Twin Research and Human Genetics
Volume 15 ▌ Number 6 ▌ pp. 767–774 ▌ © The Authors 2012 ▌ doi:10.1017/thg.2012.62

The Minnesota Center for Twin and Family Research Genome-Wide Association Study

Michael B. Miller,[1] Saonli Basu,[2] Julie Cunningham,[3] Eleazar Eskin,[4] Steven M. Malone,[1] William S. Oetting,[5] Nicholas Schork,[6] Jae Hoon Sul,[4] William G. Iacono,[1] and Matt McGue[1,7]

[1]Minnesota Center for Twin and Family Research, Department of Psychology, University of Minnesota, Minneapolis, MN, USA
[2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA
[3]Department of Laboratory Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA
[4]Departments of Computer Science and Human Genetics, University of California, Los Angeles, CA, USA
[5]College of Pharmacy, Department of Experimental and Clinical Pharmacology, and the Institute of Human Genetics, University of Minnesota, Minneapolis, MN, USA
[6]Department of Molecular and Experimental Medicine, The Scripps Research Institute and The Scripps Translational Science Institute, La Jolla, CA, USA
[7]Institute of Public Health, University of Southern Denmark, Odense, Denmark

As part of the Genes, Environment and Development Initiative, the Minnesota Center for Twin and Family Research (MCTFR) undertook a genome-wide association study, which we describe here. A total of 8,405 research participants, clustered in four-member families, have been successfully genotyped on 527,829 single nucleotide polymorphism (SNP) markers using Illumina's Human660W-Quad array. Quality control screening of samples and markers as well as SNP imputation procedures are described. We also describe methods for ancestry control and how the familial clustering of the MCTFR sample can be accounted for in the analysis using a Rapid Feasible Generalized Least Squares algorithm. The rich longitudinal MCTFR assessments provide numerous opportunities for collaboration.

■ **Keywords:** GWAS, rapid feasible generalized least squares, Minnesota Center for Twin and Family Research

Increases in the efficiency of high-throughput genotyping have made it feasible for human geneticists to survey large numbers of genetic markers at a relatively low cost. A genome-wide association study (GWAS) is based on genotyping each individual in a sample on 100,000 or more common single nucleotide polymorphisms (SNPs) and then investigating the association of these SNPs with phenotypes of interest (McCarthy et al., 2008). The initial round of GWAS has been successful in identifying hundreds of genetic variants associated with a wide range of complex phenotypes (Hindorff et al., 2009; Visscher et al., 2012). The SNP variants that have been discovered have, however, characteristically had only small phenotypic effects. Consequently, very large samples are needed to ensure adequate statistical power, especially given the multiple testing burden (Manolio et al., 2009). As a consequence, consortia involving pooled GWAS samples, in some cases with sample sizes (N) exceeding 100,000, have been organized for several complex human phenotypes, including lipids (Kathiresan et al., 2009), height (Allen et al., 2010), body mass index (Speliotes et al., 2010), and age at menarche (Elks et al., 2010).

GWAS on behavioral phenotypes suggests that effect sizes are likely to be similarly small for intellectual ability (Davies et al., 2011), personality (De Moor et al., 2010), and psychopathology (Sullivan et al., 2012). Pooled GWAS samples are clearly also needed in the behavioral domain to achieve adequate statistical power to identify the specific genetic variants implied to exist by twin, adoption, and family studies. Consortia are already organizing around several key behavioral traits, including educational attainment, intellectual achievement, personality, and a range of mental health and substance use disorders (Cichon et al., 2009). The purpose of this paper was to describe a relatively large ($N > 8{,}000$) sample that has been deeply phenotyped longitudinally at a single center and which has recently been genotyped on a GWAS array. The genotyping and initial

RECEIVED 11 August 2012; ACCEPTED 16 August 2012.

ADDRESS FOR CORRESPONDENCE: Michael B. Miller, Department of Psychology/Elliott Hall, University of Minnesota, Minneapolis, MN 55455. E-mail: mbmiller@umn.edu

TABLE 1

Descriptive Characteristics of the Minnesota Center for Twin and Family Research Longitudinal Samples

Assessment	MTFS younger cohort	MTFS older cohort	ES	SIBS
Offspring				
Intake				
Mean (*SD*) age	11.7 (0.4)	17.5 (0.5)	11.9 (0.4)	14.9 (1.9)
Min, max age	10.7, 12.8	16.5, 18.5	10.9, 13.0	10.7, 20.9
No. of participants	1,519	1,252	998	1,232
Follow-up 1				
Mean (*SD*) age	14.8 (0.5)	20.7 (0.6)	15.1 (0.6)	18.3 (2.1)
Min, max age	13.6, 16.9	19.4, 22.7	13.6, 17.0	13.7, 25.4
No. of participants	1,409	1,111	930	1,158
Follow-up 2				
Mean (*SD*) age	18.2 (0.7)	24.7 (1.0)	In progress	In progress
Min, max age	16.6, 20.3	22.6, 29.3		
No. of participants	1,325	1,167		
Follow-up 3				
Mean (*SD*) age	21.5 (0.8)	29.6 (0.6)	In progress	NA
Min, max age	19.6, 24.3	28.4, 32.4		
No. of participants	1,339	1,168		
Follow-up 4				
Mean (*SD*) age	25.3 (0.7)	NA	NA	NA
Min, max age	23.7, 27.9			
No. of participants	1,332			
Follow-up 5				
Mean (*SD*) age				
Min, max age	In progress	NA	NA	NA
No. of participants				
Parents				
Intake				
Mean (*SD*) age	40.5 (5.2)	45.2 (5.3)	42.3 (5.5)	47.4 (4.4)
Min, max age	22.8, 65.3	29.9, 66.1	26.4, 42.3	35.3, 64.3
No. of participants	1,505	1,221	936	1,164

Note: MTFS = Minnesota Twin Family Study; ES = Enrichment Study; SIBS = Sibling Interaction and Behavior Study. Assessments currently in progress are designated as such; follow-up assessments that have not yet been undertaken are designated as NA.

analysis of the data were supported through the US National Institute on Drug Abuse's Genes, Environment and Development Initiative (GEDI). GEDI's major goal is to characterize the nature of gene–environment interplay in the development of substance use disorders. The sample provides numerous opportunities for collaborative research.

The Minnesota Center for Twin and Family Research

Minnesota Center for Twin and Family Research Longitudinal Samples

The GWAS sample is drawn from participants in one of three longitudinal studies undertaken under the auspices of the Minnesota Center for Twin and Family Research (MCTFR): (1) the Minnesota Twin Family Study (MTFS; Iacono et al., 1999); (2) the Sibling Interaction and Behavior Study (SIBS; McGue et al., 2007); and (3) the Enrichment Study (ES; Keyes et al., 2009). These studies utilized similar assessment protocols and a common sampling unit, a four-member family consisting of a pair of siblings and their rearing parents. The offspring in all three samples were initially assessed in adolescence and followed into at least early adulthood. In total, 9,827 individuals (5,001 offspring and 4,826 parents) have completed an MCTFR intake assessment. Table 1 provides a descriptive overview of the MCTFR samples.

The MTFS sample consists of 1,197 monozygotic (MZ) and 684 like-sex dizygotic (DZ) twin pairs (including five additional members from triplet sets). All pairs were ascertained from Minnesota state birth records between 1971–1985 and 1988–1994. The sample includes two cohorts: one initially assessed at age 11 years (the younger cohort) and the other initially assessed at age 17 years (the older cohort). Intake and follow-up assessments of the twins were scheduled to coincide with major transitions in the lives of these adolescents and young adults. Target ages (rate of follow-up participation) are 11, 14 (92.9%), 17 (87.3%), 20 (88.6%), 24 (89.1%), and 29 (in progress) for the younger cohort and 17, 20 (88.7%), 24 (93.8%), and 29 (94.2%) for the older cohort. At intake, approximately 18% of the recruited MCTFR families refused our invitation to participate and, based on a brief survey with non-participants, we found that non-participating families differed minimally in parental education, parental occupational status, or parental mental health from participating families (Iacono et al., 1999). The sample is, thus, broadly representative of the population of the state of Minnesota.

The SIBS sample includes 409 adoptive and 208 non-adoptive families, each consisting of a pair of adolescent

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press

siblings and their rearing parents. Adoptive families were recruited from records of the three largest adoption agencies in Minnesota, and non-adoptive families were recruited from Minnesota birth records. At least one offspring in every adoptive family was not genetically related to other family members, but many adoptive families had a biological offspring of one or both of the rearing parents. All offspring in non-adoptive families are full biological siblings. Among those families that were eligible to participate, 63% of adoptive and 57% of non-adoptive families completed an intake assessment. There are minimal differences in socio-economic status and in offspring mental health between participating and non-participating but eligible families (McGue et al., 2007). Unlike the MTFS, it was not logistically possible to link SIBS assessments with specific targeted ages. Rather, on average, offspring in SIBS families were in mid-adolescence at intake, late adolescence at their first follow-up and early adulthood at their second follow-up. The first follow-up of the SIBS sample is complete (with a 94.2% participation rate) and the second follow-up is in progress.

The ES was designed to extend the MTFS by oversampling 11-year-old twins likely to develop substance use disorders by virtue of being high on an index of externalizing behavior. Like the MTFS twins, ES twins were located through Minnesota state birth records (1988–1994). However, roughly half of the ES sample was selected after the mothers in those families completed a screening interview that established that at least one member of the twin pair was high on the externalizing dimension. The final ES sample included 300 MZ and 199 like-sex DZ twin pairs and their parents. The scheduling of intake and follow-up assessments of the ES twins parallels that for the younger MTFS cohort, with assessments targeted at ages 11, 14 (93.2% participation), 17 (in progress), and 20 (in progress). Among families eligible for participation in ES, 75.4% completed an intake assessment and participation rates did not vary by screening status. Differences in socio-economic and demographic factors between participating and eligible but non-participating families were generally minimal (Keyes et al., 2009). In addition, 48 pairs of MZ twins aged 14–16 at initial assessment were recruited from the same population as the ES families and assessed 1 year apart. The twins in these families were part of a pilot study on adolescent brain development (AdBrain) and completed a clinical assessment similar to that used in ES along with both structural and functional MRIs. Twins in this study were also included in the GWAS sample.

MCTFR Assessments

Although not identical, the assessment protocols in the three MCTFR studies overlap extensively. For all three studies, the intake assessment involved a daylong visit to our labs at the University of Minnesota. Assessments common to all three studies include (1) interview assessment of common child and adult mental and substance use disorders based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-III-R and DSM-IV); (2) quantity and frequency assessment of licit (e.g., tobacco and alcohol) and illicit drug use; (3) self-report assessment of personality; (4) individually administered Wechsler IQ tests and reading scales from the Wide Range Achievement Test; (5) teacher ratings of behavioral problems, personality, and academic progress and grades; (6) anthropometric measures; and (7) self-report of environmental risk and protective factors, including peer group characteristics, family functioning, and socio-economic status. In addition, the MTFS and ES include a half-day psychophysiological assessment that includes electroencephalography, event-related potentials, and autonomic nervous system activity. The SIBS, which does not include a psychophysiological assessment, includes videotaped family and sibling interaction sessions.

DNA Samples

Although collection of DNA samples has been a routine component of the MCTFR assessments for many years, to meet GEDI requirements and supply a fresh DNA sample to the Rutgers University Cell and DNA Repository (RUCDR), it was necessary to obtain new DNA samples. Among the 9,515 MCTFR participants who were eligible to provide a DNA sample (e.g., were alive and had not withdrawn from the study), 7,278 (76.5%) provided a blood sample and an additional 567 (6.0%) provided a saliva sample. Most of the individuals who did not provide a DNA sample could either not be contacted within the time frame of the GEDI project or had privacy or general concerns over providing a DNA sample for a repository. The RUCDR followed standard DNA extraction and storage procedures in processing the MCTFR samples (Sahota et al., 2007).

Genotyping Method and Quality Control Filters

Genotyping

Genome-wide genotyping was carried out using the Illumina Human660W-Quad array (Illumina, Inc., San Diego, CA) according to the manufacturer's protocol, as described in Li et al. (2010b). This Infinium HD Beadchip required 200ng DNA per sample and contains 657,366 variants, including 95,876 intensity-only markers for calling copy number variants that are not considered here. For quality control (QC) purposes, each 96-well plate included DNA from two members of a single three-member Centre d'Etude du Polymorphisme Humain (CEPH) family, with the two members rotating across plates, as well as a duplicate of another randomly selected sample on that plate. In addition to standard QC filters, we used GenCall scores, metrics of genotype reliability generated by the BeadStudio software (Illumina Corporation, San Diego, California), to assess sample quality (Cunningham et al., 2008). Specifically,

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press

both the GenCall_10, representing the 10th percentile, and GenCall_50, 50th percentile, scores were used to assess sample quality following standard guidelines. In addition, each sample was genotyped on a custom 96-plex panel using Illumina VeraCode chemistry (Lin et al., 2009). This panel contains SNPs present on the Human660W-Quad and served as an additional check on quality control.

QC of Samples

Genotyping was attempted with a total of 7,681 samples, including 83 samples that were included as within-plate duplicates, 160 CEPH control samples, and 60 samples that were repeats of samples that had failed an initial genotyping attempt due to low call rate. Of the 60 repeated samples, 47 (78%) were successfully genotyped on the second pass, including 16 (76%) of 21 retested samples that had failed completely on initial pass with no genotype calls. The 7,438 non-control samples (i.e., including those that failed the first attempt at genotyping and were re-genotyped) were subjected to five separate QC screens. These screens, along with the number (%) of samples failing each are (1) non-calls in more than 5,000 markers ($N = 130$, 1.7%); (2) GenCall_50 score < 0.9009 ($N = 8$, 0.1%); (3) GenCall_10 score <0.75 ($N = 6$, $<0.1\%$); (4) extreme heterozygosity or homozygosity ($N = 1$, $<0.1\%$); (5) sample mix-ups or failure to confirm genetic relationship with other genotyped relatives ($N = 15$, 0.2%). In total, 160 (2.2%) of the non-control samples were eliminated, leaving a cleaned sample of 7,278.

QC of SNP Markers

There are a total of 561,490 SNP markers (including sex chromosome and mitochondrial markers) on the Illumina Human660W-Quad array, but a set of 1,508 SNPs could not be called in any batch. The remaining 559,982 markers were subjected to nine separate QC filters. These QC filters along with both the number and percentage of markers failing each are listed here: (1) identified by Illumina as a bad marker on the array ($N = 23$, $<0.1\%$); (2) more than one mismatch in duplicated samples ($N = 70$, $<0.1\%$); (3) call rate $< 99\%$ ($N = 3{,}924$, 0.7%); (4) minor allele frequency less than 1% ($N = 19{,}999$, 3.6%); (5) more than two Mendelian inconsistencies across families ($N = 9{,}117$, 1.6%); (6) significant deviation from Hardy–Weinberg genotype frequencies in the White sample at $p < 10^{-7}$ ($N = 1{,}200$, 0.2%); (7) autosomal or X marker associated with participant sex at $p < 10^{-7}$ ($N = 16$, $<0.1\%$); (8) significantly associated with batch at $p < 10^{-7}$ ($N = 17$, $<0.1\%$); and (9) markers with more than two heterozygous calls if they were on the X chromosome in male samples or from mitochondrial DNA in the total sample ($N = 160$, $<0.1\%$). A total of 32,153 markers, 5.7% of the markers attempted, failed one or more of these QC filters, leaving 527,829 markers that passed all QC filters.

TABLE 2

Number of Families and Individuals in Final GWAS Sample

	Total sample		White sample	
Family type	Families	Individuals	Families	Individuals
MZ twin	1,156	4,194	1,109	4,066
DZ twin	665	2,240	577	2,143
Adopt/Adopt	269	880	210	498
Bio/Bio	191	622	162	604
Adopt/Bio	109	368	93	294
Other	NA	101	NA	97
Total	2,390	8,405	2,151	7,702

Note: All families consist of two parents and two offspring, but some members of some families are missing genotype data. The terms 'Bio' and 'Adopt' are used with the SIBS families (see text) where offspring raised as siblings were either the biological offspring of the parents (Bio) or were adopted (Adopt). Adopt/Adopt means that the parents adopted both of their children, Bio/Bio means that both children were biological offspring of the parents, and Adopt/Bio means that one child was adopted and the other was the biological offspring of the parents.

Final GWAS Sample

Among the 7,278 genotyped samples that passed all QC filters were 1,127 samples from individuals with an MZ twin who had not been genotyped. The genotypes of the non-genotyped twins were set equal to the genotypes of their genotyped MZ co-twins, resulting in a final GWAS sample of 8,405 individuals. The genotyped sample includes 3,924 (47%) males and 4,481 (53%) females; 4,434 (53%) of the sample are in the offspring generation and 3,971 (47%) are in the parent generation. In total, 2,390 families are represented in the analysis. For purposes of analysis (see later), four-member families were divided among MZ twin families ($N = 1156$), DZ twin families ($N = 665$), SIBS families with two adopted offspring ($N = 269$), SIBS families with two biological offspring ($N = 191$), and SIBS families with mixed adopted and biological offspring ($N = 109$). In addition, there were 101 genotyped individuals, usually step-parents, who did not fit into any one of these five four-member family types. Table 2 provides an overview of the number of participants in each category. The family structure of the sample produces several unique features. In particular, it allows us to confirm the genetic relationships of a large proportion of the sample, providing an additional level of QC screen not existing in most GWAS samples. For 6,919 (82.3%) of the genotyped participants (not counting the 1,127 non-genotyped MZ co-twins), we were able to confirm their relationship with at least one genetic relative in the sample (e.g., DZ twin, parent–offspring). Even though there are 2,390 families represented in the GWAS sample, because of spouse pairs and adoptive families the maximum number of genetically unrelated individuals in the sample is 4,706. There are 1,631 genotyped spouse pairs and 1,404 families where both genetic parents and at least one offspring have been genotyped.

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press

Genetic Ancestry

In the MCTFR, ancestry was based initially on the ethnicity specified on a birth certificate, adoption records, or by self-report. Of the 8,405 GWAS samples, 7,599 (90.4%) self-reported as having primarily European ancestry (i.e., 'White'), 382 (4.5%) as Asian, 83 (1%) as African American (i.e., 'Black'), and 127 (1.5%) reported mixed ancestry. All other ethnicities had a self-reported frequency of <1% and there were approximately 1% with a missing self-report. Because the genetics of complex phenotypes can vary across different ancestral groups (Bamshad, 2005), the primary sample used in our GWAS analysis will be comprised of individuals of European ancestry. However, to improve on the accuracy of the self-report data and to deal with missing self-report data, we ran EIGENSTRAT, http://genepath.med.harvard.edu/~reich/Software.htm (Price et al., 2006) analysis, extracting the first 10 principal components (PCs), to aid in the identification of a cluster of individuals with European ancestry.

The principal component analysis implemented in EIGENSTRAT is sensitive to pairs of close relatives, so one member of every such pair was excluded in initial computations, but the genotypes of relatives were then projected onto the components constructed from the set of unrelated subjects (this was accomplished using the '-w poplist' option to smartpca.perl). We identified relative pairs, even a few between-family pairs, by running a PLINK '–genome' analysis to produce a matrix of 'Z1' identity-by-descent (IBD) statistics for all pairs of individuals. The Z1 statistic is the estimated probability that two individuals share exactly one allele from a common ancestor, averaged across the autosomal genome. We chose to classify pairs of individuals as unrelated if their Z1 was less than 0.10, which would be roughly midway between first-cousin once-removed and second-cousin except that the ethnic structure of our sample tends to amplify distant relationships in the PLINK analysis (inflating Z1 somewhat, but inflating Z2 and Pi-Hat much more), especially for members of minority groups. The matrix of Z1 statistics was used to select the largest possible subgroup of unrelated individuals. This is known as the maximum clique problem in computer science and mathematics. To find the guaranteed maximum is extremely computationally demanding (it is an NP complete problem), so we used a greedy algorithm to achieve a reasonable answer in only a few seconds. This solution provides a subset of subjects such that no pairs have Z1 greater than 0.10. It is conceivable that a larger subset could be produced with much more computational effort, but the gain is likely to be minimal.

To identify a core sample of individuals with European ancestry for genetic analysis, we computed the Mahalanobis D^2 for each genotyped individual from the centroid of self-reported Whites using all 10 PCs. Figure 1 displays the distribution of the resulting D^2 values. The cluster of data points centered at a D^2 value of approximately 3,400 represents individuals who self-reported as Asian. We defined the European cluster as including all those individuals with $D^2 < 84$. We also confirmed that full siblings and DZ twins were never in separate clusters. After adding 1,087 MZ cotwins, the final sample of 7,702 individuals with European ancestry (91.6% of the total sample) included 101 who had missing self-reported ethnicity data and 46 whose ethnicity had been recorded as other than White (several of those cases were found to have been caused by data entry errors). Figure 2 provides a plot of the first two PCs from the EIGENSTRAT analysis. The first PC is a dimension anchored at one end by individuals of self-reported East Asian ancestry and at the other end by individuals of self-reported European ancestry. The second PC is a dimension that differentiates individuals of self-reported African ancestry from those of European ancestry. Additional PCs tended to distinguish other groups such as the Native Americans or those from India.

Imputation

Untyped SNP genotypes were imputed using HapMap2 as the reference panel. Samples were first phased using Beagle (Browning & Browning, 2009), which takes into account the family structure of the data. Genotypes were then imputed from the phased data using Minimac, which is a computationally efficient version of the MACH program (Li et al., 2010a). HapMap2 provided 2,543,887 autosomal SNPs in the r22 reference panel and 64,621 X-chromosome SNPs in the r21 reference panel. Of those SNPs, 501,912 autosomal markers and 11,685 X-chromosome markers had been genotyped for our samples on the Human660W-Quad platform, thus leaving 2,041,975 autosomal and 52,936 X SNPs to be imputed. Of those 2,094,911 imputed SNPs, 96.4% had $r^2 > 0.5$, 90.9% had $r^2 > 0.8$, 85.2% had $r^2 > 0.9$, and 77.2% had $r^2 > 0.95$ in the European American sample.

Analytical Strategy

Most GWAS involve independently sampled individuals or families of fixed structure (e.g., parent–offspring trios), which can be analyzed using freely available software, such as PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/ (Purcell et al., 2007). The five different types of four-member families that comprise the MCTFR sample, however, present analytical challenges that cannot be handled efficiently within PLINK or other existing software for GWAS analysis. Consequently, we developed an efficient analytical approach for the MCTFR GWAS data based on a Rapid Feasible Generalized Least Squares (RFGLS) algorithm (Li et al., 2011). Briefly, the analysis involves regressing a phenotype on a set of pre-specified covariates and a single SNP genotype (coded 0-1-2), repeated for each SNP in the GWAS. Because of the family structure, the residuals

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press

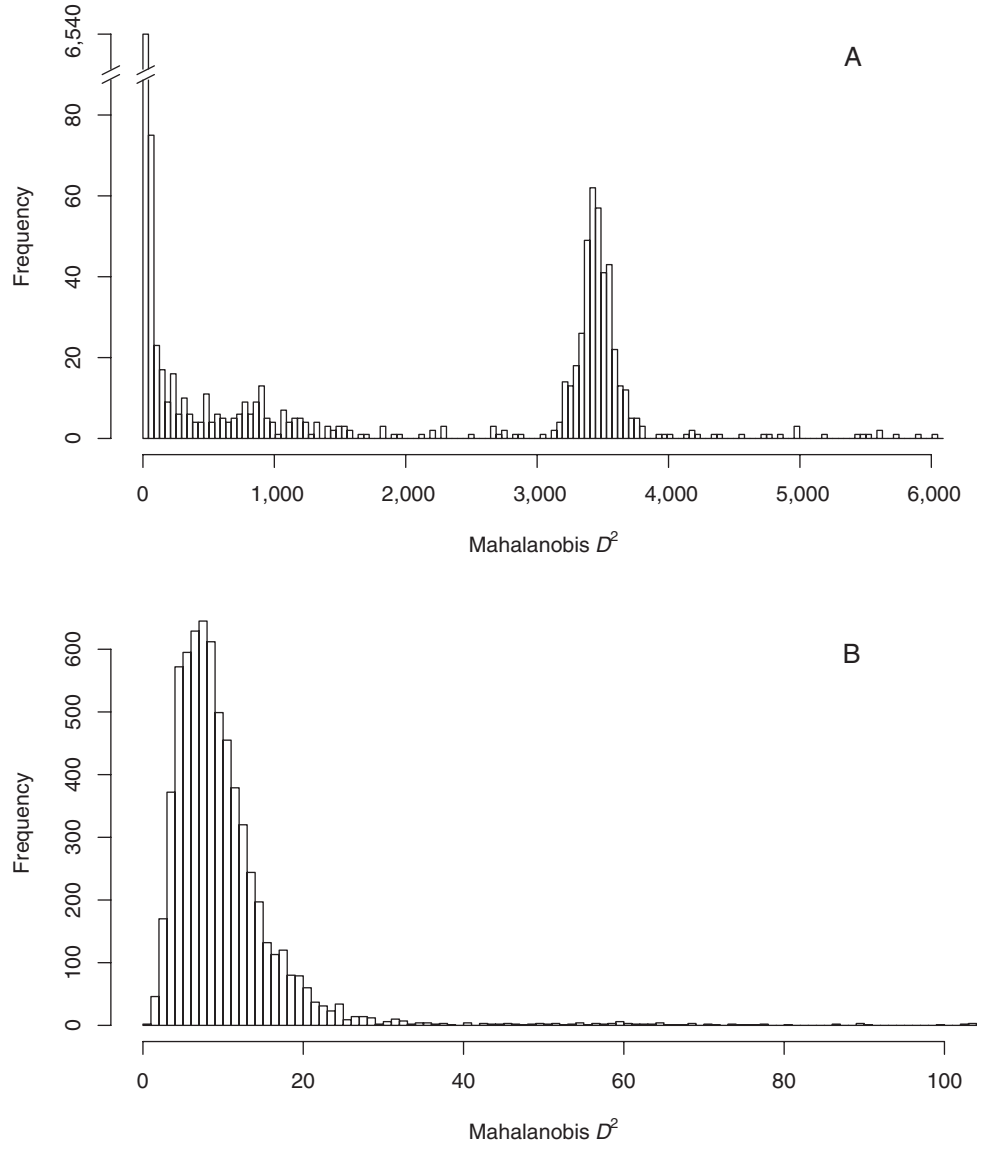

FIGURE 1

Two histograms of Mahalanobis distances (D^2) to the centroid of the European American ('White') group. Both histograms display the same data, but with different axis ranges and bin widths. Histogram A uses bins of width 42 so that the first two bars on the left include all 6,615 White subjects and no others. The subjects with D^2 near 1000 are mostly mixed-race, Hispanic, and Native American. The peak between 3,000 and 4,000 was caused by 378 Asian subjects, mostly Korean adoptees, and the subjects with D^2 exceeding 4,000 are Black. Histogram B used bin widths of 1 so that all visible bars (values less than 84 on the abscissa) represent counts of White subjects. All subjects with D^2 less than 84 were used in the core European American group for all initial GWAS analyses.

from the regression are non-independently distributed with a variance–covariance structure that depends on family type (e.g., it is not the same for MZ and DZ twin families). In RFGLS, rather than estimate the residual variance–covariance matrix for each family type in each SNP regression, we estimate it only once for each family type based on the model with covariates but no SNP effect. The residual variance–covariance matrix in the analysis of individual SNP effects is then fixed at the estimates from the model without any SNP effects. As we show in Li et al. (2011), because any SNP effect is likely to be very small, fixing the variance–covariance matrix in this way has no effect on the test statistics but greatly increases the speed of the analysis.

Future Plans and Summary

The MCTFR GWAS sample is large, >8,400, deeply phenotyped and longitudinally informative. We recently demonstrated the developmental utility of the MCTFR GWAS sample using height as a proof of principle (Vrieze et al., 2011). Specifically, we used 176 SNP height variants identified in the GIANT consortium (Allen et al., 2010) to

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press

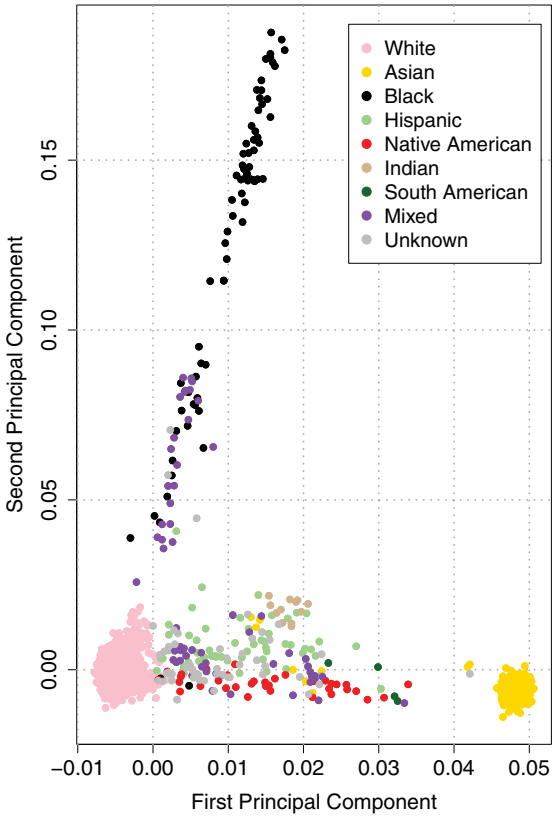

FIGURE 2

(Colour online) Scatter plot of the first two principal components from an EIGENSTRAT analysis of 10,000 markers with color coding to show reported ethnicity. Some of the gray dots of 'Unknown' ancestry had been reported as White, but they were sufficiently far from the White centroid that they have been recoded as Unknown. Note that 'Mixed' was a vague category that seems to include individuals with some European ancestry mixed with either African, Asian, or Native American ancestry. Several subjects have been recoded as White because of Mahalanobis D^2 less than 84, indicating close proximity to the centroid of the White group in principal component space (see Figure 1). These included 2 of 73 reportedly Black subjects, 5 of 60 Hispanic, 17 of 77 Mixed, 9 of 44 Native American, and 59 of 66 of Unknown ancestry. An additional 41 of 6,564 subjects reported to be White were recoded as Unknown based on the same analysis. The graph reflects the final ethnic classifications that followed the Mahalanobis distance analysis described in the text and in Figure 1.

create a polygenic score that accounted for 9.2% of adult height variance in our sample (comparable to the 10.5% accounted for in the original study). When the polygenic score was investigated longitudinally, however, we were able to show that almost all of its effect was on pre-pubertal height; that is, it was not significantly predictive of pubertal growth spurt. As consortia publish findings from pooled GWAS for other phenotypes, the rich longitudinal assessments in the MCTFR will allow us to similarly characterize these effects developmentally. The MCTFR also includes a rich assessment of environmental risk and protective factors (Hicks et al., 2009), which will provide an opportunity to explore genotype–environment interaction effects as findings emerge from GWAS consortia.

The MCTFR sample will also soon be genotyped on Illumina's HumanExome BeadChip, which includes more than 240,000 coding sequence variants. These data, when combined with the original GWAS data, will provide extensive coverage of both common and relatively rare genetic variants in key genetic regions. We are also at the initial stages of planning for whole genome sequencing a large number of MCTFR participants, as well as obtaining additional relevant phenotype information. In this era of large-scale consortia, the extensive phenotypic, genetic, and environmental assessments in the MCTFR will provide numerous opportunities for collaboration.

Acknowledgment

This research was supported in part by USPHS Grants from the National Institute on Alcohol Abuse and Alcoholism (AA09367 and AA11886), the National Institute on Drug Abuse (DA05147, DA13240, and DA024417), and the National Institute on Mental Health (MH066140).

References

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., . . . Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838.

Bamshad, M. (2005). Genetic influences on health—Does race matter? *JAMA*, *294*(8), 937–946.

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, *84*(2), 210–223.

Cichon, S., Craddock, N., Daly, M., Faraone, S. V., Gejman, P. V., Kelsoe, J., . . . Psychiatric GWAS Consortium Coordinating Committee. (2009). Genomewide Association Studies: History, rationale, and prospects for psychiatric disorders. *American Journal of Psychiatry*, *166*(5), 540–556. doi:10.1176/appi.ajp.2008.08091354.

Cunningham, J. M., Sellers, T. A., Schildkraut, J. M., Fredericksen, Z. S., Vierkant, R. A., Kelemen, L. E., . . . Goode, E. L. (2008). Performance of amplified DNA in an Illumina GoldenGate BeadArray assay. *Cancer Epidemiology Biomarkers & Prevention*, *17*(7), 1781–1789. doi:10.1158/1055-9965.epi-07-2849.

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., . . . Deary, I. J. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular Psychiatry*, *16*(10), 996–1005.

De Moor, M. H. M., Costa, P. T., Terracciano, A., Krueger, R. F., de Geus, E. J., Toshiko, T., . . . Boomsma, D. I. (2010). Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry*, *17*(3), 337–349.

Elks, C. E., Perry, J. R. B., Sulem, P., Chasman, D. I., Franceschini, N., He, C. Y., . . . Murray, A. (2010). Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nature Genetics*, *42*(12), 1077–1085.

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press

Hicks, B. M., South, S. C., DiRago, A. C., Iacono, W. G., & McGue, M. (2009). Environmental adversity and increasing genetic risk for externalizing disorders. *Archives of General Psychiatry*, *66*(6), 640–648.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–9367.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., & McGue, M. (1999). Behavioral disinhibition and the development of substance use disorders: Findings from the Minnesota Twin Family Study. *Development and Psychopathology*, *11*, 869–900.

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., ... Cupples, L. A. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics*, *41*(1), 56–65.

Keyes, M. A., Malone, S. M., Elkins, I. J., Legrand, L. N., McGue, M., & Iacono, W. G. (2009). The enrichment study of the minnesota twin family study: Increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Research and Human Genetics*, *12*(5), 489–501.

Li, X., Basu, S., Miller, M. B., Iacono, W. G., & McGue, M. (2011). A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families. *Human Heredity*, *71*(1), 67–82.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010a). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834.

Li, Y. F., Sheu, C. C., Ye, Y. Q., de Andrade, M., Wang, L., Chang, S. C., ... Yang, P. (2010b). Genetic variants and risk of lung cancer in never smokers: A genome-wide association study. *Lancet Oncology*, *11*(4), 321–330. doi:10.1016/s1470-2045(10)70042-5.

Lin, C. H., Yeakley, J. M., McDaniel, T. K., & Shen, R. (2009). Medium- to high-throughput SNP genotyping using VeraCode microbeads. *Methods in Molecular Biology*, *496*, 129–142.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. doi:10.1038/nrg2344.

McGue, M., Keyes, M., Sharma, A., Elkins, I., Legrand, L., Johnson, W., & Iacono, W. G. (2007). The environments of adopted and non-adopted youth: Evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). *Behavior Genetics*, *37*(3), 449–462.

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.

Sahota, A., Brooks, A. I., & Tischfield, J. A. (2007). Protocol 6: Preparing DNA from blood: Large-scale extraction. In M. P. Weiner, S. B. Gabriel & J. C. Stephens (Eds.), *Genetic Variation: A Laboratory Manual* (pp. 124–128). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., ... Loos, R. J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948.

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Gemetic architectures of psychiatric disorders: The emerging picture and its implications. *Nature Reviews Genetics*, *13*(7), 537–551.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery [Review]. *American Journal of Human Genetics*, *90*(1), 7–24. doi:10.1016/j.ajhg.2011.11.029.

Vrieze, S. I., McGue, M., Miller, M. B., Legrand, L. N., Schork, N. J., & Iacono, W. G. (2011). An assessment of the individual and collective effects of variants on height using twins and a developmentally informative study design. *Plos Genetics*, *7*(12), e1002413.

https://doi.org/10.1017/thg.2012.62 Published online by Cambridge University Press