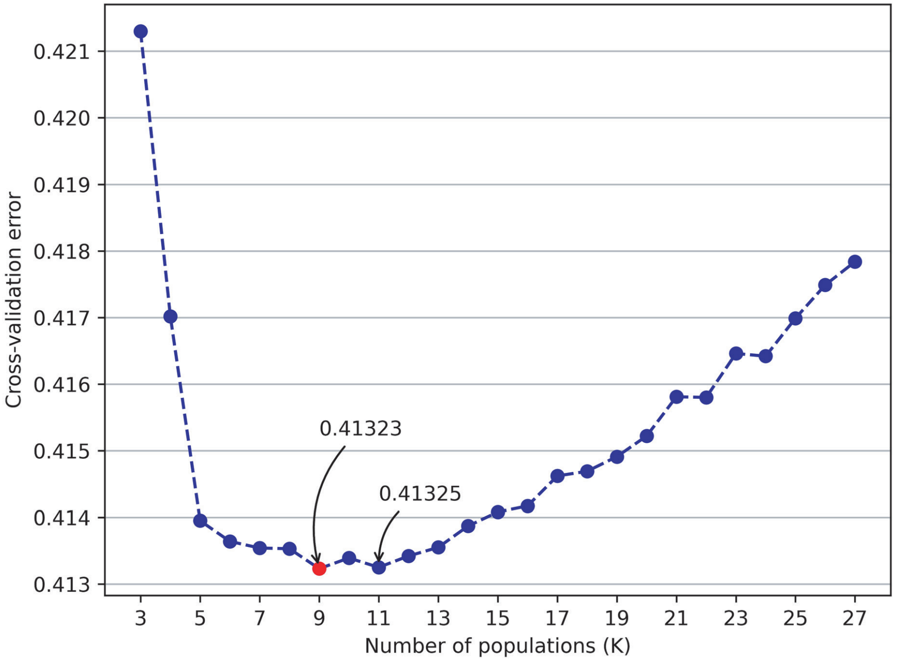


Figure 6.3 - Results of the cross-validation procedure determined by ADMIXTURE using the 1000G and GoNL reference datasets

RESULTS

Principal Components Analysis

Supplementary Material Table 6.1 contains descriptive statistics of PCs for all genotyped NTR individuals across the three genotyping platforms and the harmonized dataset. In general, the PCs between platforms are similar but are not identical since the input SNPs of each dataset vary. As a result, mean values and ranges exhibited minimal variation.

Visualization of the projected PCs 1-10 can be found in Supplementary Material Figure 6.1. For the most part, the scatter distributions of PCs 1-2 across platforms are nicely superimposed, reaffirming the similarity of calculated PCs independent of the genotyping platform. Although the same analytical procedures were applied to each dataset, the set of input SNPs for PCA varied. Therefore, the shift of plotted PCs is likely due to differences in input SNPs depending on the genotyping platform. Shifts are more pronounced in the plots of PC3 vs. PC4 and PC5 vs. PC6. These PCs may be capturing variation attributable to platform-specific SNPs included in PCA. For the axes showing the most variation, it is plausible that the ILLGSA axes are simply reversed compared to AFFY6 and AXIOM.

To examine the relationship of PCs across datasets representing distinct genotyping arrays, we calculated correlations of PCs 1-10 within and between datasets using results of array-mimicked reference populations as input. NTR samples were excluded so that PC values were representative of the 1000G and GoNL reference populations and so that PCs within a platform would not be correlated with each other. As expected, the correlations of PCs 1-10 within each genotyping platform show no correlation, reflecting the inherent statistical property of PCs in that they are uncorrelated (Supplementary Material Figure 6.2). The correlations of the same PC across platforms are near one. Inverse correlations become apparent for PC3 and PC4 between AFFY6/AXIOM and ILLGSA platforms. Further divergence of correlations is observed between PCs 6-8, potentially attributable to variation of platform SNPs.

We next investigated PC variation by assessing the differences between MZ twin and DZ twin/sibling pairs. Since sibling offspring have the same parents, it was expected that the differences in PCs between siblings would be near zero. Since MZ twins arise from the same fertilized egg, the expectation for their PC differences is zero, with non-random values reflecting measurement error. However, we recognize that post-splitting / somatic mutations can contribute to differences in DNA sequence between the twins [35]. The results of comparisons for MZ twin and DZ twin/sibling pairs are shown in Tables 6.2 and 6.3. Mean differences in PCs between MZ twins were near zero across all ten PCs irrespective of genotyping array. As expected, the mean differences between DZ twins/sibling pairs were also near zero across all genotyping platforms. With few exceptions, the absolute mean differences in PCs between MZ twins were less than DZ twin/sibling pairs across all 10 PCs and genotyping platforms. The standard deviation of the PC differences in MZ twins is always smaller than DZ twins/siblings, reflecting slightly increased variation in PC estimates between non-identical twin siblings.

To further examine PC estimates, we calculated Euclidean distance measures of PCs 1-10 within pairs of family members using Formula 6.1. Euclidean distances quantify differences in the multidimensional data between individuals with a singular metric. Within pairs, differences in respective PCs were squared and then summed over all PCs. The Euclidean distance was calculated by taking the square root of the summed squared differences. In this manner, smaller Euclidean distances represent pairs of more similar individuals across all ten PCs, whereas larger Euclidean distances indicate greater dissimilarity across all PCs. According to the relatedness group and dataset, distance measures are shown in the boxplots in Figure 6.4 (right panel). Euclidean distances were $\log_{10}$ transformed to aid in visualization. Euclidean distances across all datasets were inversely related to the genetic relatedness between pairs (Supplemental

Figure 6.3). That is, highly genetically similar/identical individuals (MZ twin pairs) have smaller Euclidean distances than DZ twin/sibling pairs, which on average, share 50% of their alleles.

Parent/offspring pairs, expected to be precisely 50% genetically similar, show more considerable variation in Euclidean distances than DZ twins/sibling pairs. Parent pairs, assumed to be unrelated, have the largest Euclidean distances.

Formula 6.1 – Formula for calculating Euclidean distances between pairs of individuals for ten PCs or nine ancestry proportions.

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}(x_j - y_j)^2}$$

$d_{x,y}$ = Euclidean distance of J between two individuals
x, y = two individuals, representing a pair within a family
J = PCs 1–10 or Q 1–9

Utilizing individuals with genetic data obtained from multiple genotyping platforms (N=751; 35 of which were genotyped on all three platforms - see Figure 6.1), we calculated Euclidean distances of PCs within individuals across genotyping arrays (Figure 6.5 right panel). We expected to see Euclidean distances near zero in this manner, similar to the Euclidean distances seen between MZ twins. The smallest distance values were obtained for individuals with genotypic data from the AFFY6 and AXIOM platforms, both Affymetrix products. Larger distances were observed for individuals with data from either the Affymetrix-manufactured array (AFFY6/AXIOM) and the Illumina platform (ILLGSA). Because the platform SNPs on which the PCs are based are not identical, the observed differences can be attributed to input SNP variation.

Table 6.2 – Within family MZ twin pair differences in PCs by genotyping array

	AFFY6 (N=1279)				AXIOM (N=433)				ILLGSA (N=1546)				Harmonized (N=3406)			
	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD
PC1	0.0000020	0.0001606	0.0001	0.0001	0.0000014	0.0001309	0	0.0001	-0.0000002	0.0001520	0	0.0001	0.0000042	0.0001432	0	0.0001
PC2	0.0000013	0.0001771	0.0001	0.0001	0.0000012	0.0001436	0.0001	0.0001	-0.0000039	0.0001722	0	0.0001	0.0000022	0.0001929	0	0.0001
PC3	-0.0000038	0.0001050	0.0001	0.0001	-0.0000025	0.0000973	0.0001	0.0001	-0.0000018	0.0000819	0	0.0001	-0.0000041	0.0001296	0.0001	0.0001
PC4	0.0000000	0.0001178	0.0001	0.0001	0	0.0000962	0.0001	0.0001	0.0000007	0.0001031	0	0.0001	-0.0000041	0.0001392	0.0001	0.0001
PC5	-0.0000085	0.0001717	0.0001	0.0002	0.0000187	0.0002288	0.0001	0.0001	-0.0000017	0.0001378	0	0.0001	-0.0000001	0.0002559	0.0001	0.0002
PC6	-0.0000010	0.0002334	0.0001	0.0001	0.0000058	0.0001777	0.0001	0.0002	0.0000026	0.0001419	0	0.0001	0.0000102	0.0003171	0.0001	0.0002
PC7	0.0000034	0.0002173	0.0001	0.0001	0.0000115	0.0002006	0.0001	0.0001	0.0000026	0.0001357	0.0001	0.0001	-0.0000076	0.0002859	0.0001	0.0002
PC8	0.0000099	0.0002456	0.0001	0.0001	0.0000088	0.0002843	0.0001	0.0002	-0.0000010	0.0001044	0	0.0001	0.0000104	0.0003477	0.0001	0.0002
PC9	-0.0000005	0.0002663	0.0001	0.0002	0.0000229	0.0002111	0.0001	0.0001	0.0000033	0.0001463	0.0001	0.0001	0.0000027	0.0003933	0.0002	0.0003
PC10	0.0000024	0.0002565	0.0001	0.0002	0.0000136	0.0002727	0.0002	0.0002	0.0000025	0.0001324	0	0.0001	0.0000046	0.0003907	0.0002	0.0003

MD and SD are the mean and standard deviation of paired principal components, MAD=median absolute difference, IQRAD=interquartile range absolute difference of quartile 1 and quartile 3.

Table 6.3 – Within family DZ twin/sibling pair differences in PCs by genotyping array

	AFFY6 (N=2966)				AXIOM (N=591)				ILLGSA (N=3689)				Harmonized (N=8464)			
	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD
PC1	-0.0000027	0.0002658	0.0002	0.0002	-0.0000249	0.0002404	0.0002	0.0001	0.0000043	0.0002733	0.0001	0.0001	-0.0000006	0.0003408	0.0002	0.0003
PC2	-0.0000179	0.0004050	0.0002	0.0003	-0.0000218	0.0003827	0.0002	0.0003	0.0000124	0.0003816	0.0002	0.0002	0.0000050	0.0005172	0.0003	0.0004
PC3	0.0000049	0.0006581	0.0004	0.0006	0.0000538	0.0006119	0.0004	0.0005	0.0000076	0.0005788	0.0004	0.0005	0.0000129	0.0009118	0.0006	0.0008
PC4	0.0000281	0.0006870	0.0005	0.0006	-0.0000294	0.0006448	0.0004	0.0006	0.0000062	0.0006056	0.0004	0.0005	0.0000008	0.0009624	0.0006	0.0008
PC5	-0.0000189	0.0015276	0.0010	0.0012	0.0000088	0.0019236	0.0014	0.0017	0.0000002	0.0017806	0.0012	0.0014	-0.0000367	0.0022051	0.0015	0.0018
PC6	-0.0000097	0.0020585	0.0014	0.0017	0.0001037	0.0015581	0.0010	0.0013	0.0000539	0.0016400	0.0011	0.0014	-0.0000418	0.0028650	0.0020	0.0024
PC7	-0.0000297	0.0013230	0.0009	0.001	0.0000091	0.0013700	0.0009	0.0011	0.0000183	0.0016578	0.0011	0.0014	0.0000016	0.0023673	0.0016	0.0019
PC8	0.0000152	0.0017449	0.0012	0.0015	-0.0000959	0.0022387	0.0016	0.002	0.0000259	0.0012354	0.0008	0.001	-0.0000485	0.0027616	0.0019	0.0023
PC9	-0.0000432	0.0022698	0.0015	0.0019	-0.0001530	0.0015666	0.0011	0.0014	-0.0000784	0.0020788	0.0014	0.0018	-0.0000309	0.0033817	0.0022	0.0027
PC10	-0.0000735	0.0021903	0.0015	0.0018	-0.0000922	0.0022229	0.0016	0.0018	-0.0000031	0.0014639	0.0010	0.0013	-0.0000133	0.0033483	0.0022	0.0027

MD and SD are the mean and standard deviation of paired principal components, MAD=median absolute difference, IQRAD=interquartile range absolute difference of quartile 1 and quartile 3.

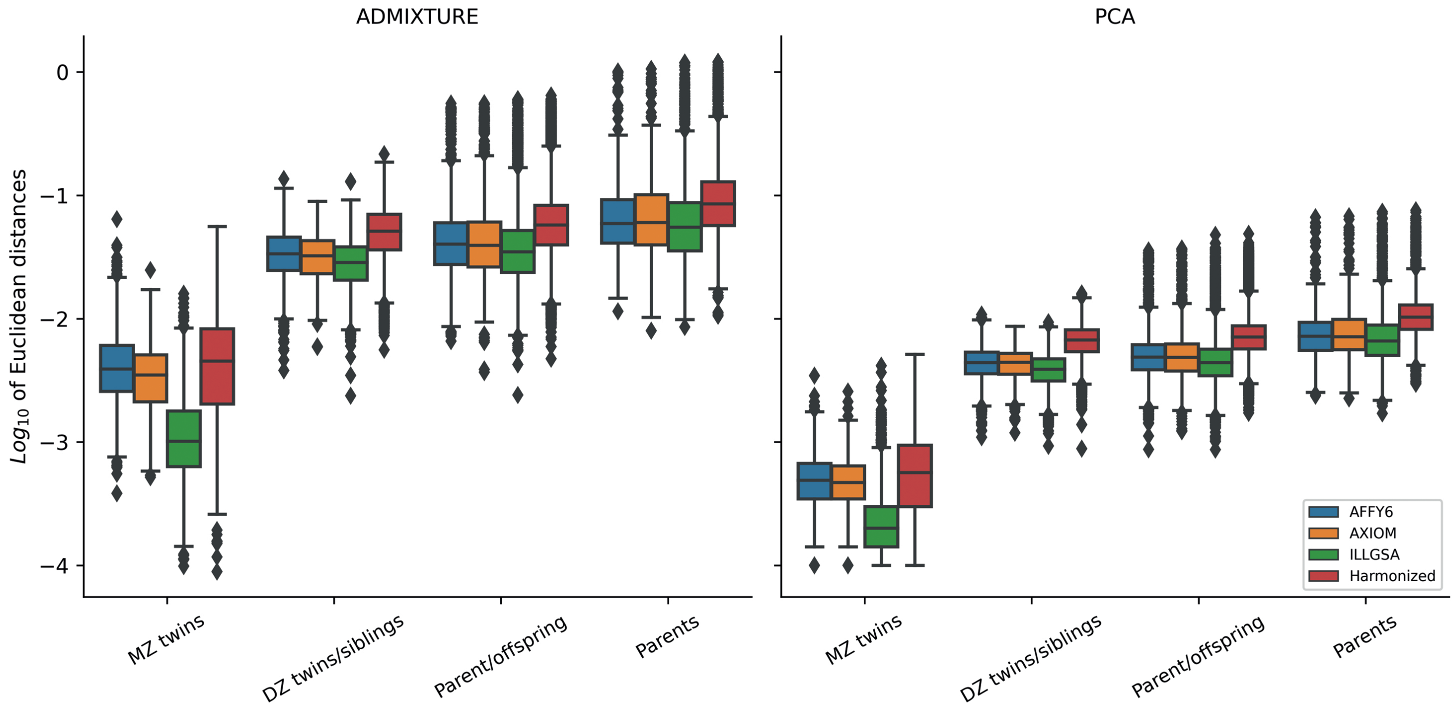


Figure 6.4 - Euclidean distance measures of ADMIXTURE ancestry proportions (Q1-Q9) and principal components (PC1-PC10) between relationship pairs

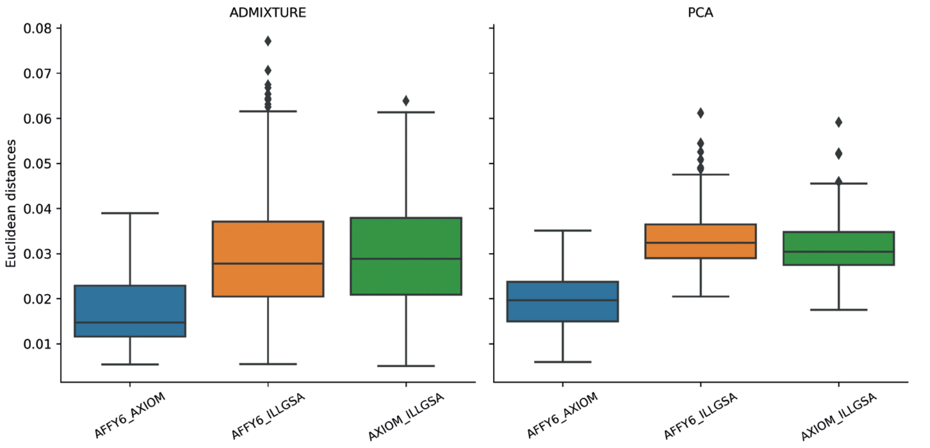


Figure 6.5 - Euclidean distances of ADMIXTURE ancestry proportions Q1-Q9 and PC1-PC10 of individuals genotyped on multiple platforms.

ADMIXTURE analysis

We ran ADMIXTURE analysis to partition global reference genetic variation into an optimal number of distinct genetic clusters. Figure 6.3 portrays the results of the cross-validation procedure as implemented in ADMIXTURE to determine the model parameter, K, with the best predictive accuracy. With the 1000G and GoNL reference populations as input, the most sensible choice of K was determined to be nine since it achieved the lowest cross-validation error compared to other tested values of K (cross-validation error=0.41323).

Global representation of the nine identified ancestral populations is illustrated in Figure 6.6. Ancestral population 1 represents an amalgam of Colombia, Italy, Puerto Rico, and Spain. Population 2 predominantly reflects Chinese regions (Beijing and Xishuangbanna) and Vietnam. Population 3 captures the Finnish population. Population 4 characterizes Northern and Western European populations from England, Scotland, and the Netherlands. Population 5 embodies Peruvian and Mexican populations. Population 6 reflects the Western African populations from Gambia and Sierra Leone. Population 7 symbolizes South Asian countries, namely Bangladesh, India, Pakistan, and Sri Lanka. Population 8 mirrors African populations from Kenya and Nigeria. Population 9 represents East Asian countries, primarily from Japan, but also from Beijing, China.

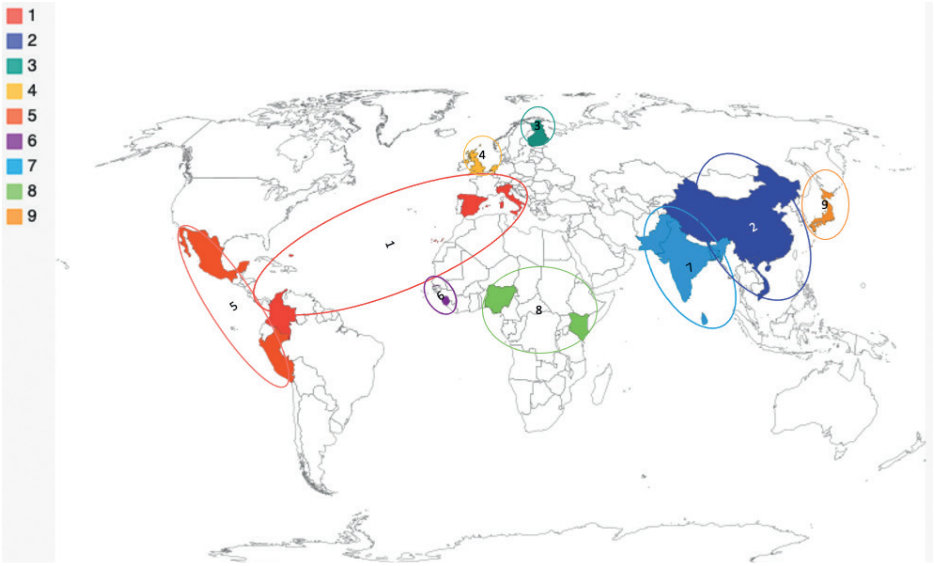


Figure 6.6 - World view of the nine populations as determined by ADMIXTURE

To estimate individual ancestry in the NTR samples, we projected them onto the population structure (allele frequencies) derived from the 1000G and GoNL reference datasets by specifying K=9. Descriptive statistics of ancestry fractions (Q1-Q9) from the three platforms and harmonized datasets of NTR participants are shown in Supplementary Material Table 6.2. There was minimal variation in mean ancestry proportions across datasets. The ancestral population labeled Q4 represents the majority ancestry fraction for NTR individuals, indicating that most genetic ancestry corresponds to the 1000G and GONL reference data obtained from Northern and Western Europeans and the Netherlands.

The stacked bar charts in Figure 6.7 A-D display the ancestry proportion estimates of each NTR individual per genotyping platform and in the harmonized dataset. Each stacked bar reflects a single individual and their ancestry fractions for the nine populations arranged in increasing order of population 4 (i.e., Northern and Western European and the Netherlands) ancestry proportion. The average proportional population 4 ancestry is 0.695, 0.687, 0.687, and 0.694 from AFFY6, AXIOM, ILLGSA, and harmonized data, respectively. Across all data, there is a modest amount of ancestry captured by Population 1 with average estimates of 0.187, 0.191, 0.191, and 0.188, correspondingly.

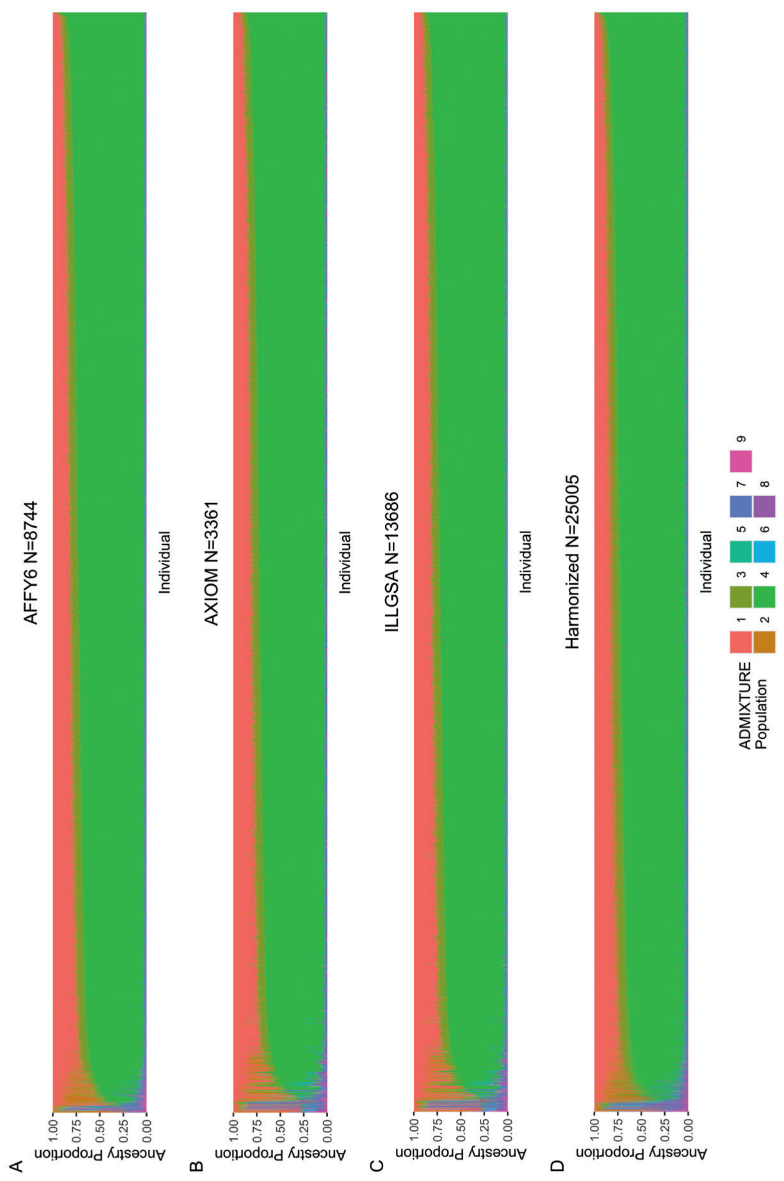


Figure 6.7 - Results of ADMIXTURE analysis for each genotyping array with K=9. Each NTR individual is represented by a thin vertical bar, partitioned into nine (K=9) colored segments representing the estimated membership proportions in K populations. For clarity, the bars are ascendingly sorted by population 4 proportions.

The ancestral fractions are nearly indistinguishable from each other for all NTR individuals across platforms. This finding highlights a relatively similar population composition of individuals genotyped on each platform, though it is imperative to consider the comparison level since the individuals live in the Netherlands. Comparatively, PCs can reveal more fine-grained differences between the same individuals, such as North-South clines. Within each dataset, a small number of genetically diverse and admixed individuals are shown on the left side of each figure. These individuals show stark variation in the ancestry proportions relative to the bulk of the NTR sample population, indicating more heterogeneous ancestry and deviation from majority Northern and Western Europe origin as captured by population 4. Similar admixture and population heterogeneity patterns among NTR samples were observed in the PCs (Supplementary Material Figure 6.1 F-J).

Correlations of ancestry proportions within and between genotyping platforms for all NTR participants are shown in Supplementary Material Figure 6.4. The estimates are strongly correlated across genotyping arrays within each ancestral population, represented as Q1-Q9. For values of Q within the genotyping platform, ancestry estimates are mostly negatively correlated or not correlated. Between values of Q and between genotyping platforms, estimates are also mainly negatively correlated or not correlated at all. Exceptions include positive correlations between Q2 and Q7 as well as Q2 and Q9, reflecting moderate overlap in South and East Asian populations. There were also slightly positive correlations between Q5 and Q8, and Q6 and Q8 obtained from AXIOM and ILLGSA arrays. The correlation between Q6 and Q8 correlation is likely due to the overlap of African populations.

We compared the estimates between MZ twins and between DZ twins/sibling pairs to examine the ancestry proportions in more detail. Results of the comparisons are shown in Tables 6.4 and 6.5. Mean differences between MZ twins were near zero across all ancestry proportions and genotyping arrays. The same was true for DZ twins/siblings. Although within respective pair differences are small, the mean differences are nearly always smaller between MZ twins than between DZ twins/siblings. Likewise, as measured by the standard deviation, the variation of the differences is less between MZ twins than between DZ twins/sibling pairs. Except for a few instances, namely AFFY6 Q6, Q8, Q9, and ILLGSA Q9, the absolute mean differences in ancestry proportions between MZ twins were less than DZ twins/siblings

Consistent with the evaluation of PCs, we calculated Euclidean distances over the nine ancestry proportions within family pairs according to Formula 6.1. Comparable to the Euclidean distances of PCs, the distances in ancestry proportions were noticeably smaller in MZ twin pairs than within DZ twins and sibling pairs across all datasets (Figure 6.4 left panel). Euclidean distances were $\log_{10}$ transformed to aid in visualization.

We also investigated the ancestry proportions of individuals with genotype information from multiple platforms. Like the Euclidean distances of PCs of individuals genotyped on multiple arrays, the smallest distances were observed for those with genetic data obtained from Affymetrix platforms (Figure 6.5 left panel). Larger distances were observed between Affymetrix and Illumina platforms.

Ancestry outliers - PCA vs. ADMIXTURE

Ancestry outliers in the NTR datasets were identified by defining thresholds based on the minimum and maximum PC and ancestry proportion values of CEU or GoNL reference populations. For PCA, thresholds were defined within each dataset (i.e., platform) since PCA projection was performed per dataset. PCs from CEU and GoNL individuals were calculated by platform-mimicked datasets. Alternatively, CEU and GoNL platform-specific thresholds were not possible for ADMIXTURE since the nine populations were determined with an LD-pruned dataset of markers present in both 1000G and GoNL panels. Each NTR dataset was projected onto the reference populations.

Ancestry outliers were defined as having PCs or ADMIXTURE proportions less than or greater than reference (i.e., CEU or GoNL) minimums or maximums, respectively. NTR individuals with values greater than or equal to the reference minimum or less than or equal to the reference maximum were considered inliers. Outliers were determined for each PC and each value of Q. The total number of outliers across all PCs and values of Q was determined by identifying unique individuals.

Table 6.6 shows the number of outliers and inliers per dataset with thresholds determined by CEU or GoNL reference populations. The number of outliers between PCA and ADMIXTURE was very similar when thresholds were defined by the larger GoNL reference population (N=498). Larger deviation in outlier counts was observed when CEU (N=99) was used for defining boundaries, which is a smaller and more ancestrally variable population than GoNL. Regardless of the reference population, there is more variation in outlier counts in the harmonized dataset, likely due to the smaller number of markers used in the calculations.

Assessment of within-family diversity

Using the calculated PCs and ADMIXTURE ancestry proportions, we also assessed if sibling offspring (non-MZ twin) of diverse parents were more dissimilar to each other than those with parents of similar ancestry. We found very modest positive correlations between Euclidean distances of parent pairs (i.e., father and mother) and averaged distances of all DZ twin and sibling pairs within a family (ADMIXTURE Spearman's rho 0.07, P-value=0.005; PCA Spearman's rho 0.04, P-value=0.122). Euclidean distances of non-identical twin offspring were averaged within a family to avoid inflating the number of comparisons in families with multiple offspring. The results are plotted in Figure 6.8, showing $\log_{10}$ transformation of Euclidean distances to aid in visualization. Though negligible correlation was observed, the Euclidean distances calculated from PCs are smaller in magnitude than those derived from ADMIXTURE proportions. Regardless of the method, the near-zero relationship indicates that sibling offspring of more diverse parents are not more dissimilar than the progeny of similar parents.

DISCUSSION

Ancestry estimation is a robust method for inferring population structure and is routinely employed in genetic association studies. With the abundance of data from various array-based genotyping technologies and the increasing popularity of within-family study designs, we sought to examine ancestry estimates as a function of genotyping array and genetic relatedness within f nuclear families. Here, we evaluated estimates of genetic ancestry obtained from PCA and ADMIXTURE in a large number of NTR twins and family members using whole-genome SNP data from three distinct genotyping platforms.

Utilizing reference data from 1000G and GoNL as global population surrogates, we demonstrated that PCs across genotyping arrays are not the same despite identical analytical strategies due to differences in platform SNPs used to calculate them. Calculation of PCs from the array-mimicked global reference data and subsequent projection onto NTR data resulted in top PCs capturing differences in ancestry. Within each platform and in the harmonized dataset, mean differences in PCs of family-matched MZ and DZ twins and siblings were near zero. Further, we calculated Euclidean distances to capture differences succinctly and quantitatively across all 10 PCs within familial pairs. Euclidean distance measures of PCs were inversely related to the degree of genetic similarity between individuals. The greater the genetic relatedness between two individuals, the smaller the Euclidean distances of their respective PCs. This finding was expected, given that the twin/sibling offspring of each pair have the same parents and possess a genetic profile derived from the same pool of segregating alleles.

Table 6.4 – Within family MZ twin differences in ancestry proportions by genotyping array

	AFFY6 (N=1279)				AXIOM (N=433)				ILLGSA (N=1546)				Harmonized (N=3406)			
	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD
Q1	−0.0000237	0.0022813	0.0013	0.0018	−0.0002526	0.0022389	0.0011	0.0016	0.0000071	0.0009782	0.0004	0.0005	−0.0000953	0.0038340	0.0015	0.0028
Q2	0.0000270	0.0012185	0.0002	0.0009	−0.0000023	0.0008463	0.0002	0.0007	−0.0000115	0.0005284	0.0001	0.0003	0.0000143	0.0014961	0.0000	0.0007
Q3	−0.0000005	0.0020525	0.0012	0.0016	−0.0000082	0.0020162	0.001	0.0015	0.0000181	0.0009788	0.0003	0.0006	0.0000385	0.0035242	0.0013	0.0024
Q4	0.0000402	0.0049464	0.0018	0.0027	0.0003037	0.0033462	0.0017	0.0025	−0.0000141	0.0013523	0.0005	0.0007	0.0001686	0.0055211	0.0020	0.0038
Q5	0.0000148	0.0007813	0.0003	0.0006	−0.0000164	0.0005789	0.0002	0.0005	−0.0000032	0.0003113	0.0001	0.0002	−0.0000189	0.0010212	0.0002	0.0007
Q6	0.0000550	0.0015490	0.0003	0.001	−0.0000545	0.0009789	0.0002	0.0006	0.0000016	0.0004610	0	0.0002	−0.0000358	0.0013595	0.0000	0.0004
Q7	−0.0000285	0.0016026	0.0007	0.0012	−0.0000099	0.0012470	0.0006	0.001	−0.0000011	0.0006435	0.0002	0.0003	−0.0000667	0.0021005	0.0005	0.0015
Q8	−0.0000545	0.0021503	0.0003	0.001	0.0000075	0.0011565	0.0001	0.0006	−0.0000215	0.0004775	0	0.0002	−0.0000230	0.0014435	0.0000	0.0005
Q9	−0.0000298	0.0010583	0.0001	0.0008	0.0000326	0.0007626	0	0.0005	0.0000245	0.0004302	0	0.0002	0.0000183	0.0014032	0.0000	0.0006

Q1-Q9 represent each of the nine ancestry populations as determined by ADMIXTURE, MD and SD are the mean and standard deviation of paired ancestry proportion differences, MAD=median absolute difference, IQRAD=interquartile range absolute difference of quartile 1 and quartile 3.

Table 6.5 – Within family DZ twin/sibling pair differences in admixture proportions by genotyping array

	AFFY6 (N=2966)				AXIOM (N=591)				ILLGSA (N=3689)				Harmonized (N=8464)			
	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD	MD	SD	MAD	IQRAD
Q1	0.0001091	0.0218368	0.0144	0.0184	−0.0010909	0.0208740	0.0144	0.0176	0.0003525	0.0183653	0.0126	0.0154	0.0007276	0.0324349	0.0219	0.0269
Q2	0.0000416	0.0066079	0.0023	0.0056	0.0004513	0.0056273	0.0023	0.0054	−0.0000435	0.0061190	0.0025	0.0051	−0.0001435	0.0088039	0.0022	0.0077
Q3	0.0000445	0.0169808	0.0115	0.0144	0.0002999	0.0164323	0.0115	0.0127	0.0002883	0.0141443	0.0094	0.012	−0.0002675	0.0279527	0.0189	0.0237
Q4	−0.0005116	0.0255068	0.0173	0.0212	0.0004375	0.0233512	0.0173	0.0202	−0.0005890	0.0202821	0.0138	0.0166	−0.0005575	0.0386821	0.0258	0.0322
Q5	0.0001855	0.0044754	0.0028	0.0039	−0.0000242	0.0042301	0.0028	0.0039	0.0001141	0.0040468	0.0024	0.0034	0.0000168	0.0065354	0.0034	0.0066
Q6	−0.0000331	0.0041066	0.0020	0.0042	0.0000620	0.0037017	0.0020	0.0039	−0.0000086	0.0036473	0.0017	0.0038	0.0000588	0.0054829	0.0013	0.0053
Q7	0.0001313	0.0094434	0.0061	0.0081	−0.0001375	0.0087924	0.0061	0.0081	−0.0002554	0.0083539	0.0055	0.007	0.0002028	0.0130998	0.0075	0.0123
Q8	0.0000234	0.0040089	0.0018	0.0042	0.0003426	0.0042484	0.0018	0.0043	0.0001301	0.0040864	0.0019	0.0038	−0.0000872	0.0056071	0.0015	0.0053
Q9	0.0000092	0.0052608	0.0020	0.0054	−0.0003406	0.0053752	0.0020	0.0053	0.0000114	0.0050228	0.0018	0.0048	0.0000496	0.0075707	0.0015	0.0069

Q1-Q9 represent each of the nine ancestry populations as determined by ADMIXTURE, MD and SD are the mean and standard deviation of paired ancestry proportion differences, MAD=median absolute difference, IQRAD=interquartile range absolute difference of quartile 1 and quartile 3.

Table 6.6 – Outliers and inliers from PCA and ADMIXTURE based on two reference datasets

	Min. and Max. thresholds determined by GONL[a]					Min. and Max. thresholds determined by CEU[b]				
	PCA[c]		ADMIXTURE[d]		Common Outliers[e]	PCA[c]		ADMIXTURE[d]		Common Outliers[e]
Dataset (N)	Outliers (%)	Inliers (%)	Outliers (%)	Inliers (%)		Outliers (%)	Inliers (%)	Outliers (%)	Inliers (%)	
AFFY6 (8744)	525 (6.0%)	8219 (94.0%)	695 (7.9%)	8049 (92.1%)	457	818 (9.4%)	7926 (90.6%)	2419 (27.7%)	6325 (72.3%)	704
AXIOM (3361)	249 (7.4%)	3112 (92.6%)	341 (10.1%)	3020 (89.9%)	224	451 (13.4%)	2910 (86.6%)	896 (26.7%)	2465 (73.3%)	382
ILLGSA (13686)	933 (6.8%)	12753 (93.2%)	954 (7.0%)	12732 (93.0%)	809	1169 (8.5%)	12517 (91.5%)	2501 (18.3%)	11185 (81.7%)	891
Harmonized (25005)	1794 (7.2%)	23211 (92.8%)	5617 (22.5%)	19388 (77.5%)	1584	3438 (13.7%)	21567 (86.3%)	12501 (50.0%)	12504 (50.0%)	2796

[a] Sample size is 498

[b] Sample size is 99

[c] Minimum and maximum thresholds determined within each dataset (498 GONL and 99 CEU reference samples mimicking the content of each dataset)

[d] Minimum and maximum thresholds determined from entire reference dataset (GONL and CEU) and not within each dataset

[e] Number of outliers in shared between each method

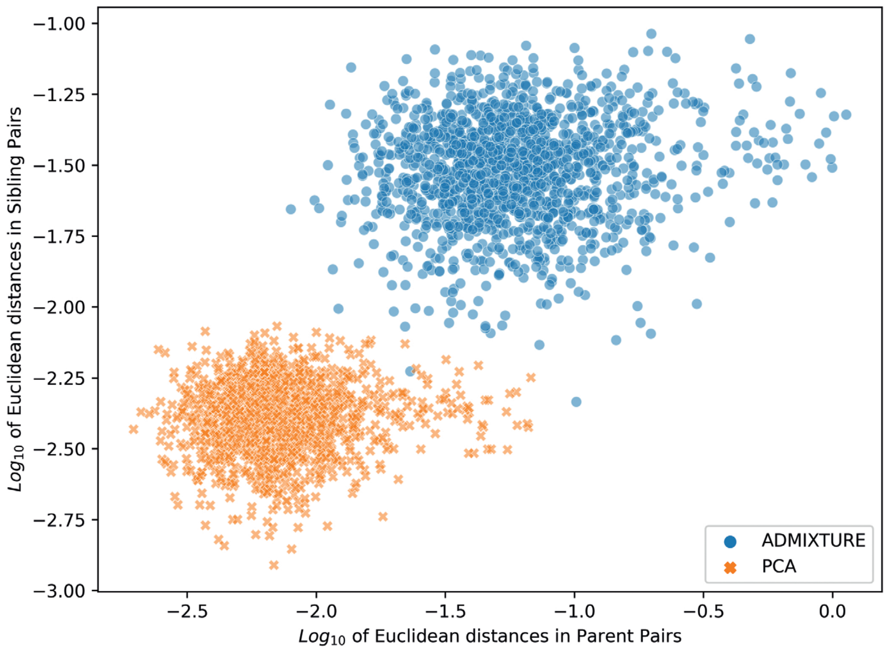


Figure 6.8 - Scatter plot of Euclidean distance measures of ADMIXTURE ancestry proportions Q1-Q9 and PC1-PC10 for parental pairs and averaged sibling pairs within families.

We used ADMIXTURE, a model-based ancestry estimation method, to detect the optimal number of ancestral populations (i.e., K) in the global reference data. The value of K with the lowest error in the cross-validation procedure was nine. We utilized a projection analysis to determine the composition of each NTR individual across the nine populations. In this manner, allele frequencies of NTR individuals were compared to the allele frequencies of the nine ancestral populations, and proportions of each population were ascribed to each NTR individual. Population 4 was the major ancestry fraction of NTR participants, representing Northern and Western Europe and the Netherlands. Akin to the results of the PCA, differences in ancestry proportions of the nine populations were near zero between MZ twins and DZ twins/siblings. As with PCs, Euclidean distance measures of ancestry proportions were inversely proportional to the amount of allele sharing between family members.

Given the uniqueness of the NTR data, we evaluated estimates of genetic ancestry using genotypic data from independently genotyped MZ twins. Importantly, within-platform differences of MZ twins were non-zero. The mean Euclidean distances in the harmonized dataset were also larger, likely because of the reduced number of SNPs used for ancestry estimation. Although measurement error may play a role, differences within MZ twin pairs cannot simply be ascribed to measurement error. One DNA sequencing study showed in a 40- and a 100-year-old MZ twin pair that somatic mosaicism leads to differences within pairs [36]. More recently, germline differences were shown in a large Icelandic study of the genomes in pedigrees of MZ twins [35]. Whether potential germline differences are the source of the variation in ancestry estimates for MZ twins remains to be determined.

We also examined ancestry estimates of individuals genotyped on multiple genotyping arrays. We observed differences larger than between MZ twins; however, this variation is mainly attributable to differences in platform-specific SNPs used in calculating ancestry. Regardless of the estimation method, the most considerable Euclidean distances were between AFFY6 and ILLGSA platforms. In the timeline of NTR genotyping efforts included in this study, AFFY6 and ILLGSA are the oldest and most recent genotyping platforms, respectively. Successively large differences were between AXIOM and ILLGSA. AFFY6 and AXIOM are Affymetrix products, whereas Illumina manufactures ILLGSA. Thus, variation in array (the molecular approach utilized to measure SNP genotypes) manufacturer and subsequent platform-specific genotype calling algorithms may contribute an effect. The potential impact of array manufacturer is also seen in the boxplots of Euclidean distances for MZ twins. On average, ancestry estimates of MZ twins obtained from ILLGSA are more similar than AXIOM, which are more than AFFY6.

Ancestry is known to exist on a continuum due to the complexity of human evolution and repeated migrations. Thus, it should be kept in mind that the spectrum of ancestry referred to in this work is constrained by the diversity represented in the surrogate samples from the 1000G and GoNL projects. As more extensive and diverse genetic datasets become available, finer resolution estimates of genetic ancestry will be possible. Another important consideration concerns the optimal number of PCs and ancestral populations when making claims regarding genetic ancestry. A variety of statistical tests have been recommended for selecting the ideal number of PCs (e.g., Tracy-Widom statistics [15]) or ancestral populations from ADMIXTURE (e.g., Bayesian information criterion [19]) to consider for downstream analysis. Still, others advise that these decisions be made based on the knowledge of the history of the study population(s) [20] or additional investigative analysis [7]. Although arbitrary, the top 10 PCs of the PCA method are often included in association studies to adjust population structure [37-39], which is the number of PCs we considered in this project. It is possible that additional examination of ancestry estimates derived from PCA, including selecting PCs that correspond best with genetic ancestry, will lead to utilizing additional (e.g., more than 10) PCs. We evaluated ancestry estimates between platforms using a harmonized dataset comprised of a modest number of overlapping genetic markers of the three different genotyping arrays. Future studies examining ancestry estimates when genotype data are coordinated and aggregated via imputation would be of merit.

Overall, we show genetic ancestry inference methods can provide reliable estimates of individual genetic ancestry across the genetic relatedness spectrum and when genetic data are sourced from various genotyping arrays. The consistency of the estimates is contingent upon the inclusion of necessary proxies of global population diversity and proper analytical execution. Genetic relatedness can confound individual ancestry estimates in the absence of reference population samples [40]. Alternative methods for handling relatedness in PCA have been proposed [41, 42], though they rely on performing PCA on diverse unrelated individuals first with subsequent PC prediction based on genetic similarities. To mitigate the concern of genetic relatedness, we utilized projection strategies to select independent SNPs for PCA and ADMIXTURE analyses based on unrelated individuals from globally diverse reference populations. We showed that PCs and ancestry proportions from ADMIXTURE show negligible differences between closely related pairs of individuals (i.e., MZ twins and DZ twins/sibling pairs) and individuals with genetic data obtained from different genotyping platforms. As expected, we observed that as the degree of relatedness between any two individuals becomes less, differences in their ancestry estimates become greater. Despite consistent results from PCA and ADMIXTURE, ancestry proportion estimates may be more favorable since they are more easily interpretable. ADMIXTURE returns membership proportions to surrogate global ancestral populations, whereas PCA simply reveals axes of variation in the data. Regardless of whichever method a researcher prefers, we show that the performance of PCA and the software ADMIXTURE for estimating genetic ancestry is comparable for downstream analyses involving families or different genotyping platforms.

REFERENCES

1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. *10 Years of GWAS Discovery: Biology, Function, and Translation.* American Journal of Human Genetics. 2017;**101**(1):5-22.
2. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet. 2006;**38**(8):904-9.
3. Novembre J, Stephens M. *Interpreting principal component analyses of spatial population genetic variation.* Nat Genet. 2008;**40**(5):646-9.
4. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. *Genes mirror geography within Europe.* Nature. 2008;**456**(7218):98-101.
5. Reich D, Price AL, Patterson N. *Principal component analysis of genetic data.* Nat Genet. 2008;**40**(5):491-2.
6. Abdellaoui A, Hottenga J-J, de Knijff P, Nivard MG, Xiao X, Scheet P, et al. *Population structure, migration, and diversifying selection in the Netherlands.* European Journal of Human Genetics : EJHG. 2013;**21**(11):1277-85.
7. Prive F, Luu K, Blum MGB, McGrath JJ, Vilhjalmsson BJ. *Efficient toolkit implementing best practices for principal component analysis of population genetic data.* Bioinformatics. 2020;**36**(16):4449-57.
8. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, et al. *Long-range LD can confound genome scans in admixed populations.* American Journal of Human Genetics. 2008;**83**(1):132-5; author reply 5-9.
9. Zou F, Lee S, Knowles MR, Wright FA. *Quantification of population structure using correlated SNPs by shrinkage principal components.* Hum Hered. 2010;**70**(1):9-22.
10. Prive F, Aschard H, Ziyatdinov A, Blum MGB. *Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr.* Bioinformatics. 2018;**34**(16):2781-7.
11. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. *Association mapping in structured populations.* American Journal of Human Genetics. 2000;**67**(1):170-81.
12. Raj A, Stephens M, Pritchard JK. *fastSTRUCTURE: variational inference of population structure in large SNP data sets.* Genetics. 2014;**197**(2):573-89.
13. Tang H, Peng J, Wang P, Risch NJ. *Estimation of individual admixture: analytical and study design considerations.* Genet Epidemiol. 2005;**28**(4):289-301.
14. Alexander DH, Lange K. *Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.* BMC Bioinformatics. 2011;12:246.
15. Patterson N, Price AL, Reich D. *Population structure and eigenanalysis.* PLoS Genet. 2006;**2**(12):e190.
16. Engelhardt BE, Stephens M. *Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis.* PLoS Genet. 2010;**6**(9):e1001117.
17. McVean G. *A genealogical interpretation of principal components analysis.* PLoS Genet. 2009;**5**(10):e1000686.
18. Ma J, Amos CI. *Principal components analysis of population admixture.* PLoS One. 2012;**7**(7):e40115.
19. Alexander DH, Novembre J, Lange K. *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res. 2009;**19**(9):1655-64.

20. Zheng X, Weir BS. *Eigenanalysis of SNP data with an identity by descent interpretation.* Theor Popul Biol. 2016;107:65-76.
21. Brumpton B, Sanderson E, Heilbron K, Hartwig FP, Harrison S, Vie GA, et al. *Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses.* Nature Communications. 2020;**11**(1):3519.
22. Abecasis GR, Cardon LR, Cookson WO. *A general test of association for quantitative traits in nuclear families.* American Journal of Human Genetics. 2000;**66**(1):279-92.
23. Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. *Within-sibship GWAS improve estimates of direct genetic effects.* bioRxiv. 2021:2021.03.05.433935.
24. Benyamin B, Visscher PM, McRae AF. *Family-based genome-wide association studies.* Pharmacogenomics. 2009;**10**(2):181-90.
25. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. *A global reference for human genetic variation.* Nature. 2015;**526**(7571):68-74.
26. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. *The Genome of the Netherlands: design, and project goals.* Eur J Hum Genet. 2014;**22**(2):221-7.
27. Genome of the Netherlands C. *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet. 2014;**46**(8):818-25.
28. Willemsen G, Vink JM, Abdellaoui A, den Braber A, van Beek JH, Draisma HH, et al. *The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection.* Twin Res Hum Genet. 2013;**16**(1):271-81.
29. van Beijsterveldt CE, Groen-Blokhuis M, Hottenga JJ, Franic S, Hudziak JJ, Lamb D, et al. *The Young Netherlands Twin Register (YNTR): longitudinal twin and family studies in over 70,000 children.* Twin Res Hum Genet. 2013;**16**(1):252-67.
30. Min JL, Lakenberg N, Bakker-Verweij M, Suchiman E, Boomsma DI, Slagboom PE, et al. *High microsatellite and SNP genotyping success rates established in a large number of genomic DNA samples extracted from mouth swabs and genotypes.* Twin Res Hum Genet. 2006;**9**(4):501-6.
31. Ehli EA, Abdellaoui A, Fedko IO, Grieser C, Nohzadeh-Malakshah S, Willemsen G, et al. *A method to customize population-specific arrays for genome-wide association testing.* Eur J Hum Genet. 2017;25(2):267-70.
32. Beck JJ, Hottenga JJ, Mbarek H, Finnicum CT, Ehli EA, Hur YM, et al. *Genetic Similarity Assessment of Twin-Family Populations by Custom-Designed Genotyping Array.* Twin Res Hum Genet. 2019:1-10.
33. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience. 2015;4:7.
34. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. *Robust relationship inference in genome-wide association studies.* Bioinformatics. 2010;**26**(22):2867-73.
35. Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, et al. *Differences between germline genomes of monozygotic twins.* Nat Genet. 2021;**53**(1):27-34.
36. Ouwens KG, Jansen R, Tolhuis B, Slagboom PE, Penninx B, Boomsma DI. *A characterization of postzygotic mutations identified in monozygotic twins.* Hum Mutat. 2018;**39**(10):1393-401.
37. Price AL, Zaitlen NA, Reich D, Patterson N. *New approaches to population stratification in genome-wide association studies.* Nat Rev Genet. 2010;**11**(7):459-63.

38. Feng Q, Abraham J, Feng T, Song Y, Elston RC, Zhu X. *A method to correct for population structure using a segregation model.* BMC Proc. 2009;3 Suppl 7:S104.

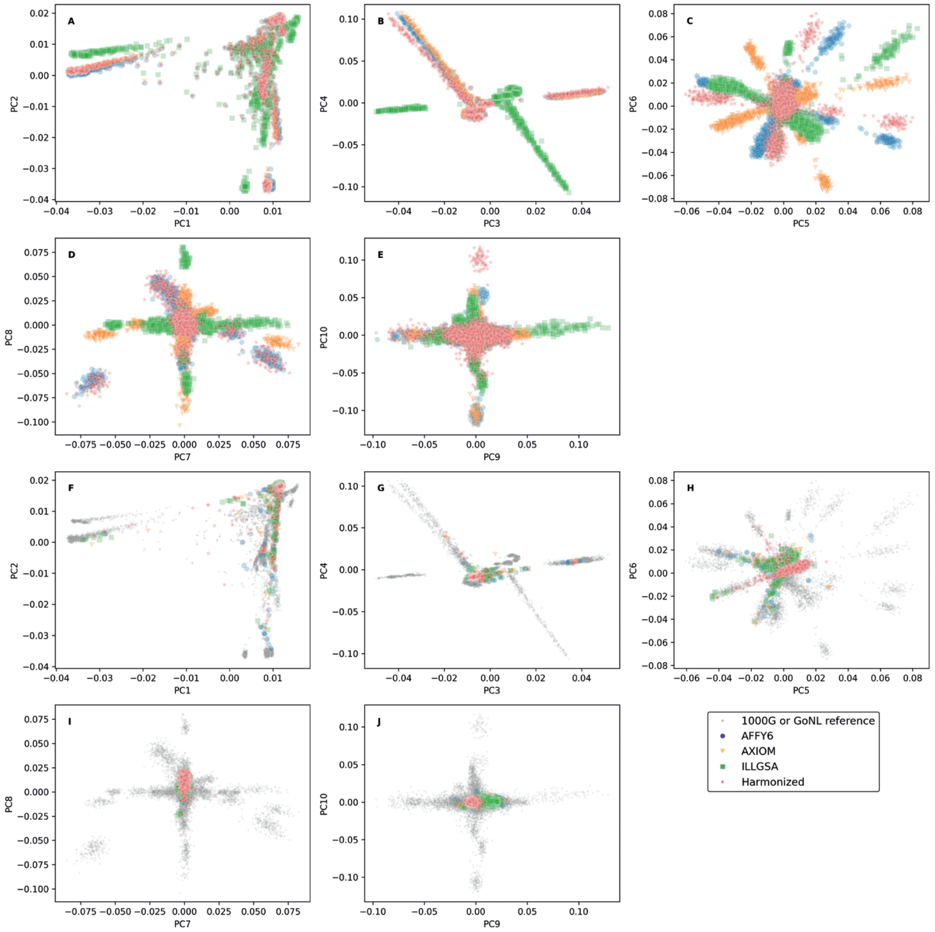
39. Kang SJ, Larkin EK, Song Y, Barnholtz-Sloan J, Baechle D, Feng T, et al. *Assessing the impact of global versus local ancestry in association studies.* BMC Proc. 2009;3 Suppl 7:S107.

40. Thornton T, Conomos MP, Sverdlov S, Blue EM, Cheung CY, Glazner CG, et al. *Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing.* BMC Proc. 2014;**8**(Suppl 1):S5.

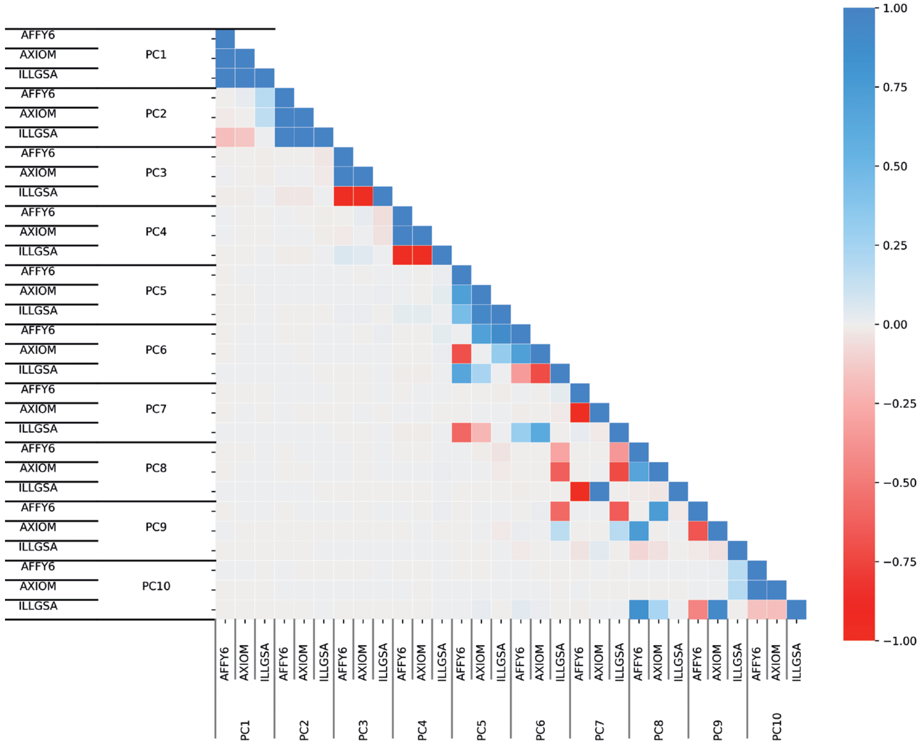
41. Zhu X, Li S, Cooper RS, Elston RC. *A unified association analysis approach for family and unrelated samples correcting for stratification.* American Journal of Human Genetics. 2008;**82**(2):352-65.

42. Conomos MP, Miller MB, Thornton TA. *Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness.* Genet Epidemiol. 2015;**39**(4):276-93.

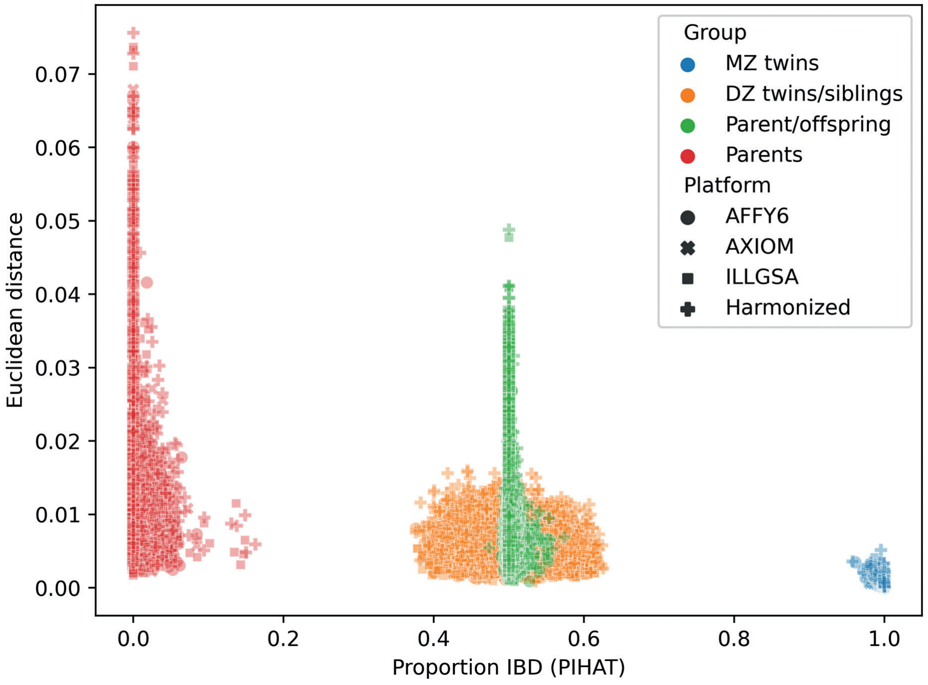
SUPPLEMENTARY MATERIALS



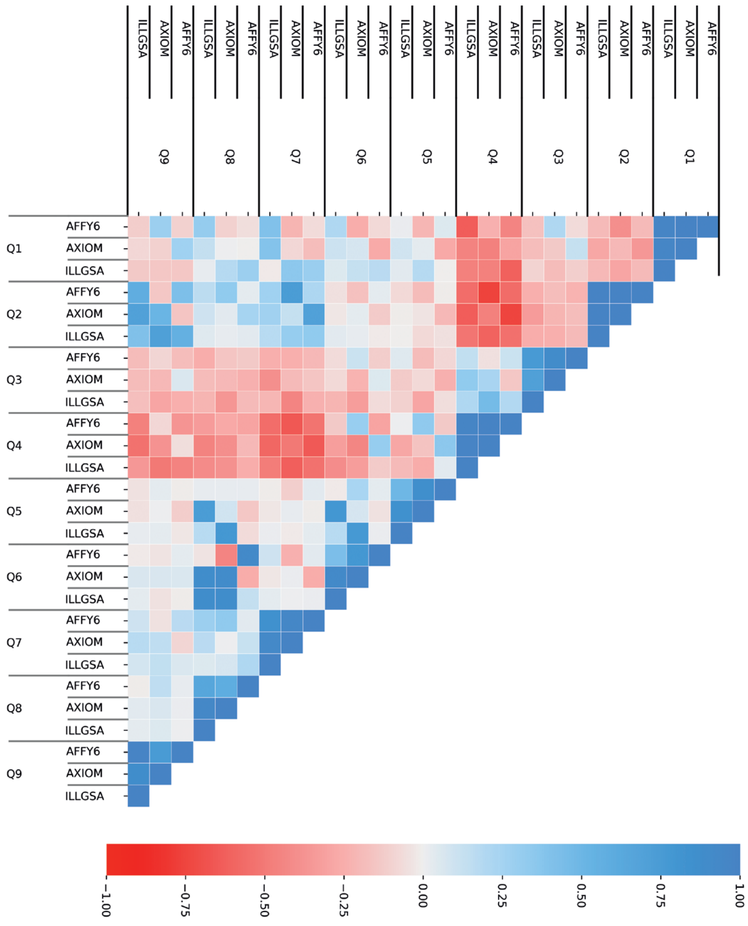
Supplementary Figure 6.1 –Scatterplots of PCs 1-10 per genotyping array. Panels A-E are for 1000G and GoNL reference samples only. Panels F-H are the NTR samples colored by genotyping array superimposed on the 1000G and GoNL reference samples (small gray points).



Supplementary Figure 6.2 – Correlations of PCs by genotyping array.



*Supplementary Figure 6.3 – Scatterplot of PC Euclidean distance as a function of IBD (identity-by-descent). PIHAT is calculated as the proportion(IBD=2) + 0.5*proportion(IBD-1).*



Supplementary Figure 6.4 – Correlations of admixture proportions by genotyping array.

Supplementary Table 6.1 – Descriptive statistics of principal components of NTR participants

	AFFY6					AXIOM					ILLGSA					HARMONIZED				
	Mean	**SD**	**Min**	**Max**	**Range**	**Mean**	**SD**	**Min**	**Max**	**Range**	**Mean**	**SD**	**Min**	**Max**	**Range**	**Mean**	**SD**	**Min**	**Max**	**Range**
PC1	0.0112	0.0015	-0.0338	0.0121	0.0459	0.0114	0.0022	-0.0313	0.0124	0.0437	0.0143	0.0022	-0.0352	0.0156	0.0508	0.0116	0.0019	-0.0359	0.0130	0.0489
PC2	0.0176	0.0037	-0.0344	0.0194	0.0538	0.0174	0.0038	-0.0342	0.0191	0.0533	0.0159	0.0037	-0.0365	0.0176	0.0541	0.0172	0.0037	-0.0368	0.0196	0.0564
PC3	-0.0070	0.0023	-0.0196	0.0395	0.0591	-0.0069	0.0029	-0.0191	0.0385	0.0576	0.0070	0.0023	-0.0422	0.0197	0.0619	-0.0068	0.0025	-0.0240	0.0431	0.0671
PC4	-0.0087	0.0015	-0.0161	0.0380	0.0541	-0.0085	0.002	-0.0147	0.0446	0.0593	0.0083	0.0017	-0.0500	0.0164	0.0664	-0.0089	0.0019	-0.0176	0.0475	0.0651
PC5	0.0020	0.0033	-0.0403	0.0284	0.0687	0.0059	0.0056	-0.0442	0.0266	0.0708	0.0061	0.0053	-0.0478	0.0501	0.0979	0.0004	0.0033	-0.0469	0.0423	0.0892
PC6	0.0069	0.0049	-0.0452	0.0327	0.0779	0.0035	0.0036	-0.0226	0.0401	0.0627	-0.0083	0.0041	-0.0350	0.0288	0.0638	0.0069	0.0057	-0.0525	0.0481	0.1006
PC7	-0.0002	0.0014	-0.0055	0.0052	0.0107	0.0002	0.0014	-0.0095	0.0222	0.0317	-0.0073	0.0041	-0.0289	0.0345	0.0634	-0.0005	0.0024	-0.0310	0.0197	0.0507
PC8	0.0049	0.0023	-0.0131	0.0134	0.0265	0.0113	0.005	-0.0251	0.0231	0.0482	-0.0002	0.0013	-0.0179	0.0332	0.0511	0.0035	0.0030	-0.0308	0.0163	0.0471
PC9	0.0112	0.0045	-0.0206	0.0303	0.0509	-0.0038	0.0022	-0.0112	0.0087	0.0199	0.0003	0.0025	-0.0181	0.0190	0.0371	0.0112	0.0054	-0.0309	0.0315	0.0624
PC10	0.0004	0.0023	-0.0085	0.0152	0.0237	0.0005	0.0022	-0.0123	0.0091	0.0214	-0.0006	0.0017	-0.0307	0.0109	0.0416	-0.0006	0.0034	-0.0174	0.0180	0.0354

PC1–PC10=principal components 1 through 10, SD=standard deviation, Min=minimum, Max=maximum

Supplementary Table 6.2 – Descriptive statistics of admixture proportions of NTR participants

	AFFY6					AXIOM					ILLGSA					HARMONIZED				
	Mean	**SD**	**Min**	**Max**	**Range**	**Mean**	**SD**	**Min**	**Max**	**Range**	**Mean**	**SD**	**Min**	**Max**	**Range**	**Mean**	**SD**	**Min**	**Max**	**Range**
Q1	0.1869	0.0587	0.0000	0.7468	0.7468	0.1908	0.0663	0.0039	0.7405	0.7366	0.1948	0.0596	0.0000	0.7194	0.7194	0.1878	0.0658	0.0000	0.7608	0.7608
Q2	0.0112	0.0527	0.0000	0.8731	0.8731	0.0114	0.0538	0.0000	0.8164	0.8164	0.0106	0.0505	0.0000	0.8937	0.8937	0.0113	0.0511	0.0000	0.8807	0.8807
Q3	0.0725	0.0220	0.0000	0.4332	0.4332	0.0726	0.0212	0.0000	0.3204	0.3204	0.0756	0.0206	0.0000	0.6184	0.6184	0.0702	0.0312	0.0000	0.6213	0.6213
Q4	0.6949	0.0919	0.0000	0.9314	0.9314	0.6868	0.1041	0.0000	0.8588	0.8588	0.6833	0.0954	0.0000	0.9218	0.9218	0.6936	0.1024	0.0000	0.9575	0.9575
Q5	0.0054	0.0100	0.0000	0.4484	0.4484	0.0054	0.0138	0.0000	0.5075	0.5074	0.0055	0.0117	0.0000	0.5525	0.5525	0.0060	0.0126	0.0000	0.5576	0.5575
Q6	0.0036	0.0154	0.0000	0.4681	0.4681	0.0049	0.0216	0.0000	0.4586	0.4585	0.0041	0.0186	0.0000	0.5581	0.5581	0.0043	0.0185	0.0000	0.5584	0.5584
Q7	0.0171	0.0419	0.0000	0.8840	0.8840	0.0189	0.0530	0.0000	0.8491	0.8491	0.0174	0.0400	0.0000	0.9001	0.9001	0.0173	0.0442	0.0000	0.9330	0.9330
Q8	0.0039	0.0167	0.0000	0.5172	0.5171	0.0049	0.0229	0.0000	0.5952	0.5952	0.0043	0.0197	0.0000	0.6412	0.6412	0.0044	0.0199	0.0000	0.6672	0.6672
Q9	0.0045	0.0144	0.0000	0.5465	0.5465	0.0042	0.0123	0.0000	0.3438	0.3438	0.0043	0.0205	0.0000	0.7515	0.7515	0.0050	0.0183	0.0000	0.7733	0.7733

Q1-Q9 represent each of the nine ancestry populations as determined by ADMIXTURE, SD=standard deviation, Min=minimum, Max=maximum

7

SUMMARY AND DISCUSSION

The amount of information obtained from DNA in molecular genetic laboratories, including our lab at the Avera Institute Human Genetics (AIHG), is enormous. The richness of these data and the conclusions that can be drawn from them have fueled applied genetic epidemiological research. With the completion of the Human Genome Project in 2003 [1] and improvements since its initial release [2, 3], genome-wide association studies (GWAS) have become the key gene discovery approach for providing insight into human traits and disease [4-6]. Now, more than 18 years after the completion of the first draft of the human genome, GWAS continue to be a primary method for identifying genetic variant-trait associations [7]. To date, more than 257,000 variant-trait associations have been identified and documented in the GWAS Catalog (https://www.ebi.ac.uk/gwas/). Despite the apparent benefit of linking genetic variants to human health and traits, the wealth of genetic data can also be leveraged to study populations. Genetic studies of populations have and will continue to inform about population substructure and inferences regarding human history [8]. Therefore, studies on the genetic composition of individuals, families, and populations will unlock additional information about the human condition.

The central aim of this thesis was to use molecular genetic data to improve our understanding of twinning, twins, their families, and the populations they represent. In this thesis, I begin with an extensive overview of the current knowledge of the biology and genetics of twinning. Next, I present a pedigree-based study that leverages whole-genome genotype and sequence data to identify novel biomarkers of human dizygotic twinning. Then, in what follows are three separate studies that incorporate large amounts of molecular genetic and phenotypic data obtained from twins and their families that participate in population-based twin registries worldwide. The common theme of the three studies is the usage of data from twins and their family members to make informed conclusions about genetic structure at the population level and the genetic architecture underlying a complex trait in twins.

In this final chapter, I summarize the main findings of each chapter of this thesis. I then conclude with a general discussion of human genetic studies, emphasizing gene finding for twinning and birth weight, as well as considerations of genetic ancestry testing. Lastly, I provide my future perspectives.

BIOLOGY AND GENETICS OF DIZYGOTIC AND MONOZYGOTIC TWINNING

Across religions, cultures, and societies, twins have sparked human curiosity for centuries. Twin births were considered an extraordinary phenomenon in ancient times, though they were likely limited due to pregnancy and birth complications. With clinical and medical improvements, successful twin pregnancies have become far more prevalent in more recent times. As a matter of fact, twin births can be the by-product of medical technologies designed to assist couples experiencing reproductive difficulties (i.e., assisted reproductive technologies) [9-11].

In the case of identical twins, inherent interest comes from their astonishing physical similarity yet unique personality traits. Alternatively, fraternal twins are captivating in their own way, representing a set of non-identical siblings born at the same time. Identical twins, which share close to 100% of their genetic material, are called monozygotic (MZ) twins, indicating they are derived from a single zygote. Fraternal twins, which on average share 50% of their genetic material, are dizygotic (DZ), meaning they formed from two independent zygotes.

In science, the degree of phenotypic resemblance between MZ and DZ twins is fundamental to studies aimed at deriving the influence of genetics on a particular trait, also known as trait heritability. In this way, twins allow for a natural experiment and form the basis of the twin study design. Studies involving twins have grown in popularity through the years, especially in the 'omics' era in which different facets of human biology are studied in the context of human traits and diseases [12]. The scientific interest in twins has largely been facilitated by the establishment of global twin registries, with the first being established in Denmark in the 1950s [13]. Since then, it has been estimated that more than 1.5 million twins, high-order multiples, and family members participate in twin registries worldwide [14, 15].

Twins are undoubtedly a valuable resource for scientific research. Twin registries have been instrumental in making significant contributions to studies of human disease susceptibility and healthy development and regarding determinants and correlates of complex traits. Because twins are born in all strata of society, they are representative of the general population [16]. Thus, twin research may also be used to further explore discoveries in the general population, as reflected in chapters 4, 5, and 6.

In chapter 2, an extensive overview of the biology and genetics of MZ and DZ twinning is provided. The focus of the chapter is a description of what is known and what remains to be understood about the human twinning process with related findings from animal studies. First, I present the biological similarities and differences of DZ and MZ twins concerning zygosity, chorionicity, and placentation and how this information relates to the traditional and widely adopted models of twinning [17]. Second, I describe the etiology of twinning from biological and genetic perspectives, with an important distinction between twin types. Third, the incidence and global variation in MZ and DZ twinning rates are outlined.

The remarkable overall consistency in MZ twinning rates worldwide suggests that MZ twinning is spontaneous and may be under the influence of some combination of genetic and non-genetic factors. Although comprehensive understanding has remained challenging, recent breakthroughs have discovered stable epigenetic signatures associated with MZ twinning that persist into adulthood [18]. The occurrence and rates of DZ twinning exhibit temporal and spatial variation, with significant variation attributable to maternal (i.e., genetic) factors. Important genetic developments and techniques have facilitated considerable scientific efforts to discover genes related to the DZ twinning process. Most notable is the identification and replication in 2016 of two genetic variants, near the *SMAD3* and within the *FSHB* genes, robustly associated with being a mother of spontaneous DZ twins (i.e., no use of assisted reproductive technologies) [19].

In conclusion, chapter 2 defines and differentiates between human MZ and DZ twinning from biological and genetic viewpoints. Even with the recent development and improvement of molecular techniques, the human twinning process remains to be fully characterized, with substantially less known about contributory factors of MZ twinning. As technological advancements continue, so too will our knowledge of the twinning process. Further elucidation of the genetic and non-genetic factors underpinning the twinning process will aid in the understanding of female fertility and improve upon the outcomes of multiple gestation pregnancies from predictive, supportive, and clinical care perspectives.

PEDIGREE BASED ANALYSIS OF HUMAN DIZYGOTIC TWINNING USING WHOLE-GENOME SEQUENCE DATA

Spontaneous DZ twinning in humans results from double ovulation and tends to run in families [20]. Although large regional differences in DZ twinning rates exist, approximately 1-4% of women worldwide are affected [21-23]. Over many decades, many have tried to elucidate its genetic basis, but with limited success. A recent gene-discovery effort brought together unique collections of mothers of DZ twins (MoDZT) to conduct the first genome-wide association meta-analysis of 1,980 MoDZT and 12,953 controls [19]. The study identified two polymorphisms, located near *FSHB* and within *SMAD3* genes, with significant effects on twinning and numerous measures of female fertility, including higher serum FSH levels, earlier age at menarche, earlier age at first child, higher lifetime parity, and earlier age at menopause. The findings from this study revealed important genetic aspects of reproductive capacity and health; however, the identified variants do not entirely capture the genetic disposition of having DZ twins. The lack of complete understanding underlying the genetics of DZ twinning provided an opportunity to uncover novel genetic variants not identifiable through common variant approaches.

As a follow-up approach to the promising results of the investigative twinning GWAS, we performed whole-genome genotyping and sequencing on mothers of DZ twins from a large multigenerational Dutch pedigree with a rich history of DZ twinning. The project's goal was to identify additional gene regions potentially containing rare and functional variants that increase the risk of DZ twinning and that index female fertility. We hypothesized that such a strategy could lead to the discovery of next-generation biomarkers of clinical importance for predicting DZ twinning and related to fertility measures.

Chapter 3 provides a detailed description of a large DZ twinning pedigree and the results of combined within-family linkage information and analysis of whole-genome sequence data. The large multigenerational pedigree contains 18 MoDZT and 21 sets of spontaneous DZ twins (no use of assisted reproductive technologies). Biological samples were available for 17 individuals (4 males and 13 females). Of the 13 females, 11 are MoDZT, of which one is part of an opposite-sex DZ twin pair. Of the MoDZT, two of them gave birth to two sets of DZ twins. DNA was extracted from the available samples and sent to the Avera Institute for Human Genetics for SNP genotyping on the Illumina GSA and whole-genome sequencing.

To locate DZ twinning-related genes in the pedigree, we employed linkage analysis using genotypic data. Broadly, linkage analysis aims to demonstrate co-segregation of genetic markers and a trait within families. Non-parametric linkage analysis with Merlin software [24] yielded a top hit on chromosome 5 (maximum LOD score=1.21, p=0.009). Though not a convincingly strong signal, the result indicated excess allele sharing among the affected MoDZT and favored the presence of linkage.

As a follow-up to the initial linkage work, four of the most distantly affected MoDZT in the pedigree were selected for whole-genome sequencing. Combined genotypic and sequence data from the four MoDZT were used for haplotype estimation and identification of shared genomic segments with Olorin software [25]. Sizeable segments present in at least three of the four MoDZT were present on nearly all chromosomes. The largest shared segments possessed by all four mothers were found on chromosomes 1, 3, 6, 11, making them regions of interest to identify rare driver variants of DZ twinning.

Identification of variants in the shared regions may reveal novel genetic biomarkers for DZ twinning. Coupling variants with corresponding functional consequence information obtained from Variant Effect Predictor may reaffirm or establish new biological mechanisms driving multiple ovulation and subsequent DZ twinning events. Screening of shared haplotypes and variants against an external cohort of MoDZT from the Genome of the Netherlands project [26] (46 MoDZT with whole genome-sequence data) will demonstrate if they are pedigree-specific or generally characteristic of being a MoDZT. Given the global variation in DZ twinning rates, it will also be impactful to assess the shared genetic regions and variants in more diverse populations, especially those with high twinning rates.

The continuing objective of this project is to use whole-genome sequencing as a follow-up approach to linkage and association studies of DZ twinning. Given the results of previous work, variants affecting multiple ovulation and subsequent DZ twinning events are likely to occur in genes and pathways controlling the synthesis of Follicle Stimulating Hormone (FSH), an important reproductive hormone. Pedigree-based whole-genome sequencing may implicate new pathways or novel regulation of FSH-related pathways. Though initial genomic regions of interest have been identified, additional work is needed to determine specific variants with meaningful biological effects. The findings from this work may someday enable new opportunities to treat infertility or optimize assisted reproductive strategies.

GENETIC SIMILARITY ASSESSMENT OF TWIN-FAMILY POPULATIONS

Genetic studies of human complex traits and disease have demonstrated the necessity of large sample sizes for achieving adequate statistical power and identifying reproducible discoveries [4, 5, 7]. One approach for attaining large sample sizes is to combine forces. That is, to aggregate data collected by different groups from around the world. Twin registries represent a premier source for acquiring data given their dedication to recruitment, longitudinal assessment, and collection of biological samples [27]. While population-based twin registers enable data aggregation, it is not always clear the extent to which these data are harmonizable or what extra precautions need to be considered. Thus, there is value in examining the degree of genetic similarity of population cohorts to be combined for genetic studies from a genetic standpoint.

In chapter 4, an assessment of genetic similarity for three global populations (Australian, Dutch, American) that are routinely combined for large association studies is described. In this chapter, empirical and quantitative measures of global genetic differentiation were estimated. Estimates of genetic similarity for Australians and the Dutch have been previously reported in a smaller study, suggesting that the populations are indeed similar, allowing for data aggregation for joint genetic analyses [28]. In their study, Sullivan et al. estimated the proportion of total genetic variability attributable to genetic differences between cohorts, known as Wright's fixation index (F_{ST}), based on 359 short tandem repeat (STR) polymorphisms and analysis of molecular variance. The F_{ST} between 519 Australian and 549 Dutch individuals was estimated at 0.30%, a value smaller than between many European groups. Their work concluded that it is reasonable to combine samples from Australian and Dutch cohorts in genetic studies. However, their results were based on modest sample sizes and numbers of genetic markers.

In chapter 4, we augmented the study design by adding a third major population from the Midwestern United States and attained much larger sample sizes. Genetic data were obtained from a DNA microarray, with more than 600,000 genetic markers interrogated per individual. With the available genome-wide data, similar methods for quantifying population genetic variability were applied. I hypothesized that estimates of genetic similarity between the three populations would be comparable to the previously published estimates, which indeed turned out to be the case.

The study design was enhanced by incorporating genotypes from a small number of individuals from Nigeria, representative of a globally distinct population compared to the three populations of predominantly European ancestry. I predicted that empirical estimates of genetic similarity between Australian, Dutch, and American populations would be more like each other than estimates with other globally diverse populations, including samples from Nigeria and a reference cohort of worldwide representative samples from the Human Genome Diversity Project (HGDP).

Another unique aspect of the study presented in chapter 4 is that although the biomaterials were collected from individuals from geographically distant areas of the globe, all genotypes were generated at the same laboratory (the AIHG) on the same DNA microarray (Illumina GSA). This particular characteristic of the study design is extremely beneficial as sample handling, shipping, and processing may affect sample quality and downstream genetic estimates [29]. Chapter 4 describes the design of the customized Illumina GSA. Included is a detailed breakdown of the content and marker selection categories, including markers related to clinical research (i.e., pharmacogenomics), human health, population research, and a core backbone of makers for optimizing imputation. Results of the initial validation studies of the Illumina GSA for imputation are presented, including coverage, overall imputation quality, and concordance checks between array-mimicked imputed data and sequence data. For these assessments, a dataset mimicking the content of the GSA was created from whole-genome sequence data of 249 unrelated females from the Genome of the Netherlands (GoNL) project. Results demonstrated that genotypes derived from the GSA could be re-imputed with high confidence, apart from very rare alleles that never impute well [30]. The imputation quality metrics of the GSA were consistent with previous reports from other custom-designed microarrays, including the earlier Affymetrix Axiom array, also collaboratively designed by the NTR and the AIHG [31].

With the genotypic data from the GSA, we employed various analytical strategies to assess population genetic similarity and differences, including principal components analysis (PCA) and F_{ST} estimates. PCA was used to visually compare the populations of interest to each other and globally diverse populations from the HGDP. After accounting for factors that can bias PCA, such as family structure and linkage disequilibrium, we found considerable overlap in the plotted top principal components (PCs) of Australian, Dutch, and American individuals. The superimposition of top PCs suggested genetic resemblance. No overlap in top PCs was observed with more diverse populations, such as those obtained from Asia or Africa. This finding agreed with previous work that has demonstrated strong correlations of top PCs with geography [32]. F_{ST} estimates were smaller between Australian, Dutch, and American populations than Nigerian populations, suggesting more genetic similarity between the three European ancestry-based populations. The F_{ST} estimates from two methods, Weir and Cockerham and Hudson estimators, were consistent with the geographical patterns observed from the PCA and with previous estimates of population genetic differentiation [26, 33].

In the context of GWAS and population structure, the findings and results of chapter 4 support the practice of aggregating data from ancestry matched (i.e., genetically similar) populations to minimize the degree of possible confounding while maximizing the power to detect association. While the populations under study showed strong genetic similarity, it is still of utmost importance to account for the genetic ancestry of individuals in a GWAS. In this regard, and when samples come from more diverse populations, correction for ancestry can be accomplished using an association model with appropriate covariates, where covariates (ideally) capture differences in genetic ancestry. This strategy is critical when more genetically diverse or admixed individuals are included to enhance detection power further.

GWAMA OF TWIN BIRTH WEIGHT AND COMPARISON TO GENETICS OF SINGLETON BIRTH WEIGHT

Studies of birth weight (BW) have been of great interest because BW is an important indicator of newborn and infant survival [34-36]. In addition, there is robust and well-replicated evidence for associations of BW extremes and adult health. For example, there is consistent evidence between low BW and adverse cardiovascular outcomes, such as heart disease, hypertension, stroke, and type 2 diabetes [37-40]. Furthermore, low BW has also been considered a risk factor for mood disorders and poorer cognitive ability [41, 42].

Given the robust relationships with various later-life conditions, studies in the last few years have sought to advance our understanding of the genetic architecture of BW through large-scale genome-wide association studies (GWAS) [43-46]. A common characteristic of these gene-finding studies is that they are performed with BW data obtained from singletons and not from twins. Twins tend to be excluded from discovery analyses because they typically have a lower BW than singletons, mainly due to a crowded intrauterine environment and a shortened gestational period. Only one genome-wide association study on BW in twins, specifically 4953 female twins, has been performed [47]. Therefore, there is a clear knowledge gap in our understanding of this trait in twins from a genetic perspective. Even more, it is unclear to what extent findings of BW from singletons can be generalized to twins and to what degree twins can contribute to gene discovery for BW. Thus, the goal of chapter 5 was to elucidate the genetics of BW in twins and to compare findings to those previously reported in singletons.

In chapter 5, a genome-wide association meta-analysis (GWAMA) was performed to gain insight into the genetics of BW in twins. In the study, eight population-based twin cohorts contributed results of a GWAS on twin BW considering important covariables, including sex, gestational age, year of birth, maternal age at birth, birth order, and other relevant study-specific metrics (e.g., ten PCs capturing ancestry differences or variation in the genotyping platform). The GWAS results from 42,212 twin individuals were subsequently meta-analyzed to compare twin and singleton BW genetics and address whether GWAS results from the two groups can be combined. A secondary goal was to identify genetic variants associated with twin BW.

In a first step, the performance of SNPs with genome-wide significant signals in the largest and most recent GWAS of singleton BW [46] was evaluated against the GWAMA results of twin BW. We compared the most strongly associated markers ($P<6.6x10^{-9}$) with singleton BW to those from our study by correlating the effect size estimates. Of the significant singleton BW variants, 150 of them overlapped with twin BW GWAMA results. We found a strong positive correlation (Pearson's r=0.66) between the 150 overlapping SNPs of the two studies. Although not in perfect unison, the strong positive linear relationship suggested that the previously reported genetic variants for BW behave similarly regardless of singleton or twin status. The likeness of the effect size estimates was the first promising indicator that the genetic influences on BW might be similar in singletons and twins.

No genome-wide significant genetic variants were identified in our GWAMA at the defined threshold of genome-wide significance ($P<5x10^{-8}$). Although two SNPs achieved a suggestive level of significance ($P<5x10^{-7}$). The two SNPs, rs10800682 and rs3845913, were located on chromosomes 1 and 3, respectively. Of the two signals, the former was independent (i.e., not in linkage disequilibrium) of sixty previously reported loci in large GWAS of singleton BW [43]. Although not significant, the novelty of rs10800682 makes this SNP a potential candidate for future studies of BW in twins. Alternatively, the latter SNP was found to be about 31 kilobases downstream of rs11719201, an intronic variant of the *ADCY5* gene, and one of 60 loci previously found to be robustly associated with BW in singletons [43]. This finding is supported by the fact that *ADCY5* and *CCNL1* were two of the first genes shown to be associated with fetal growth and BW [48]. Exactly how and through which gene(s) rs10800682 and rs3845913 may impart variation in BW is a promising avenue for future research.

For discerning the degree of genetic overlap between BW in twins and singletons, genetic correlations were calculated. In the simplest sense, genetic correlations represent the extent to which the same genes influence two traits. The two traits can be different phenotypes (e.g., BMI and blood pressure), the same trait measured at different ages, or the same trait measured in different groups (e.g., twins and non-twins). In this context, positive genetic correlations between twin and singleton BW indicate shared genetic contribution to BW irrespective of twin status. Strongly positive genetic correlations ($r_g>0.9$) were observed between the results of our GWAMA of twin BW and previously published GWAS results of BW in singletons from three sequential studies [43, 44, 46].

In the next step, we also looked at genetic correlations between BW in twins and a series of other health-related traits previously compared to BW in singletons. We found remarkably similar genetic correlations with these traits, particularly anthropometric traits, akin to those reported for singleton BW.

As a follow-up to the exciting genetic correlation findings, polygenic scores (PGS) for BW were also calculated. PGS represent the cumulative estimated effect of many genetic variants on an individual's phenotype, calculated as a weighted sum of trait-associated variants. Summary statistics (i.e., the effect size estimates per genetic variant) were derived from a UK Biobank discovery study on BW and were used to calculate PGS in an independent target sample of Dutch twins and singletons. The Dutch twins and singletons participate in the Netherlands Twin Register and had BW data available. The PGS were used to predict BW separately in twins and singletons after considering important confounding factors, such as family relationships, sex, year of birth, gestational age, genotyping platform, and principal components. Similar predictability of BW in singletons and twins with the optimal PGS was found. The model including PGS and important covariates explained 2% and 3% of BW variation in twins and non-twins, respectively. The likeness in the predictiveness provided yet another piece of evidence supporting a similar genetic architecture of BW in twins and singletons.

In conclusion, chapter 5 supports the inclusion of twins in genetic studies of BW, assuming proper analytical strategies are employed to account for the known differences in BW. The compelling results from various analytical approaches suggested a similar genetic profile of BW in twins and singletons. Given the continually expanding efforts to identify genetic variants associated with BW by initiatives such as the Early Growth Genetics Consortium (EGG) [49], the results of chapter 5 may considerably aid in detection efforts by boosting sample sizes through the incorporation of twins.

GENETIC ANCESTRY ESTIMATES IN TWINS AND FAMILIES

The popularity of population-based genetic association studies has increased throughout the years [5]. Their success is rooted in the continuing value of identifying genetic variant-trait associations related to human health and disease. One necessary component of genetic association studies is inferring ancestry from genetic data to correct for population stratification [50]. That is, allele frequency differences due to systematic ancestry differences. Population stratification can confound association studies and give rise to spurious results [51]. Many methods can determine genetic ancestry, but the approaches can generally be categorized into algorithmic or model-based approaches.

In chapter 6, the focus was on two scenarios that often arise in association studies. How do ancestry estimates using algorithmic and model-based approaches perform when family members participate in a study and is it possible that different decisions would be taken for individuals from the same family? We are especially interested in addressing this question in those individuals that are ancestrally diverse or admixed. Secondly, how comparable are ancestry estimates from those methods when study participants have been genotyped across multiple DNA microarrays? This question is relevant since large cohort studies are often faced with genotyping data obtained from different (e.g., earlier and later) microarrays. Genome-wide SNP data from twins and their family members participating in the Netherland's Twin Register were analyzed to answer these questions. All genotyping was performed at the AIHG on at least one of three genotyping platforms, namely Affymetrix6 (AFFY6), Affymetrix Axiom (AXIOM), or Illumina GSA (GSA). Genetic ancestry was determined using PCA (algorithmic) and a model-based approach, exemplified by the ADMIXTURE software package [52, 53].

PCA is a mathematical data dimensionality reduction technique. When employed on genetic data, PCA transforms a large set of correlated variables (i.e., genetic variants) into a smaller group of uncorrelated principal components (PCs) that capture most of the variation in the data. All individuals in the analysis then get a series of scores corresponding to each of the selected number of PCs. In genetic association studies, top PCs typically reflect population structure among the individuals in the analysis. PCs often correlate with geography, reflecting decreasing genetic similarity with increasing geographic distance [54, 55]. Recently, the correlation of PCs with geography and the general use of PCA in genetics has been scrutinized, albeit more of a warning of PCA abuse and misuse [56]. Thus, if PCA is improperly executed, PCs can reflect other artifacts of the data. As an example, rather than capturing population structure, PCs may also capture linkage disequilibrium structure [32, 57-59]. In the past several years, PCA has been increasingly utilized in population genetics for inferring genetic ancestry [50, 60], correcting for confounding due to population stratification [50], and understanding population composition and migration [32, 54, 61].

Model-based approaches can also elucidate population structure, one example being the software program ADMIXTURE. The premise of model-based methods is to calculate relative proportions of ancestry, knowing that the genetic composition of an individual is a mosaic of the ancestral populations they originate from. These approaches mostly use Bayesian or maximum likelihood estimation approaches to optimize the probability of observed genotypes by modeling ancestry proportions and population allele frequencies. The culminating result is ancestry proportions for each individual, where each proportion corresponds to the percentage of each ancestral population. In this way, ancestry proportions from ADMIXTURE are more directly interpretable than PCs since they resemble actual populations rather than some arbitrary axes of variation in the data (i.e., PCs).

Using global reference data from the 1000 Genomes Project [62] and the Genome of the Netherlands [26, 63], PCA and ADMIXTURE projection analyses were completed on NTR families, including MZ and DZ twins, siblings, parents, and parent-offspring pairs. Euclidean distances of PCs and ancestry proportions were calculated for pairs of family members. In this manner, Euclidean distances provided a singular quantitative value between two individuals across all PCs or ancestry proportions. Larger Euclidean distances indicated increasing dissimilarity across all PCs or ancestry proportions, whereas smaller Euclidean distances suggested more similarity. Even in diverse and admixed families, Euclidean distances of PCs and ancestry proportions closely resembled the degree of relatedness between individuals. That is, two genetically similar individuals (i.e., MZ twins share ~%100 of segregating alleles) had smaller Euclidean distances than less related family members (i.e., parent-offspring pairs that share precisely 50% of segregating alleles. Despite underlying analytical differences, the results were consistent with reports of high concordance between algorithmic and model-based approaches [53].

Ancestry estimates obtained from PCA and ADMIXTURE showed a negligible difference for individuals with genotypic information obtained from three different genotyping platforms. However, slightly larger differences were observed across array manufacturers (Affymetrix and Illumina) than within (Affymetrix 6.0 and Affymetrix Axiom). This finding, coupled with the results from within family pairs, demonstrated that ancestry analyses utilizing PCA or

ADMIXTURE software are robust to genotyping platforms, assuming the proper/same analytical steps are employed. This result is encouraging for cohorts having to aggregate genotypic data from different platforms.

In conclusion, chapter 6 assessed genetic ancestry estimates in many twin and family participants of the NTR. Genetic data generated at the AIHG from three different genotyping platforms were used to calculate genetic ancestry estimates using algorithmic (i.e., PCA) and model-based approaches, such as the one implemented in the software ADMIXTURE. When combined with genetic data from global reference panels in projection analyses, population structure estimates in the NTR between twins and family members reflected the genetic relatedness among them as would be expected given the law of segregation of alleles. The results provide confidence in commonly employed ancestry estimation strategies for family-based studies and those that combine genetic data from multiple microarrays.

GENERAL DISCUSSION

The emphasis of this thesis concerns genetically informed studies of twinning, twins, families, and populations from genome-wide molecular data. In what follows, I begin by discussing recent findings involving the twinning process since I have a particularly keen interest in gene-finding efforts related to dizygotic (DZ) twinning. With respect to these findings, I transition into recent developments that hold promise for supplementing and improving our understanding of the biology of the twinning process. Next, I turn to birth weight, focusing on more recent analytical strategies for improving our understanding of the phenotype. Then I broadly discuss genetic ancestry testing and the implications of results from these analyses for individuals and society. Following the ancestry discussion, I reflect on the importance of including cohorts of diverse ancestry in human genetic studies. Lastly, I close with my take on future perspectives of the field of human genetics.

GENE-FINDING FOR HUMAN TWINNING

Throughout my entire Ph.D. trajectory, I have worked with twins and their families in some capacity. When I first began my graduate studies and research projects at the Avera Institute for Human Genetics (AIHG), I dedicated a significant amount of time genotyping DNA extracted from cheek swabs of twins on the Affymetrix Axiom array. Around the same time, the AIHG established its twin register, the Avera Twin Register [64]. The formation of the twin register filled the local community with excitement and raised awareness for the importance of twins in genetic studies. The public attention resulted in fielding a wide array of questions about twins. A popular inquiry was about the likelihood of conceiving twins, a question that twin researchers and geneticists have been trying to answer for more than 40 years [65].

Fortuitously, a landmark paper was published shortly thereafter, which provided compelling and replicable evidence for two genetic variants involved in spontaneous dizygotic (DZ) twinning. The study, spearheaded by colleagues at the Netherlands Twin Register, identified the first common genetic variants associated with being a mother of spontaneous DZ twins (MoDZT), but that also appeared to influence many other female reproductive traits [19]. The variants, rs11031006 and rs17293443, are located near *FSHB* and within *SMAD3*, respectively. *FSHB* had been hypothesized but never shown to be implicated in DZ twinning. Alternatively, *SMAD3*, which regulates ovarian responsiveness to FSH, was not previously implicated in twinning. The findings were remarkable, replicable, robust, and identified the first common genetic variants related to female fertility, which could partly explain the inheritance of DZ twinning.

Around the same time as the identification of DZ twinning genes, I enrolled in a course on female biology and endocrinology to learn more about female fertility. Specifically, I desired to put in context the recent genetic findings related to twinning. I sought to understand better the biological mechanisms related to multiple ovulation and the potential for conceiving twins. While enrolled in the course, I began to work on a project associated with DZ twinning, focusing on identifying less common genetic variants associated with the trait. I utilized samples obtained from a large Dutch pedigree with a rich history of DZ twinning to complete SNP genotyping and whole-genome sequencing experiments. I presented preliminary sequencing experiment results at a global meeting in Singapore in November 2019 dedicated to research on the etiology of DZ and MZ twinning.

Since then, efforts to characterize the twinning phenotype from a genetic standpoint have expanded. For example, in 2019, I assisted with a follow-up project of UK Biobank participants that reported being part of a multiple birth [66]. The study replicated previously discovered DZ twinning genes, namely *FSHB* and *SMAD3*. A novel genetic variant (rs428022), close to two genes *PIAS1* and *SKOR1*, was also associated with multiple birth. *PIAS1*, a protein inhibitor of activated *STAT 1*, regulates the androgen receptor and has been implicated in prostate cancer [67-69]. It has also been shown that inhibitors of PIAS proteins interact with the transforming growth factor-beta pathway and regulate transcriptional activity mediated by SMAD proteins [70, 71]. Likewise, *SKOR1*, known as functional SMAD-suppressing element on chromosome 15, interacts with SMAD1, SMAD2, and SMAD3 to regulate bone morphogenetic protein (BMP) signaling [72]. The BMP protein family regulates many biological and developmental aspects of the reproductive system and has been shown to increase ovulation rate in sheep [73] and the marmoset monkey, which exhibits a high twinning rate [74]. The study reiterated some previous findings and provided additional insight into the biological and genetic etiology of multiple births and fertility.

Past epidemiological research has firmly established the complex inheritance of familial DZ twinning [75, 76], a hallmark and defining characteristic compared to monozygotic (MZ) twinning. Thus, discovery efforts have extended the phenotype to more than just MoDZT. Current and ongoing studies of DZ twinning now include the proxy phenotype of "Are you a DZ twin?". The combined set, including MoDZT and now DZ twins, helps improve the ability to detect new signals and identify novel biological and genetic associations through combined meta-analysis. Still an ongoing effort, multiple new loci have been identified, which all seem to have apparent implications in female reproductive function and endocrinological processes. Generally, identified genes converge on pathways involving hormone ligand-binding receptors and the ovulation cycle, including the hypothalamic-pituitary-gonadotrophin axis and intra-ovarian signaling. In a biological context, the findings from the combined meta-analysis are encouraging and provide reassurance for identifying genes that influence DZ twinning.

In contrast to the ongoing efforts and the promising results related to gene-finding for DZ twinning, the etiology of MZ twinning remains much less clear. Consistent with the prevailing hypothesis that MZ twinning occurs at random, attempts to identify genes associated with "being a MZ twin" have been less successful. Instead, attention has shifted to epigenetics since the MZ twinning event occurs early in development, coinciding with the same time as major epigenetic reprogramming. In fact, immediately following fertilization, the pre-implantation embryo undergoes multiple waves of global DNA methylation followed by *de novo* methylation. The methylation changes occur during the differentiation of pluripotent cells to specific cellular lineages and are crucial for embryonic development [77]. Given the substantial overlap, there has been recent interest in identifying DNA methylation signatures associated with MZ twinning.

Efforts to characterize epigenetic influences on MZ twinning have employed an epigenome-wide association study (EWAS) design. Analogous to GWAS, an EWAS is a powerful approach for identifying epigenetic signatures, such as DNA methylation, associated with a particular trait [78, 79]. The first EWAS for MZ twinning has generated a plethora of promising results [18]. Using DZ twins as controls to account for the unique prenatal effects of womb sharing, a strong association was found between MZ twinning and a DNA methylation signature in adult somatic tissues. The results revealed 834 differentially methylated sites associated with MZ twinning, which were not randomly distributed across the genome. The differential methylation existed near telomeres and centromeres, in transcriptionally repressed regions, and at putative sites of known inter-individual epigenetic modification (e.g., metastable epialleles). The robust DNA methylation signature can be used for retrospective diagnosis of MZ twinning, which could aid in investigations of known links between congenital disorders and MZ twinning. The first MZ twinning EWAS results are promising. They are beginning to illuminate potential biological mechanisms related to MZ twinning, a phenotype that has evaded scientific efforts to uncover its genetic basis.

Efforts to identify and characterize the genes for DZ twinning continue. The results of these investigations will have a profound impact on society since the global number of twins is increasing at an unprecedented level [21]. Much of the sharp rise in recent decades is due to medically assisted reproduction [23, 80]. However, increased family size and delayed childbearing (i.e., advanced maternal age at birth) have also contributed, particularly in higher-income areas of the world [81, 82]. Regardless, the surge is highly relevant since twin births are associated with higher infant and child mortality rates and increased prenatal and perinatal complications for the mother and fetus [83-85], especially in low-income countries [83]. Given the health implications for twins and mothers, a complete understanding of the twinning process will be imperative for predictive capabilities and improving the outcomes of multiple gestation pregnancies.

GENETICS AND (EPI)GENETICS OF BIRTH WEIGHT

An essential factor in newborn and infant survival is weight at birth [86]. It is well established that newborns at the high (i.e., macrosomia) and low (i.e., fetal growth restriction) ends of the population distribution for birth weight (BW) will have an increased risk of adverse health outcomes in adulthood [39, 87-90]. For this reason, BW has been studied extensively, with substantial effort dedicated to discerning the relative genetic and environmental influences on BW variation (Table 7.1). Genetic studies of BW are not straightforward since genetic effects can be direct because of the fetal genotype and indirect, acting through the maternal genotype (i.e., non-transmitted alleles) [91-95]. Of the conjoined effects, the latter represents genes that act via the intrauterine environment, more generally serving as a proxy for the environment.

Table 7.1 – Studies of genetic and environmental influences on BW variation

Contribution to BW (%)						
Fetal Genes	**Maternal Genes**	**Environment**	**Interaction**	**Other**	**Notes**	**Ref**
>50	<20	20–30 (random)	None			[96]
69.4	3	19 (random) and 8.6 (common to sibs)				[97]
70	12	18 (random)				[98]
60	None detected			Sex-limited effects (35 of males and 26 of females)	Null maternal effect due to sample	[99]
39 (heritability estimate)		5.2 (common, including maternal genotype)			Variation not associated with maternal and gestational age	[100]
42 (heritability estimate)		56 (random) and 3 (common)			Study design did not permit separation of fetal/maternal genetic effects	[101]
31.0	22.1	14.8 (common) and 32.2 (random)				[102]
25 (heritability estimate)						[103]
	40–52	Of the maternal influences, 72 (common) and 28 (random)				[104]
18	52	30 (unknown environment)				[105]
7.9	5.5				'Fetal and maternal factors'	[106]
0.32–1.52					7 SNPs	[44]
24	4		4 (fetal/ maternal covariance)			[43]

Table 7.1 – Studies of genetic and environmental influences on BW variation (continued)

Contribution to BW (%)						
Fetal Genes	**Maternal Genes**	**Environment**	**Interaction**	**Other**	**Notes**	**Ref**
1.4 (10 loci) and 11.1 (all autosomal variants)				7 of 10 identified maternal loci act via the intrauterine environment rather than via effects of shared alleles with the fetus		[45]
6	2		-0.5% fetal/ maternal covariance		209 lead SNPs	[46]
28.5	7.6		3.7 fetal/ maternal covariance		Genome-wide	[91]

To date, GWAS have identified and implicated 190 genomic loci with BW [43, 45, 46]. Classification of associated variants suggests that 75% of the identified loci show direct effects of the fetal genotype, a small proportion of which also exhibit maternal effects. The remaining SNPs show indirect effects of the maternal genotype, considered the non-transmitted alleles. These findings are consistent with a more recent haplotype-based approach in mother-child pairs, demonstrating that fetal size is primarily the result of the fetal genome, whereas the maternal genome determines the duration of gestation [107]. Further efforts have discerned that maternally non-transmitted alleles influencing offspring birthweight through the intrauterine environment are unlikely to be major determinants of later-life adverse cardiometabolic outcomes [94].

Although many genetic loci have been found to be associated with BW, the causal variants have not yet been identified. Many of the GWAS hits exist outside protein-coding regions, so it remains uncertain whether the functional variant they tag exerts its effect through the most proximal gene or some other gene(s) located elsewhere. Knowledge of the causal biological and genetic pathways underpinning BW will help us understand the life-course associations between infant BW and adult morbidity.

Like scientific investigations of twinning, recent work on BW has also sought to understand other sources of BW variation influencing the later-life risk of non-communicable disease, including epigenetic processes. DNA methylation represents one plausible mechanism linking BW to adult health outcomes, including advanced aging. Prenatal exposures and stressors may alter methylation status, detrimentally affecting development, and lead to preterm birth and reduced BW. Preterm birth has been shown to increase sensitivity to long-term epigenetic effects [108], including accelerated aging [109].

Knowledge of methylation status across the genome has been shown to yield promising predictive results for BW. For example, methylation scores for BW, weighted sums of the individual's methylation levels at selected sites within the genome, have been shown to explain nearly 2% of BW variance, compared to 0.4% captured by genetics alone [110]. Another study found that methylation and polygenic scores could capture 0.4% and 1.5% of BW variation, respectively [111]. Together these findings corroborate some equivocal combination of genetic and epigenetic factors influencing variation in BW.

Methylation scores extend beyond genetic vulnerability, such as that captured by polygenic scores. In addition to genetic influences on the trait, methylation scores may also capture environmental and stochastic influences and the effect of the trait itself on the score, reflecting a reciprocal effect. Thus, methylation status may be informative as biomarkers of environmental influences on BW. It could also serve as a potential target for therapies since methylation status is known to be modifiable. For BW, these therapies could eventually help attenuate the extremes of BW distribution.

Epigenetic studies of BW aim to describe the intrauterine environment by quantitatively characterizing (un)favorable developmental conditions leading to variation in BW. One study, an EWAS of a Scottish birth cohort of 1757 individuals with BW and DNA methylation data from whole blood, yielded the identification of one significant methylation site associated with BW [112]. In addition to discovering a local effect, the study also supported an association with global DNA methylation, as determined by associations between BW and epigenetic measures of biological age (e.g., telomere length). An important caveat of this study is that blood-based methylation was assessed during adulthood. For one, the findings may not generalize to other tissues. However, blood measures can track biological processes and contain biomarkers of inflammation, cardiovascular disease, cardiometabolic disease, all of which are relevant to BW. Secondly, the epigenetic marks were evaluated in adulthood, representing a snapshot of epigenetic signatures that are known to be temporally dynamic. Previous work has demonstrated that DNA methylation associated with BW may fade away with age [113]. Other studies have reported differences in methylation being a consequence of later life obesity rather than a cause [114]. Evaluation of persistent and variable methylation signatures across time could be achieved with longitudinal data, ultimately providing more reliable associations.

Another EWAS utilized DNA methylation data from 1,040 infants from the United Kingdom to investigate epigenetic signatures of BW in cord blood [115]. The result was the identification of 236 and 1230 differentially methylated sites and regions associated with BW, respectively. Many of the associated methylation markers were enriched for methylation sites previously associated with pregnancy complications and exposures, including gestational hypertension/pre-eclampsia, smoking, and maternal folic acid levels during pregnancy. These findings are important since pre-eclampsia is itself associated with reduced BW. The results of this study provided insight into developmental pathways affecting BW, over and above what has been elucidated from purely genetic studies. Furthermore, epigenetic studies of BW may suggest potential surrogate markers for identifying prenatal exposures, which may assist in individual risk stratification for later-life non-communicable disease.

Localization and characterization efforts of BW-associated loci are ongoing and will continue to improve our knowledge of genetic and environmental influences on the trait. Further assessment of the separation of fetal and maternal genetic effects will enhance our understanding of the regulation of BW and its connections with adult health outcomes, including cardiometabolic health. Investigations of the epigenetic mechanisms influencing BW may help elucidate later-life disease pathways and potentially lead to targetable and modifiable treatment strategies. Combined studies of various 'omics' data will help define links between the biological mechanisms affecting BW variation and later life disease and may even provide new insights for improved clinical management of in-utero development.

INFERENCE OF GENETIC ANCESTRY – APPLICATIONS AND CHALLENGES

Recent advances in genetics and genomics have brought forth new opportunities to study the ancestry composition of individuals and populations. Genetic ancestry testing has been applied in a variety of ways, including biomedical research [116], forensics [117], and genealogical research [118]. Knowledge of ancestry is also used for clinical decision-making, and pharmacogenomics [119] since ancestry is a potential risk factor for disease [120, 121]. Ancestry testing has also gained a foothold in the commercial genetics enterprise through direct-to-consumer (DTC) testing. DTC genetic testing offers an opportunity for individuals to learn about their athletic aptitude, ancestry, dating compatibility, health, nutrition, physical traits, wine preference, and a seemingly endless list of other personal attributes. Despite the popularity and excitement that health-related DTC genetic testing offers, ongoing concern has been expressed concerning ancestry testing initiatives [122-125].

From a research standpoint, anthropologists and population geneticists leverage genetic data and information in genetic databases to draw conclusions about population structure, history, and evolution. In this context, inferences have been made regarding human migratory events and differentiating selective pressures from demographic changes [32, 126-129]. Depending on the application, research-oriented ancestry testing usually makes ancestry inferences at the population level rather than the individual level, which is the basis of commercial ancestry testing. Consequently, ancestry inferences at the population level are inherently more robust to the imprecision and limitations associated with ancestry testing in individuals.

Genetic epidemiologists routinely employ some form of ancestry inference from genetic data for predominantly analytical reasons. Population stratification, the systematic differences in allele frequencies of (sub)populations can lead to confounding in association studies. Therefore, statistical biases related to population stratification need to be controlled for. In this context, principal component analysis (PCA) has been widely applied [50, 61, 130]. The expectation of PCA is that a small number of resulting coordinates (i.e., principal components) relate to the geographic origin of each individual for which highly-dimensional genetic data are supplied [131]. PCA has been firmly established in population genetics, given the remarkable genealogical interpretation of principal components [132]. Although commonly utilized, PCA is not without limitation. As one example, PCA will incorrectly ascribe a single origin that is intermediate of parental source populations for an admixed individual (i.e., offspring of parents from disparate populations). Drawbacks of PCA, along with recent denunciation of its misuse in population genetic studies [56], warrant the need for alternate ancestry estimation approaches, such as admixture estimation.

Inference of genetic ancestry from autosomal genetic data using admixture estimation strategies almost always relies on a model of discrete demes that individuals inherit proportions of their genome from. The demes are often called 'ancestral' or 'parental' populations. Thus, the goal of the admixture model approach is to estimate individual admixture proportions for each ancestral population. Despite being more robust than PCA in dealing with admixture, these models also have intrinsic limitations. For example, not all ancestral populations may be observable since the proxy reference populations strictly define them. For instance, Yoruba samples are frequently used for inferring African American ancestry, even though most African Americans derive their ancestry from other (West) African populations [133, 134]. Thus, the current Yoruba proxy population may not be well representative of diverse African ancestry. In instances like this, a poor proxy would result in compensation of proportions by adding to another ancestral population. Another caveat is when ancestral populations are missing. The admixture estimation algorithms of commonly utilized software programs may skew the results and force proportions depending on the applied reference populations. Regardless of its limitations, admixture estimation has dramatically advanced the field of ancestry inference from genetic data resulting in numerous scientific works regarding population structure and history.

The relationship of genetic ancestry to individual and population health is still not well understood. Increasing efforts to apply ancestry-specific [121, 135] and trans-ancestry [136, 137] study designs is helping close this gap. However, the currently limited understanding of the ancestry and health connection can manifest in severe consequences. From a basic biological standpoint, the extent to which disease risk is due to DNA sequence versus gene expression is not always apparent. Concerning DTC genetic testing, a customer may share the results with a healthcare provider and assume that the information is considered when receiving care. Going forward, the healthcare community must play an integral role in genetic ancestry inference since this practice may become more widespread given the increasing popularity of DTC ancestry testing. In short, it is clear that the relationships among genetic variation, genetic ancestry, ethnicity, and health are complex, highlighting an exciting area for investigation.

In summary, genetic ancestry inference requires enormous care both in commercial and research applications. Along with technical and analytical limitations, a host of social, ethical, and even psychological concerns on the individual and global level arise when considering the interpretation of ancestry testing results. From a commercial standpoint, variation exists in how the estimates are determined and presented to the customer. To operate in the customer's best interest, additional policy and regulation have been proposed in this space [123], though lab certification and accreditation are still not required. In the academic research realm, accountability and rigor are equally as crucial for genetic ancestry estimation. The field of population genetics is evolving to standardize and improve upon terminology, methodologies, and communication of research conclusions based on ancestry testing. Though challenges surrounding ancestry inference continue to persist, assessment of genetic variation undoubtedly provides a window into human history.

FUTURE PERSPECTIVES

The fields of genetic epidemiology and population genetics are advancing knowledge of the human condition at an astounding rate. With each passing year, progressively more exciting developments and remarkable discoveries are put forth. The range of these findings is extensive, and the last year and a half have been no exception despite the havoc caused by the coronavirus outbreak. The incredible response by the scientific community, including our lab at the AIHG, has been nothing short of astonishing. I will never forget the opportunity to immediately impact our local community by providing COVID-19 testing for the Avera Healthcare System and the state of South Dakota. Even with the chaos surrounding the coronavirus pandemic, many other fascinating findings, technological developments, and methodological improvements in the field of genetics have amounted. In what follows, I provide future perspectives broadly related to the work presented in this thesis.

TWINNING

Ongoing meta- and mega-analyses of different aspects of human twinning, especially DZ twinning, will continue to reveal insight into common genetic variation underlying this trait. Moving forward, increased sample sizes and proxy phenotypes, including mothers of DZ twins and DZ twins themselves, will augment the ability to identify genetic associations. The effects of current findings related to *FSHB* and *SMAD3* have been proposed describing fluctuations in population twinning rates based on the number of risk alleles possessed by females [138]. However, these translations are population estimates and are not yet suited for prediction at the individual level. More extensive studies may facilitate these predictions in the future.

Rare and structural genetic variation associated with human twining also represents a promising area of investigation. Sequencing projects of mothers of DZ twins, particularly in large, carefully phenotyped pedigrees, may be a suitable approach for identifying rare variants associated with human DZ twinning. In this manner, the entire genomes of mothers of twins and potential carriers could be analyzed to uncover genetic variants that cannot be identified with the current microarray and imputation strategies. Larger scale sequencing studies, those extending beyond pedigrees, may further aid in the elucidation of rare alleles driving DZ twinning. If performed in diverse populations, we may uncover rare genomic sites or regions that impact the global variation in DZ twinning rates.

BIRTH WEIGHT

BW-oriented consortia, such as the Early Growth Genetics Consortium (http://egg-consortium.org/), continue to expand and increase sample numbers for improved power for detection for genomic loci impacting early growth. I sincerely hope that twins will be incorporated into future gene-finding association studies, given the strong genetic correlations we reported in chapter 4. The addition of twins will drastically bolster sample sizes because of diligent genotyping and phenotyping efforts of twin registers from around the world. Although an obvious benefit for increasing statistical power, careful analytical decisions must be made to obtain reliable results. Examples include correcting for familial relatedness and accounting for apparent differences in BW between twins and singletons. I envision the optimal strategy consisting of separate GWAS in singletons and twins, which would then be combined for meta-analysis of P-values since effect size estimates will drastically vary between groups. An alternate option would be to standardize BW within each group before meta-analysis. Regardless of the approach, careful consideration will be needed to properly account for the known BW differences between twins and singletons and the implications therein.

We continue to learn more about the maternal and fetal genetic effects influencing BW by employing clever and informative study designs. In addition to the strategies already implemented, it would be interesting to predict BW using polygenic scores (PGS) of both transmitted and untransmitted alleles from maternal and paternal lineages. In this way, direct genetic effects could be distinguished from maternal and paternal indirect genetic effects, of which the latter would be expected to be (close to) null. This analytical strategy would permit the estimation of environmentally mediated effects of PGS on BW.

In chapter 4, a lower effect of the PGS was observed for twins compared to singletons. This potentially indicates a form of sibling competition or interaction. Detection strategies for such competition/interaction have been applied to educational attainment, though no evidence was found to support this phenomenon [139]. Concerning BW, the genes of one (larger) twin could influence the ability of the co-twin to achieve its full genetic potential for growth. The larger twin thereby limits the space available for the development of the co-twin, consequently decreasing its BW. The result is dampened predictive power of the PGS in twins where this type of competition potentially occurs. Future studies using PGS predictions may expand on our initial findings of the possibility of sibling competition/interaction in terms of BW.

GENETIC ANCESTRY

The opportunities, challenges, and implications of genetic ancestry testing are apparent and continually changing. The importance of including ancestrally diverse cohorts in large-scale analyses has been made abundantly clear in genetics research. Historically, most genomic studies have examined individuals of European ancestry; however, recent findings from trans-ancestry studies indicate that more association signals are discoverable when diverse populations are included. The ongoing disparity has resulted in a lack of sufficient data for genomics and health-related research. Therefore, it is imperative that studies include more diverse populations to improve our understanding of different traits and diseases, not just in specific populations but in general. The inclusion of globally diverse or comparatively under-studied populations will facilitate trans-ancestry comparisons that will produce relevant results worldwide.

REFERENCES

1. Green, E.D., J.D. Watson, and F.S. Collins, *Human Genome Project: Twenty-five years of big biology.* Nature, 2015. **526**(7571): p. 29-31.
2. Nurk, S., et al., *The complete sequence of a human genome.* bioRxiv, 2021: p. 2021.05.26.445798.
3. Aganezov, S., et al., *A complete reference genome improves analysis of human genetic variation.* bioRxiv, 2021: p. 2021.07.12.452063.
4. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
5. Visscher, P.M., et al., *10 Years of GWAS Discovery: Biology, Function, and Translation.* Am J Hum Genet, 2017. **101**(1): p. 5-22.
6. Tam, V., et al., *Benefits and limitations of genome-wide association studies.* Nat Rev Genet, 2019.
7. Loos, R.J.F., *15 years of genome-wide association studies and no signs of slowing down.* Nat Commun, 2020. **11**(1): p. 5900.
8. Cavalli-Sforza, L.L., *Interview with Luigi Luca Cavalli-Sforza: past research and directions for future investigations in human population genetics. Interview by Franz Manni.* Hum Biol, 2010. **82**(3): p. 245-66.
9. Fauser, B.C., P. Devroey, and N.S. Macklon, *Multiple birth resulting from ovarian stimulation for subfertility treatment.* Lancet, 2005. **365**(9473): p. 1807-16.
10. Martin, J.A., et al., *Births: final data for 2003.* Natl Vital Stat Rep, 2005. **54**(2): p. 1-116.
11. Sunderam, S., et al., *Assisted Reproductive Technology Surveillance - United States, 2015.* MMWR Surveill Summ, 2018. **67**(3): p. 1-28.
12. van Dongen, J., et al., *The continuing value of twin studies in the omics era.* Nat Rev Genet, 2012. **13**(9): p. 640-53.
13. Pedersen, D.A., et al., *The Danish Twin Registry: An Updated Overview.* Twin Res Hum Genet, 2019. **22**(6): p. 499-507.
14. Hur, Y.M., et al., *Twin Family Registries Worldwide: An Important Resource for Scientific Research.* Twin Res Hum Genet, 2019. **22**(6): p. 427-437.
15. Hur, Y.M. and J.M. Craig, *Twin registries worldwide: an important resource for scientific research.* Twin Res Hum Genet, 2013. **16**(1): p. 1-12.
16. Martin, N., D. Boomsma, and G. Machin, *A twin-pronged attack on complex traits.* Nat Genet, 1997. **17**(4): p. 387-92.
17. McNamara, H.C., et al., *A review of the mechanisms and evidence for typical and atypical twinning.* Am J Obstet Gynecol, 2016. **214**(2): p. 172-191.
18. van Dongen, J., et al., *Identical twins carry a persistent epigenetic signature of early genome programming.* Nat Commun, 2021. **12**(1): p. 5618.
19. Mbarek, H., et al., *Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility.* Am J Hum Genet, 2016. **98**(5): p. 898-908.
20. Hoekstra, C., et al., *Body composition, smoking, and spontaneous dizygotic twinning.* Fertil Steril, 2010. **93**(3): p. 885-93.
21. Monden, C., G. Pison, and J. Smits, *Twin Peaks: more twinning in humans than ever before.* Hum Reprod, 2021.

22. Smits, J. and C. Monden, *Twinning across the Developing World.* PLoS One, 2011. **6**(9): p. e25239.
23. Hoekstra, C., et al., *Dizygotic twinning.* Hum Reprod Update, 2008. **14**(1): p. 37-47.
24. Abecasis, G.R., et al., *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees.* Nat Genet, 2002. **30**(1): p. 97-101.
25. Morris, J.A. and J.C. Barrett, *Olorin: combining gene flow with exome sequencing in large family studies of complex disease.* Bioinformatics, 2012. **28**(24): p. 3320-1.
26. Genome of the Netherlands, C., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet, 2014. **46**(8): p. 818-25.
27. Odintsova, V.V., et al., *Establishing a Twin Register: An Invaluable Resource for (Behavior) Genetic, Epidemiological, Biomarker, and 'Omics' Studies.* Twin Res Hum Genet, 2018. **21**(3): p. 239-252.
28. Sullivan, P.F., et al., *Empirical evaluation of the genetic similarity of samples from twin registries in Australia and the Netherlands using 359 STRP markers.* Twin Res Hum Genet, 2006. **9**(4): p. 600-2.
29. Finnicum, C.T., et al., *Relative Telomere Repeat Mass in Buccal and Leukocyte-Derived DNA.* PLoS One, 2017. **12**(1): p. e0170765.
30. Zheng, H.F., et al., *Performance of genotype imputation for low frequency and rare variants from the 1000 genomes.* PLoS One, 2015. **10**(1): p. e0116487.
31. Ehli, E.A., et al., *A method to customize population-specific arrays for genome-wide association testing.* Eur J Hum Genet, 2017. **25**(2): p. 267-270.
32. Abdellaoui, A., et al., *Population structure, migration, and diversifying selection in the Netherlands.* European journal of human genetics : EJHG, 2013. **21**(11): p. 1277-85.
33. International HapMap, C., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-8.
34. Wilcox, A.J., *On the importance--and the unimportance--of birthweight.* Int J Epidemiol, 2001. **30**(6): p. 1233-41.
35. Wilcox, A.J., *Birth weight and perinatal mortality: the effect of maternal smoking.* Am J Epidemiol, 1993. **137**(10): p. 1098-104.
36. Wilcox, A.J. and I.T. Russell, *Birthweight and perinatal mortality: II. On weight-specific mortality.* Int J Epidemiol, 1983. **12**(3): p. 319-25.
37. Huxley, R., et al., *Is birth weight a risk factor for ischemic heart disease in later life?* Am J Clin Nutr, 2007. **85**(5): p. 1244-50.
38. Zanetti, D., et al., *Birthweight, Type 2 Diabetes Mellitus, and Cardiovascular Disease: Addressing the Barker Hypothesis With Mendelian Randomization.* Circ Genom Precis Med, 2018. **11**(6): p. e002054.
39. Mu, M., et al., *Birth weight and subsequent blood pressure: a meta-analysis.* Arch Cardiovasc Dis, 2012. **105**(2): p. 99-113.
40. Rich-Edwards, J.W., et al., *Longitudinal study of birth weight and adult body mass index in predicting risk of coronary heart disease and stroke in women.* BMJ, 2005. **330**(7500): p. 1115.
41. Shenkin, S.D., J.M. Starr, and I.J. Deary, *Birth weight and cognitive ability in childhood: a systematic review.* Psychol Bull, 2004. **130**(6): p. 989-1013.
42. Wojcik, W., et al., *Foetal origins of depression? A systematic review and meta-analysis of low birth weight and later depression.* Psychol Med, 2013. **43**(1): p. 1-12.

43. Horikoshi, M., et al., *Genome-wide associations for birth weight and correlations with adult disease.* Nature, 2016. **538**(7624): p. 248-252.
44. Horikoshi, M., et al., *New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism.* Nature genetics, 2013. **45**(1): p. 76-82.
45. Beaumont, R.N., et al., *Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics.* Hum Mol Genet, 2018. **27**(4): p. 742-756.
46. Warrington, N.M., et al., *Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors.* Nat Genet, 2019. **51**(5): p. 804-814.
47. Metrustry, S.J., et al., *Variants close to NTRK2 gene are associated with birth weight in female twins.* Twin research and human genetics : the official journal of the International Society for Twin Studies, 2014. **17**(4): p. 254-61.
48. Freathy, R.M., et al., *Variants in ADCY5 and near CCNL1 are associated with fetal growth and birth weight.* Nature genetics, 2010. **42**(5): p. 430-5.
49. Middeldorp, C.M., et al., *The Early Growth Genetics (EGG) and EArly Genetics and Lifecourse Epidemiology (EAGLE) consortia: design, results and future prospects.* Eur J Epidemiol, 2019. **34**(3): p. 279-300.
50. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
51. Hellwege, J.N., et al., *Population Stratification in Genetic Association Studies.* Curr Protoc Hum Genet, 2017. **95**: p. 1 22 1-1 22 23.
52. Alexander, D.H. and K. Lange, *Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.* BMC Bioinformatics, 2011. **12**: p. 246.
53. Alexander, D.H., J. Novembre, and K. Lange, *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res, 2009. **19**(9): p. 1655-64.
54. Novembre, J., et al., *Genes mirror geography within Europe.* Nature, 2008. **456**(7218): p. 98-101.
55. Yang, W.Y., et al., *A model-based approach for analysis of spatial structure in genetic data.* Nat Genet, 2012. **44**(6): p. 725-31.
56. Elhaik, E., *Why most Principal Component Analyses (PCA) in population genetic studies are wrong.* bioRxiv, 2021: p. 2021.04.11.439381.
57. Prive, F., et al., *Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr.* Bioinformatics, 2018. **34**(16): p. 2781-2787.
58. Zou, F., et al., *Quantification of population structure using correlated SNPs by shrinkage principal components.* Hum Hered, 2010. **70**(1): p. 9-22.
59. Price, A.L., et al., *Long-range LD can confound genome scans in admixed populations.* Am J Hum Genet, 2008. **83**(1): p. 132-5; author reply 135-9.
60. Novembre, J. and A. Di Rienzo, *Spatial patterns of variation due to natural selection in humans.* Nat Rev Genet, 2009. **10**(11): p. 745-55.
61. Reich, D., A.L. Price, and N. Patterson, *Principal component analysis of genetic data.* Nat Genet, 2008. **40**(5): p. 491-2.
62. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
63. Boomsma, D.I., et al., *The Genome of the Netherlands: design, and project goals.* Eur J Hum Genet, 2014. **22**(2): p. 221-7.

64. Kittelsrud, J., et al., *Establishment of the Avera Twin Register in the Midwest USA.* Twin Res Hum Genet, 2017. **20**(5): p. 414-418.

65. Boomsma, D.I., *The Genetics of Human DZ Twinning.* Twin Res Hum Genet, 2020. **23**(2): p. 74-76.

66. Mbarek, H., et al., *Biological insights into multiple birth: genetic findings from UK Biobank.* Eur J Hum Genet, 2019. **27**(6): p. 970-979.

67. Heinlein, C.A. and C. Chang, *Androgen receptor in prostate cancer.* Endocr Rev, 2004. **25**(2): p. 276-308.

68. Hoefer, J., et al., *PIAS1 is increased in human prostate cancer and enhances proliferation through inhibition of p21.* Am J Pathol, 2012. **180**(5): p. 2097-107.

69. Puhr, M., et al., *PIAS1 is a determinant of poor survival and acts as a positive feedback regulator of AR signaling through enhanced AR stabilization in prostate cancer.* Oncogene, 2016. **35**(18): p. 2322-32.

70. Liang, M., et al., *Regulation of Smad4 sumoylation and transforming growth factor-beta signaling by protein inhibitor of activated STAT1.* J Biol Chem, 2004. **279**(22): p. 22857-65.

71. Long, J., et al., *Repression of Smad transcriptional activity by PIASy, an inhibitor of activated STAT.* Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9791-6.

72. Arndt, S., et al., *Fussel-15, a novel Ski/Sno homolog protein, antagonizes BMP signaling.* Mol Cell Neurosci, 2007. **34**(4): p. 603-11.

73. Fabre, S., et al., *Regulation of ovulation rate in mammals: contribution of sheep genetic models.* Reprod Biol Endocrinol, 2006. **4**: p. 20.

74. Harris, R.A., et al., *Evolutionary genetics and implications of small size and twinning in callitrichine primates.* Proc Natl Acad Sci U S A, 2014. **111**(4): p. 1467-72.

75. Painter, J.N., et al., *A genome wide linkage scan for dizygotic twinning in 525 families of mothers of dizygotic twins.* Hum Reprod, 2010. **25**(6): p. 1569-80.

76. Hoekstra, C., et al., *Familial twinning and fertility in Dutch mothers of twins.* Am J Med Genet A, 2008. **146A**(24): p. 3147-56.

77. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.

78. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-41.

79. Lappalainen, T. and J.M. Greally, *Associating cellular epigenetic models with human phenotypes.* Nat Rev Genet, 2017. **18**(7): p. 441-451.

80. de Mouzon, J., et al., *Assisted reproductive technology in Europe, 2007: results generated from European registers by ESHRE.* Hum Reprod, 2012. **27**(4): p. 954-66.

81. Duncan, J.M., *On Some Laws of the Production of Twins.* Edinb Med J, 1865. **10**(9): p. 767-781.

82. Pison, G. and A.V. D'Addato, *Frequency of twin births in developed countries.* Twin Res Hum Genet, 2006. **9**(2): p. 250-9.

83. Monden, C.W.S. and J. Smits, *Mortality among twins and singletons in sub-Saharan Africa between 1995 and 2014: a pooled analysis of data from 90 Demographic and Health Surveys in 30 countries.* Lancet Glob Health, 2017. **5**(7): p. e673-e679.

84. Santana, D.S., et al., *Twin Pregnancy and Severe Maternal Outcomes: The World Health Organization Multicountry Survey on Maternal and Newborn Health.* Obstet Gynecol, 2016. **127**(4): p. 631-641.

85. Santana, D.S., et al., *Perinatal outcomes in twin pregnancies complicated by maternal morbidity: evidence from the WHO Multicountry Survey on Maternal and Newborn Health.* BMC Pregnancy Childbirth, 2018. **18**(1): p. 449.

86. Lawn, J.E., et al., *Every Newborn: progress, priorities, and potential beyond survival.* Lancet, 2014. **384**(9938): p. 189-205.

87. Wang, S.F., et al., *Birth weight and risk of coronary heart disease in adults: a meta-analysis of prospective cohort studies.* J Dev Orig Health Dis, 2014. **5**(6): p. 408-19.

88. Wang, T., et al., *Birth Weight and Stroke in Adult Life: Genetic Correlation and Causal Inference With Genome-Wide Association Data Sets.* Front Neurosci, 2020. **14**: p. 479.

89. Whincup, P.H., et al., *Birth weight and risk of type 2 diabetes: a systematic review.* JAMA, 2008. **300**(24): p. 2886-97.

90. Zhao, Y., et al., *Birth weight and overweight/obesity in adults: a meta-analysis.* Eur J Pediatr, 2012. **171**(12): p. 1737-46.

91. Warrington, N.M., et al., *Using structural equation modelling to jointly estimate maternal and fetal effects on birthweight in the UK Biobank.* Int J Epidemiol, 2018. **47**(4): p. 1229-1241.

92. Eaves, L.J., et al., *Resolving the effects of maternal and offspring genotype on dyadic outcomes in genome wide complex trait analysis ("M-GCTA").* Behav Genet, 2014. **44**(5): p. 445-55.

93. Chen, J., et al., *Dissecting maternal and fetal genetic effects underlying the associations between maternal phenotypes, birth outcomes, and adult phenotypes: A mendelian-randomization and haplotype-based genetic score analysis in 10,734 mother-infant pairs.* PLoS Med, 2020. **17**(8): p. e1003305.

94. Moen, G.H., et al., *Mendelian randomization study of maternal influences on birthweight and future cardiometabolic risk in the HUNT cohort.* Nat Commun, 2020. **11**(1): p. 5404.

95. Evans, D.M., et al., *Elucidating the role of maternal environmental exposures on offspring health and disease using two-sample Mendelian randomization.* Int J Epidemiol, 2019. **48**(3): p. 861-875.

96. Magnus, P., *Causes of variation in birth weight: a study of offspring of twins.* Clinical genetics, 1984. **25**(1): p. 15-24.

97. Magnus, P., *Further evidence for a significant effect of fetal genes on variation in birth weight.* Clin Genet, 1984. **26**(4): p. 289-96.

98. Magnus, P., *Distinguishing fetal and maternal genetic effects on variation in birth weight.* Acta Genet Med Gemellol (Roma), 1984. **33**(3): p. 481-6.

99. Magnus, P., et al., *Parental determinants of birth weight.* Clin Genet, 1984. **26**(5): p. 397-405.

100. Vlietinck, R., et al., *Genetic and environmental variation in the birth weight of twins.* Behav Genet, 1989. **19**(1): p. 151-61.

101. Clausson, B., P. Lichtenstein, and S. Cnattingius, *Genetic influence on birthweight and gestational length determined by studies in offspring of twins.* BJOG : an international journal of obstetrics and gynaecology, 2000. **107**(3): p. 375-81.

102. Lunde, A., et al., *Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data.* Am J Epidemiol, 2007. **165**(7): p. 734-41.

103. Magnus, P., et al., *Paternal contribution to birth weight.* J Epidemiol Community Health, 2001. **55**(12): p. 873-7.

104. Nance, W.E., et al., *A causal analysis of birth weight in the offspring of monozygotic twins.* Am J Hum Genet, 1983. **35**(6): p. 1211-23.

105. Penrose, L.S., *Some recent trends in human genetics.* Caryologia, 1954. **6 (suppl)**: p. 520-530.

106. Langhoff-Roos, J., et al., *Relative effect of parental birth weight on infant birth weight at term.* Clin Genet, 1987. **32**(4): p. 240-8.

107. Srivastava, A.K., et al., *Haplotype-based heritability estimations reveal gestational duration as a maternal trait and fetal size measurements at birth as fetal traits in human pregnancy.* bioRxiv, 2020: p. 2020.05.12.079863.

108. Mathewson, K.J., et al., *DNA methylation profiles in adults born at extremely low birth weight.* Dev Psychopathol, 2020: p. 1-18.

109. Van Lieshout, R.J., et al., *Extremely Low Birth Weight and Accelerated Biological Aging.* Pediatrics, 2021. **147**(6).

110. Reed, Z.E., et al., *The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers.* Clin Epigenetics, 2020. **12**(1): p. 50.

111. Odintsova, V.V., et al., *Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles.* Frontiers in Psychiatry, 2021. **12**: p. 1141.

112. Madden, R.A., et al., *Birth weight associations with DNA methylation differences in an adult population.* Epigenetics, 2020: p. 1-14.

113. Kupers, L.K., et al., *Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight.* Nat Commun, 2019. **10**(1): p. 1893.

114. Vehmeijer, F.O.L., et al., *DNA methylation and body mass index from birth to adolescence: meta-analyses of epigenome-wide association studies.* Genome Med, 2020. **12**(1): p. 105.

115. Antoun, E., et al., *DNA methylation signatures in cord blood associated with birthweight are enriched for dmCpGs previously associated with maternal hypertension or pre-eclampsia, smoking and folic acid intake.* Epigenetics, 2021: p. 1-17.

116. Choudhry, S., et al., *Genome-wide screen for asthma in Puerto Ricans: evidence for association with 5q23 region.* Hum Genet, 2008. **123**(5): p. 455-68.

117. Budowle, B. and A. van Daal, *Forensically relevant SNP classes.* Biotechniques, 2008. **44**(5): p. 603-8, 610.

118. Shriver, M.D. and R.A. Kittles, *Genetic ancestry and the search for personalized genetic histories.* Nat Rev Genet, 2004. **5**(8): p. 611-8.

119. Yang, H.C., et al., *Genetic ancestry plays a central role in population pharmacogenomics.* Commun Biol, 2021. **4**(1): p. 171.

120. Flack, J.M. and M. Hamaty, *Difficult-to-treat hypertensive populations: focus on African-Americans and people with type 2 diabetes.* J Hypertens Suppl, 1999. **17**(1): p. S19-24.

121. Machipisa, T., et al., *Association of Novel Locus With Rheumatic Heart Disease in Black African Individuals: Findings From the RHDGen Study.* JAMA Cardiology, 2021.

122. Bolnick, D.A., et al., *Genetics. The science and business of genetic ancestry testing.* Science, 2007. **318**(5849): p. 399-400.

123. Lee, S.S., et al., *Genetics. The illusive gold standard in genetic ancestry testing.* Science, 2009. **325**(5936): p. 38-9.

124. Nelson, A., *Bio science: genetic genealogy testing and the pursuit of African ancestry.* Soc Stud Sci, 2008. **38**(5): p. 759-83.

125. Royal, C.D., et al., *Inferring genetic ancestry: opportunities, challenges, and implications.* Am J Hum Genet, 2010. **86**(5): p. 661-73.

126. Akey, J.M., et al., *Population history and natural selection shape patterns of genetic variation in 132 genes.* PLoS Biol, 2004. **2**(10): p. e286.

127. Sabeti, P.C., et al., *Positive natural selection in the human lineage.* Science, 2006. **312**(5780): p. 1614-20.

128. Nielsen, R., et al., *Darwinian and demographic forces affecting human protein coding genes.* Genome Res, 2009. **19**(5): p. 838-49.

129. Li, J.Z., et al., *Worldwide human relationships inferred from genome-wide patterns of variation.* Science, 2008. **319**(5866): p. 1100-4.

130. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.

131. Novembre, J. and M. Stephens, *Interpreting principal component analyses of spatial population genetic variation.* Nat Genet, 2008. **40**(5): p. 646-9.

132. McVean, G., *A genealogical interpretation of principal components analysis.* PLoS Genet, 2009. **5**(10): p. e1000686.

133. Salas, A., et al., *Charting the ancestry of African Americans.* Am J Hum Genet, 2005. **77**(4): p. 676-80.

134. Tishkoff, S.A., et al., *The genetic structure and history of Africans and African Americans.* Science, 2009. **324**(5930): p. 1035-44.

135. Li, H.J., et al., *Novel Risk Loci Associated With Genetic Risk for Bipolar Disorder Among Han Chinese Individuals: A Genome-Wide Association Study and Meta-analysis.* JAMA Psychiatry, 2021. **78**(3): p. 320-330.

136. Chen, J., et al., *The trans-ancestral genomic architecture of glycemic traits.* Nat Genet, 2021. **53**(6): p. 840-860.

137. Conti, D.V., et al., *Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction.* Nat Genet, 2021. **53**(1): p. 65-75.

138. Mbarek, H., C.V. Dolan, and D.I. Boomsma, *Two SNPs Associated With Spontaneous Dizygotic Twinning: Effect Sizes and How We Communicate Them.* Twin Res Hum Genet, 2016. **19**(5): p. 418-21.

139. Demange, P.A., et al., *Parental influences on offspring education: indirect genetic effects of non-cognitive skills.* bioRxiv, 2020: p. 2020.09.15.296236.

8

GENERAL SUMMARY

The work presented in this thesis broadly covers the genetics of twinning, genetic influences on birth weight in twins compared to singletons, and considerations of analytical strategies for genetically informed study designs. I first present an overview of the current understanding of the biology and genetics of the human twinning process, highlighting critical differences between monozygotic (MZ) and dizygotic (DZ) twins. Next, I provide a molecular genetic study of DZ twinning in a large pedigree to identify novel genetic associations. Then, I report on the use of data from twin-family cohorts to genetically compare worldwide populations. The extent to which populations are genetically homogenous or comparable is an essential consideration of genome-wide association studies (GWAS) and meta-analysis. In what follows, I detail the findings of a meta-analysis of GWAS of twin birth weight from eight global twin cohorts. In this study, I compared the genetic influences on birth weight in twins to those previously reported in singletons. In the final study, I evaluate genetic ancestry estimates in twins and their family members with routinely employed methods for inferring population structure in GWAS.

In what follows, I provide concise summaries for each main chapter of the thesis.

Chapter 2 summarizes the current understanding of the biology, epidemiology, genetics, and incidence of twins. In this chapter, the central theme was differentiating between the twin types, MZ and DZ, by characteristically defining each group. The key distinguishing feature between the types of twins is the mechanism of embryo formation. Whereas MZ twinning results from splitting a single fertilized ovum, DZ twinning arises when two independent ova are fertilized. In the remainder of the chapter, other hallmark characteristics of twins are described, including the variation of incidence rates and other influential factors known to affect the twinning process, particularly for DZ twins. Several factors influence the DZ twinning process, including maternal traits (age, height, weight, parity, family history of twinning) and the use of assisted reproductive technologies. Enhanced knowledge of the biological and genetic aspects underlying DZ twinning has been tremendously improved with advancements in molecular techniques. These developments enabled the recent identification of genetic factors robustly associated with DZ twinning, namely variation in the *FSHB* and *SMAD3* genes. Despite the strong associations, our knowledge of the etiology of DZ twinning is not yet comprehensive, and even less is known about factors giving rise to MZ twins. Regardless, it is understood that there are many genetic and non-genetic factors contributing to the human twinning process. Efforts to characterize these influences are ongoing and will have important contributions in healthcare for improving fertility and predicting, managing, and improving the outcomes of twin pregnancies.

Chapter 3 follows up on the genetic findings related to DZ twinning presented in the previous chapter. This study aimed to identify rare and potentially novel genetic variants influencing the likelihood of a mother to conceive DZ twins. I leveraged genome-wide genotypic and sequence data from selected members of a sizable DZ-twinning pedigree to identify genomic regions shared by mothers of DZ twins. In a first step, I used nonparametric linkage analysis to locate trait-specific genes by demonstrating co-segregation of genetic markers and being a mother of DZ twins. Albeit modest, a region on chromosome 5 exhibited excess allele sharing amongst selected mothers (maximum LOD score 1.21, p=0.009), favoring the presence of linkage in this region. To pursue the linkage results further, I performed whole-genome sequencing on four of the pedigree's most distantly affected mothers of DZ twins. With available genotype and sequence data, I used haplotype estimation strategies to identify large shared genomic segments possessed by affected mothers with hopes of discovering rare/novel variants harbored in these regions. The largest areas shared by the four mothers with whole-genome sequence data were found to be on chromosomes 11, 1, 3, and 6. Within these regions, I propose to investigate further the variants identified through analysis of the sequence data. Functional consequence prediction of these variants may help uncover novel biological functions related to DZ twinning. It remains to be determined whether the shared regions and the genetic variants therein are specific to the mothers of the pedigree or possessed by all mothers of DZ twins more generally.

Chapter 4 focuses on designing and applying a DNA microarray, the Illumina Global Screening Array (GSA). The GSA was designed to provide reliable genotype calls in a high-throughput manner capable of fostering genetic studies of human disease, complex traits, and population genetics. Per individual, the GSA interrogates nearly 700,000 genetic loci, known as single nucleotide polymorphisms (SNPs). The array content contains a core backbone of genetic markers useful for imputation and thousands of other markers concerning human disease, drug metabolism, fertility, and twinning. Chapter 4 provides a detailed description of the GSA content selection and subsequent bioinformatic validation steps. I used genetic data from the GSA to compare the genetic compositions of populations possessing twin registers that routinely contribute to large-scale genetic association studies. Samples were obtained from twin-family participants representing Australian, Dutch, and Midwestern American populations. Principal Component Analysis (PCA), where the top components reflect genetic ancestry, enabled the visualization of genetic similarity. Visual inspection of the PCA results revealed superimposition of population clusters, suggesting genetic similarity. The addition of genetic data from globally diverse populations, including Nigerian samples genotyped on

GSA, further augmented this finding by providing broad resolution comparisons. Measures of population differentiation were quantified with F_{ST} (ranging from 0-1), where high values reflect a considerable degree of differentiation among populations. The analysis revealed very small F_{ST} between the three populations of interest, which were much less than the F_{ST} between each population and the Nigerian population. Together, the results suggest that Australian, Dutch, and Midwestern American populations have slight genetic differences but are genetically alike overall. The findings indicate that genetic data obtained from these populations can be combined for large-scale human genetic studies if proper analytical steps are employed. Overall, the GSA demonstrates excellent utility in providing robust genotype calls across many populations, pivotal to nearly all projects in the thesis.

Chapter 5 explores the genetic architecture of birth weight (BW) in twins to determine if twins can contribute to genome-wide association studies (GWAS) of BW, typically performed only with singletons. Here, the underlying question was whether the genetic factors responsible for differences in BW in singletons also explain variation in BW in twins. In this study, I addressed this question by first meta-analyzing the GWAS results performed in eight global twin cohorts, comprising 42,212 twin individuals. The meta-analysis effect size estimates were strongly correlated (Pearson's r=0.66) with the effect sizes of 150 genetic variants recently found to be significantly associated with BW in singletons. This finding provided the first indication that the genetic profile of BW of twins and singletons may be similar despite known differences in BW between the groups. I followed up this finding by computing genome-wide genetic correlations to discern the degree of genetic overlap between traits. Robust positive genetic correlations (r_g>0.9) were observed between the meta-analysis results of twin BW and previously published GWAS results of singleton BW. I also computed genetic correlations between BW in twins and a series of other health-related traits. The results were remarkably similar to previously reported correlations between the same traits and BW in singletons. In the last step, we utilized summary statistics from a discovery GWAS of BW in the UK Biobank to construct BW polygenic scores (PGS) in a Dutch target population. The target population consisted of singletons and twins participating in the Netherlands Twin Register. Using the same fraction of genetic markers captured by the optimal PGS, we found similar predictability of BW in twins and singletons. Taken together, the results of chapter 5 provide compelling evidence that the genetic profile underlying variation in BW is very similar between twins and singletons. The genetics of BW is critical to understand since BW is an important indicator of newborn health and survival. At the extremes, there are strong associations with adverse health outcomes later in life. Thus, ongoing efforts to identify genetic variants associated with BW will benefit from including twins through improved statistical power afforded by increased sample sizes.

Chapter 6 evaluates genetic ancestry inference in twins and family members. Estimating genetic ancestry is essential for population-based association studies to account for population heterogeneity and (sub)structure. I looked at two scenarios that may impact the results of genetic ancestry estimates. The first arises when family members participate in a study, and ancestry estimates differ for siblings. We hypothesized that this situation might mainly occur in ancestrally diverse or admixed families. The second scenario is when study participants have been genotyped across multiple different microarrays. To address these queries, I analyzed genome-wide SNP data of families participating in the Netherlands Twin Register. Family members included independently genotyped MZ and DZ twins, siblings, and parents, constituting 21,117 unique individuals belonging to 6,361 unique families. Participants were genotyped on one of three genotyping platforms: Affymetrix 6.0, Affymetrix Axiom, and Illumina GSA. A modest number of individuals were genotyped on at least two arrays (N=751), 35 of which genotyped on all three arrays, facilitating cross-platform comparisons. Estimates of genetic ancestry were determined from model-based (ADMIXTURE software) and algorithmic (PCA) approaches. Ancestry estimates were evaluated by comparing Euclidean distances, representing quantitative measures that summarize each method's ancestry estimates. Euclidean distances between pairs of family members closely resembled the degree of genetic relatedness between them, even in more diverse families. That is, the more closely two individuals are related, the smaller the Euclidean distances between them. However, the magnitude of the ancestry differences was larger when calculated with ADMIXTURE software. The differences were also larger for Affymetrix arrays (Affymetrix 6.0 and Axiom) than for the Illumina GSA. Across all platforms, ancestry estimates of individuals genotyped on multiple microarrays were similar. Slightly larger differences were found between Affymetrix and Illumina arrays than those with genotypes from Affymetrix arrays only. The differences can, in part, be attributed to the platform-specific SNPs used as input for PCA or ADMIXTURE software. Overall, the results of this study are promising and suggest that reliable estimates of ancestry can be obtained with PCA or ADMIXTURE software and that estimates in families are robust, even in diverse families. The findings from this study may have implications in the rapidly progressing area of trans-ancestry association studies.

Chapter 7 provides a thorough summary of each proceeding chapter and a broad discussion of the topics therein. Following the chapter-specific outlines, I overview current strategies and findings related to gene-finding for the human twinning process and how these developments are guiding future studies. A discussion of recent results related to BW genetics is then provided, emphasizing epigenetic processes that may help explain the link between BW extremes and later-life disease. Here, the underlying theme is expanding research to include combined 'omics' approaches, which will help elucidate additional biological mechanisms involved in human health and disease. Then, I put forth ideas related to the application, challenges, and implications of genetic ancestry testing from commercial and research perspectives. Lastly, I provide my take on future research opportunities related to twinning, birth weight, and genetic ancestry. I specifically highlight the need to include cohorts of diverse ancestry and under-studied populations to provide novel insights into different diseases for specific populations and in general.

APPENDIX

LIST OF PUBLICATIONS

1. Finnicum CT, Dolan CV, Willemsen G, Weber ZM, Petersen JL, **Beck JJ**, Codd V, Boomsma DI, Davies GE, Ehli EA. Relative Telomere Repeat Mass in Buccal and Leukocyte-Derived DNA. *PLoS One*. 2017 Jan 26;12(1):e0170765. doi: 10.1371/journal.pone.0170765. PMID: 28125671; PMCID: PMC5268389.

2. Finnicum CT, Doornweerd S, Dolan CV, Luningham JM, **Beck JJ**, Willemsen G, Ehli EA, Boomsma DI, Ijzerman RG, Davies GE, de Geus EJC. Metataxonomic Analysis of Individuals at BMI Extremes and Monozygotic Twins Discordant for BMI. *Twin Res. Hum. Genet*. 2018 Jun;21(3):203-213. doi: 10.1017/thg.2018.26. PMID: 29792248.

3. Mbarek H, van de Weijer MP, van der Zee MD, Ip HF, **Beck JJ**, Abdellaoui A, Ehli EA, Davies GE, Baselmans BML, Nivard MG, Bartels M, de Geus EJ, Boomsma DI. Biological insights into multiple birth: genetic findings from UK Biobank. *Eur. J. Hum. Genet*. 2019 Jun;27(6):970-979. doi: 10.1038/s41431-019-0355-z. Epub 2019 Feb 13. PMID: 30760885; PMCID: PMC6777609.

4. **Beck JJ**, Hottenga JJ, Mbarek H, Finnicum CT, Ehli EA, Hur YM, Martin NG, de Geus EJC, Boomsma DI, Davies GE. Genetic Similarity Assessment of Twin-Family Populations by Custom-Designed Genotyping Array. *Twin Res. Hum. Genet*. 2019 Aug;22(4):210-219. doi: 10.1017/thg.2019.41. Epub 2019 Aug 5. PMID: 31379313.

5. Kittelsrud JM, Ehli EA, Petersen V, Jung T, **Beck JJ**, Kallsen N, Huizenga P, Holm B, Davies GE. Avera Twin Register Growing Through Online Consenting and Survey Collection. *Twin Res. Hum. Genet*. 2019 Dec;22(6):686-690. doi: 10.1017/thg.2019.73. Epub 2019 Oct 14. PMID: 31608846.

6. Finnicum CT, **Beck JJ**, Dolan CV, Davis C, Willemsen G, Ehli EA, Boomsma DI, Davies GE, de Geus EJC. Cohabitation is associated with a greater resemblance in gut microbiota which can impact cardiometabolic and inflammatory risk. *BMC Microbiol*. 2019 Oct 22;19(1):230. doi: 10.1186/s12866-019-1602-8. PMID: 31640566; PMCID: PMC6805388.

7. Hur YM, Jeong HU, Kang MC, Ajose F, Kim JW, **Beck JJ**, Hottenga JJ, Mbarek H, Finnicum CT, Ehli EA, Martin NG, de Geus EJ, Boomsma DI, Davies GE, Bates T. The Nigerian Twin and Sibling Registry: An Update. *Twin Res. Hum. Genet*. 2019 Dec;22(6):637-640. doi: 10.1017/thg.2019.110. Epub 2019 Dec 3. PMID: 31796140.

8. **Beck JJ**, Bruins S, Mbarek H, Davies GE, Boomsma DI. Biology and Genetics of Dizygotic and Monozygotic Twinning. In: Khalil, A., Lewi, L., Lopriore, E. (*Eds*) *Twin and higher-order pregnancies*, Springer Nature, 2021. doi.org 10.1007/978-3-030-47652-6

9. **Beck JJ**, Pool R, van de Weijer M, Chen X, Krapohl E, Gordon SD, Nygaard M, Debrabant B, Palviainen T, van der Zee MD, Baselmans B, Finnicum CT, Yi L, Lundström S, van Beijsterveldt T, Christiansen L, Heikkilä K, Kittelsrud J, Loukola A, Ollikainen M, Christensen K, Martin NG, Plomin R, Nivard M, Bartels M, Dolan C, Willemsen G, de Geus E, Almqvist C, Magnusson PKE, Mbarek H, Ehli EA, Boomsma DI, Hottenga JJ. Genetic Meta-Analysis of Twin Birth Weight Shows High Genetic Correlation with Singleton Birth Weight. *Hum. Mol. Genet*. 2021 May. 6:ddab121. doi: 10.1093/hmg/ddab121. Epub ahead of print. PMID: 33955455.

10. Slunecka JL, van der Zee MD, **Beck JJ**, Johnson BN, Finnicum CT, Pool R, Hottenga JJ, de Geus EJC, Ehli EA. Implementation and implications for polygenic risk scores in healthcare. *Hum. Genomics*. 2021 June. Accepted.

11. Odintsova VV, Odintsova VV, Rebattu V, Hagenbeek FA, Pool R, **Beck JJ**, Ehli EA, van Beijsterveldt CEM, Ligthart L, Willemsen G, de Geus EJC, Hottenga JJ, Boomsma DI, van Dongen J. Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles. *Front. Psychiatry*. 2021 June. Accepted.

12. Odintsova VV, Sundermann M, Hagenbeek FA, Caramaschi D, Hottenga JJ, Pool R, Dolan C, Ligthart L, van Beijsterveldt CEM, Willemsen G, de Geus EJC, **Beck JJ**, Ehli EA, Cuellar-Partida G, Evans D, Medland S, Relton C, Boomsma DI, van Dongen J. DNA methylation signatures of left-handedness. *Nat. Hum. Behav*. 2021 July. Submitted.

GOOGLE SCHOLAR PROFILE