Published in final edited form as:
Behav Genet. 2015 January ; 45(1): 12–23. doi:10.1007/s10519-014-9692-4.

The National Longitudinal Study of Adolescent to Adult Health (Add Health) Sibling Pairs Genome-Wide Data

Matthew B. McQueen[a,*], **Jason D. Boardman**[b], **Benjamin W. Domingue**[b], **Andrew Smolen**[c], **Joyce Tabor**[d], **Ley Killeya-Jones**[d], **Carolyn T. Halpern**[d,e], **Eric A. Whitsel**[f,g], and **Kathleen MullanHarris**[d,h,i]

[a]Department of Integrative Physiology, University of Colorado Boulder

[b]Institute of Behavioral Science, University of Colorado Boulder

[c]Institute for Behavioral Genetics, University of Colorado Boulder

[d]Carolina Population Center, University of North Carolina at Chapel Hill

[e]Department of Maternal and Child Health, Gillings School of Global Public Health, University of North Carolina at Chapel Hill

[f]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill

[g]Department of Medicine, School of Medicine, University of North Carolina at Chapel Hill

[h]Department of Sociology, University of North Carolina at Chapel Hill

[i]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill

Abstract

Here we provide a detailed description of the genome-wide information available on the National Longitudinal Study of Adolescent to Adult Health (Add Health) sibling pair subsample (Harris et al., 2012). A total of 2020 samples were genotyped (including duplicates) arising from 1946 Add Health individuals from the sibling pairs subsample. After various steps for quality control (QC) and quality assurance (QA), we have high quality genome-wide data available on 1,888 individuals. In this report, we first highlight theQC and QA steps that were taken to prune the data of poorly performing samples and genetic markers. We further estimate the pairwise biological relationships using genome-wide data and compare those estimates to the assumed relationships in Add Health. Additionally, using genome-wide data from knownregional reference populations from Europe, West Africa, North and South America, Japan and China, weestimate the relative genetic ancestry of the respondents. Finally, rather than conducting a traditional cross-sectional genome-wide association study (GWAS) of body mass index (BMI), we opted to utilize the extensivepublicly available genome-wide information to conduct a weighted genome-wide association study (GWAS) of longitudinal BMI while accounting for both family and ethnic variation.

[*]Address all correspondence to: Matthew B. McQueen, University of Colorado Boulder, Department of Integrative Physiology, 354 UCB; ph: 303-735-5158; fx: 303-492-4009. matt.mcqueen@colorado.edu..

Keywords

National Longitudinal Study of Adolescent to Adult Health; Add Health; GWAS; BMI; Obesity; Sibling Pairs

Introduction

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a nationally representative longitudinal study including over 20,000 adolescents originally sampled in Grades 7-12 in the United States between 1994 and 1995. Add Health respondents have been followed through adolescence and into early adulthood with four in-home interviews (1995, 1996, 2001-2002 and 2008-2009). The Add Health design included the oversampling of approximately 3,000 pairs of individuals who were raised in the same household. These pairs of individuals are biologically related to varying degrees including monozygotic (MZ) and dizygotic (DZ) twins, full siblings, half siblings and unrelated. For further details on the study design and sampling scheme for the Add Health Sibling Pairs Sample, including phenotypic, environmental and biological assessments, see Harris et al. (2013). During the fourth in-home visit (Wave IV; 2008-2009), Add Health collected saliva on the entire sample of Add Health respondents (N=15,701), including the sibling pairs subsample. Consent rates (consent to provide saliva for DNA extraction) among the sibling pairs subsample for Wave IV saliva collection was an impressive 96%, which was similar to the consent rate for the entire Add Health sample. See Harris et al. (2013) for additional information on the Add Health Study design and genetic data.

Genome-wide association studies have largely been conducted using case-control and/or cross-sectional study designs primarily due to efficiency and ease of collection. The integration of genome-wide data into well-characterized longitudinal and prospective cohort studies that *include biological relationships* such as the Add Health sibling pair subsample has been much more limited. Notable exceptions include the Framingham Heart Study [NHLBI SNP-Health Association Resource (SHARe)] that follows multi-generational samples prospectively, and the Health and Retirement Study (HRS) that surveys a representative sample of individuals over the age of 50 every two years and follows them prospectively.Utilizing genetic data from longitudinal and prospective cohort studies has many potential advantages including refinement of phenotypic endpoints, phenotypic change and trajectory. Within the context of the ethnically diverse Add Health sibling pairs subsample of adolescents and young adults, there are additional advantages including family-based assessment and utilizing measured environmental and social factors collected over time.

Here, we provide a description of the genome-wide data that were generated on the Add Health sibling pairs subsample. In particular, we focus on describing the targeted sample for genotyping, the quality control (QC) and quality assurance (QA) steps that were taken and how putative biological relationships were assessed. Using genome-wide data from known reference populations, we alsoshow the genetic ancestry of the Add Health sibling pairs subsample. We also explore the genetic heritability of body mass index (BMI) using the

genome-wide data from the Add Health sibling pairs subsample. Finally, rather than conducting a traditional cross-sectional genome-wide association study (GWAS) of BMI, we opt to utilize the rich genome-wide information publicly available to conduct a weightedgenome-wide association study (GWAS) of longitudinal BMI while accounting for both family and ethnic variation. Funding for the genotyping of the sibling pairs subsample was provided by the National Institutes of Child Health and Human Development (R01 HD060726).

Materials andMethods

Quality Control and Quality Assurance

The QA/QC Report for the Add Health Sibling Pairs Sample is provided in the Supplemental Materials. Briefly, there we describe how the sample was selected, prepared and genotyped, the number of markers removed, the number of samples removed, sex checks and duplicate concordance. The number of individual samples deemed of high quality for subsequent relationship testing, ancestry estimation and genome-wide analysis is N=1,888. The number of SNP markers (chromosomes 1-22 and X) with a genotyping call rate of at least 95% is N=940,862.

Computer Software

For biological relationship testing, PLINK (Purcell et al., 2007) and Kinship-based Inference for GWAS (KING; Manichaikul et al., 2010) were used. For genetic ancestry estimation, we used KING (Manichaikul et al., 2010) and ADMIXTURE (Alexander et al., 2009). R (R Core Team, 2013) was used for graphical display of ancestry information. For the estimation of heritability using genome-wide data, genome-wide complex trait analysis (GCTA; Yang et al., 2011) was used. For the genome-wide association study (GWAS) we used SAS 9.3 (SAS Institute, Cary, NC, USA) and R. Once again, R was usedfor graphical display of the genome-wide association results.

Estimation of Genetic Relatedness

Using information from chromosomes 1-22 (919,509 SNP markers) on the clean set of 1,888individual samples, we estimated Identity by State (IBS) and Identity by Descent (IBD) using PLINK (Purcell et al., 2007) as well asthe Kinship Coefficient using KING (Manichaikul et al., 2010). These measures are used to test duplicate concordance, confirm expected biological relationships, identify unknown or cryptic relatedness in the sample and provide the information necessary to assess genetic ancestry. The relationship measures are calculated pairwise for all individuals in the dataset. As generally recommended, we pruned autosomal SNPs to establish an approximately independent set of SNP markers to be used for IBS, IBD and Kinship Coefficient estimation. We used a linkage disequilibrium threshold (r^2) of 0.20 with a SNP window size of 50 and number of SNPs to shift window at each step of 5 *(PLINK command: --indep-pairwise 50 5 0.20*). After pruning, a set of 231,649 autosomal SNP markers in approximate linkage equilibrium was used to estimate the relationship measures. Pairwise mean IBD was estimated using PLINK (“PI_HAT”). However, PLINK’s estimates of IBD may be biased in stratified (multiethnic) samples

(Manichaikul et al., 2010 and Thorton et al., 2012). Therefore, we relied upon the KING package to provide estimates of relationship (Kinship) that are robust to stratification.

Estimation of Genetic Ancestry

We explored genetic ancestry in two different ways. Note that the sample of N=1,888 individuals with clean genotypes includes two MZ twin pairs. For the purposes of estimating genetic ancestry, we removed one individual (randomly) from each of the two MZ twin pairs resulting in a final analysis sample of N=1,886. For our first approach to estimating genetic ancestry, we used KING (Manichaikul et al., 2010) to identify clusters of individuals based upon genetic similarity. KING uses multidimensional scaling (MDS) with Euclidean distance to generate principal coordinates (PCs) that can be used to identify population substructure. For the KING procedure, we used the same set of 231,649autosomal SNP markers in approximate linkage equilibrium that was used for the estimation of genetic relatedness.

Second, we explored genetic ancestry using the software package, ADMIXTURE (Alexander et al., 2009). ADMIXTURE uses an efficient likelihood model-based estimation of genetic ancestry using genome-wide data. For the ADMIXTURE procedure, we opted for a supervised analysis utilizing a series of known genetic ancestry populations as fixed groups to estimate the proportion of ancestry that individuals from the Add Health sibling pairs subsample share with each ancestral reference population. The ancestral populations used were derived from the Human Genome Diversity Project (HGDP; Li et al., 2008) and International Haplotype Map Project (HapMap; International HapMap 3 Consortium, 2010). Specifically, we utilized 108 samples from the HGDP to represent the Americas (Surui, Maya, Karitiana, Pima and Colombian), and 402 samples from HapMap to represent Europe (CEU), Africa (YRI), China (CHB) and Japan (JPT). In all, we identified 257,035 SNP markers that overlap across the Add Health sibling pairs subsample, the HGDP sample and the HapMap sample. For efficiency using the program ADMIXTURE, we created an autosomal SNP marker set that was in approximate linkage equilibrium (123,198 SNPs) to estimate ancestry.

GCTA Heritability of BMI

We used the GCTA software (Yang et al., 2011) to estimate heritability of body mass index (BMI) as measured in Add Health as part of the Wave 2, Wave III and Wave IV data collection. BMI was calculated using the standard formula of mass (kg) divided by height (m)squared (kg/m^2) for each respondent. GCTA works by first estimating the genetic relatedness between all possible pairs of individuals. The genetic relatedness measures are known to be sensitive to population stratification, so for this application, we restricted the analysis to white respondents only. The subsequent step in the GCTA process is the estimation of a random effects model, where the random effects have a covariance structure based on the estimated genetic relatedness values. The percentage of total variance associated with the genetic random effects is considered the estimated heritability. For this particular study, we removed all pairwise relationship measures above 0.025.

Genome-Wide Association Approach

To conduct SNP-by-SNP genome-wide association analysis of BMI, we started initially with919,509 autosomal markers with a genotyping call rate greater than 95%. Further steps involved removing SNP markers that show evidence of deviation from Hardy-Weinberg Equilibrium (HWE) in 492 unrelated, self-identified white individuals extracted from the entire sample. These 492 individuals were selected via a two-step process. First, we focused on the homogenous self-identified white sample followed by the random selection of one individual from each biological relationship pair. In all, 6,237 autosomal SNPs were flagged for potential deviation from HWE ($p < 0.001$) and removed from the genome-wide association analysis. On the basis of minor allele frequency (MAF), we further removed SNP markers with an MAF < 0.01 (32,313). Therefore, the final genome-wide association marker set includes 880,959 autosomal SNPs.As noted previously, the sample of N=1,888 individuals with clean genotypes includes two MZ twin pairs. For the purposes of the genome-wide association analysis, we removed one individual (randomly) from each of the two MZ twins pair resulting in a final analysis sample of N=1,886.

To optimize statistical power, rather than conduct a traditional family-based association analysis on the related sets of individuals, we opted for a more flexible linear mixed effects model (Bates et al., 2014). This approach allowed us to model longitudinal measures of BMI (Waves II, III and IV) from all 1,886 individuals while accounting for biological relationships (if present) and within-individual variation in BMI as well as controlling for age, sex and MDS-derived components of ancestry. BMI measures from women who were pregnant were excluded from this analysis. Note that only 5 respondents did not have BMI measures across all three waves.

Weighted Association

A major issue plaguing genome-wide studies is multiple testing that arises from testing hundreds of thousands (if not millions) of SNP markers for association with the disease or trait of interest. In response to this issue, many investigators have advocated the use of a Bonferroni-correction to limit the probability of committing type-I errors. However, this comes at a cost of simultaneously increasing the probability of committing type-II errors, thereby diminishing the opportunity of detecting true association signals. This is particularly true of smaller genome-wide association datasets such as the sibling pairs samples. One solution is to utilize prior information into the association scan. In this study, we use a weighted association approach as implemented by Roeder et al., 2006 to accomplish this. While there are a variety of ways to construct weights, there are only two criteria that must be met. First, each weight must be greater than 0 and the mean of the weights must be 1. There are numerous sources of prior information that can motivate the weighting scheme including linkage scans, bioinformatics information, as well as previously conducted (and independent) genome-wide association signals (Roeder et al., 2007; Roeder & Wasserman, 2009). Further, the prior information can be in the form of test statistics (i.e. LOD scores, Z scores) or p-values (Roeder et al., 2006).

The weights for this study were derived from the GWAS on BMI as conducted by the Genetic Investigation of Anthropometric Traits (GIANT) consortium (Speliotes et al., 2010).

Details on the sample and the analysis procedures can be found elsewhere (Speliotes et al., 2010). Briefly, the GIANT consortium conducting a GWAS on BMI using 249,796 individuals and made the association signals for each of the ~2.8M SNP markers available to the public. In particular, the p-values from the GWAS served as the prior information used to devise the weighting scheme for the genome-wide association scan from this study. In the original introduction to this approach, Roeder et al. (2006) introduced exponential and cumulative weighting procedures. We opted for a cumulated weighting scheme that can be less sensitive to large prior association signals and we also used a scaling factor (B) of 2 (Roeder et al., 2006).

We focused on markers that either overlapped between the GIANT consortium and this study or GIANT consortium markers that were in reasonable linkage disequilibrium (r^2> 0.80)with SNP markers from this study. In all, we identified 717,411 markers to be tested for association that also have corresponding weights from the GIANT consortium. As a result, p-values from this study may be up-weighted or down-weighted depending upon the association signal from the GIANT GWAS. More specifically, the unweighted (nominal) p-values from this study are divided by the weights as assigned through the GIANT GWAS to generate the weighted p-values. The weighting procedure was conducted in R using the "weighted_FDR.R" script that can be found at http://www.wpic.pitt.edu/wpiccompgen/fdr/.

Results

Biological Relationships

Table 1uses the Kinship Coefficient generated from KING to tabulate the observed pairwise relationship status based upon genome-wide data versus the expected pairwise relationships based upon information from Add Health (using the Add Health variable, "sibcl4"). The sibcl4 variable is one of the classification variables available for the Add Health sibling pairs subsample. This particular classification designates pairs of respondents into monozygotic twin pair (MZ), dizygotic twin pair (DZ), full sibling pair (FS), half-sibling pair (HS), cousin pair (CO), unrelated pair (UN) and undetermined relationship (UD). A total of 1,781,328 ($_{1888}C_2$) pairwise relationship comparisons were conducted. As recommended by the authors of KING (Manichaikul et al., 2010), a Kinship Coefficient greater than 0.354 is categorized as an MZ twin pair (duplicates have been removed), between 0.177 and 0.354 as 1st degree relationship (DZ twin pairsand full sibling pairs - no parent-offspring are present), between 0.0884 and 0.177 as 2nd degree relationship (half-sibling pair and avuncular), between 0.0442 and 0.0884 as 3rd degree relationship (half-avuncular and first cousin) and less than 0.0442 as not related (NR).Note that the expected values for the Kinship Coefficient are 0.50, 0.25, 0.125, 0.0625 and 0.0 for MZ, DZ/FS, HS, CO and UN respectively and the boundaries suggested by the authors of KING are to account for the variability in the estimated Kinship based upon genome-wide data. As can be seen in table 1, the vast majority of expected relationships are consistent with the observed genetic relationships. However, there are notable discrepancies. For example, there are 33 expected full sibling pairs that are consistent with being half-sibling pairs according to the genetic data. Further, there are two pairs of MZ twins - one set of MZ twin pairs were thought to be a DZ twin pair while the other set were unknown prior to this study.

Additionally, a pair of individuals who were thought to be an MZ twin pair is likely an DZ twin pair.As can also been seen in the table we have detected 266 pairs of individuals thought to be unrelated who are at least distally related.The majority of these pairs (264) are 3rd degree relationships (i.e., cousins). In total, there are 664 full sibling/DZ twin pairs (1328 individuals) that would be utilized for studies employing a traditional sibling pair family-based design. Based upon these relatedness measures, we have created a new variable available in the Add Health data sources called "sibcl**g**" that specifies the biological relationship based upon the genome-wide information as presented here.This variable will be made available to Add Health researchers through the Restricted-Use Data mechanism of Add Health and will be included with other variables related to the sibling pair data.

Self-Reported Ethnicity and Multidimensional Scaling (MDS)

We used the Add Health "ah_race" variable for self-report ethnicity. This variable includes five categories: White, Black, Native American, Asian and Hispanic. Add Health respondents who self-reported as Hispanic were included in the Hispanic category irrespective of whether they also self-reported as White, Black, etc. Of the 1,886 individuals included in this analysis, there are 917 who self-identify as White, 677 who self-identify as Black, 209 who self-identify as Hispanic, 73 who self-identify as Asian, 8 who self-identify as Native American; 2 individuals have unknown self-identified ethnicity (see table 2). Supplemental figure 3 shows the first 5 principal coordinate (PC) MDS estimates from KING, color-coded by self-identified ancestry. As can be seen in supplemental figure 3A, the first principal coordinate (PC1) distinguishes the European ancestry from African ancestry while the second principal coordinate (PC2) captures variation between European, Asian and to an extent, Hispanic ancestry. In supplemental figure 3B we see variation from Hispanic samples including a small set of self-identified Native American individuals. Supplemental figures 3C and 3D further distinguish between and within self-identified ethnic groups.

Self-Reported Ethnicity and Genetic Ancestry

Table 2 and supplemental figure 4 displays the proportion of ancestry shared with each of the reference populations of each individual from the Add Health sibling pairs sample. As can be seen in table 2 and supplemental figure 4A, the vast majority of individuals who self-identify as White have predominately European ancestry (CEU). Supplemental figure 4B illustrates the admixed ancestry typical of African Americans, self-identifying as Black. Self-identified Black individuals have a varying degree of African (YRI), European (CEU) and American (AMR) ancestry as can be seen in table 2. Likewise, self-identified Hispanic and Native American individuals (supplemental figures 4C and 4D) display an admixed ancestry largely comprised of American (AMR), European (CEU), African (YRI) and to a lesser extent, Chinese (CHB) and Japanese (JPT) ancestry. Finally, supplemental figure 4E includes self-identified Asians. As can also be seen in table 2 the ancestry of this subgroup is primarily of Chinese origin (CHB) but with measurable admixture of Japanese (JPT) and European (CEU).

GCTA Heritability and Weighted Genome-Wide Association Analysis of BMI

The GCTA heritability of Wave II, Wave III and Wave IV BMI based upon a sample of whiterespondents was estimated to be 0.82 (SE=0.081), 0.71 (SE=0.091), and0.67 (SE=0.084) respectively. Using a bivariate approach, the GCTA genetic correlation between Waves II/III, Waves II/IV and Waves III/IV were estimated to be 0.95 (SE=0.031), 0.85 (SE=0.042) and 0.99 (SE=0.033) respectively. The degree of relationship among the 1,886 individuals used in the genome-wide association analysis varies. In all, there are 614 individuals who are not 1st degree relatives (siblings) of others in the sample (N=614), 609 sibling pairs (N=1,218 individuals) and 18 sibling trios (N=54). Tables 2 and 3 provide the characteristics of the sample used for the genome-wide analysis. As noted previously and seen in table 2, approximately half of the sample self-identifies as White, 36% Black, 11% Hispanic, 4% Asian and less than 1% as Native American. Table 3 shows the frequency of males (48%) and females (52%) as well as the mean age and BMI of the sample at each wave of collection. Consistent with other studies, the mean (and standard deviation) of BMI increases throughout young adulthood.

To assess for the presence of systematic biases in genome-wide analyses, we generated a quantile-quantile(Q-Q) plot of the unweighted p-values (supplemental figure 5). As can be seen in the Q-Q plot, there is no evidence of widespread bias that is generating the associations.To illustrate the distribution of the weights as derived from the GIANT consortium, we provide asimple histogram (figure 1). As can be seen in figure 1, the vast majority of the p-values genome-wide are effectively down-weighted (<1; gray bar, N=522,093) while a fair number of p-values are substantially up-weighted (>1; colored bars, N=195,318). These weights were applied to each of the nominal p-values generated via the linear mixed effect model of longitudinal BMI.The resulting $-\log_{10}$ weighted p-values are displayed using a traditional Manhattan Plot (figure 2). The red horizontal line on figure 2 represents a genome-wide *significant* threshold ($p=5\times10^{-8}$) while the blue horizontal line represents a threshold of $p=5\times10^{-5}$. Table 4 displays the 39 SNP associations achieving$p< 5.0\times10^{-5}$ ordered by genomic location. We chose a threshold of $p< 5.0\times10^{-5}$ as a reasonable, albeit arbitrary, threshold for association signals that warrants potential follow-up as other studies have done (e.g. see Carty et al., 2012). In the table, we report the SNP marker name, chromosome, base pair location, and the nearest gene and where that SNP is located relative to that gene. We also report the allele conferring risk (increasing BMI units), the frequency of that allele and the other allele present in the data (reported as forward strand). Finally, we report the results of the analysis including the linear mixed model coefficient (corresponding 95% confidence interval), the unweighted p-value, the weight applied to that SNP marker and the corresponding weighted *p*-value. In this analysis, the linear mixed model coefficient may be interpreted as the additive effect of the risk allele on body mass index in units of kg/m^2. For example, the C allele of SNP rs1421085 (FTO gene) is associated with a 0.27 kg/m^2 increase in BMI. Overall, the list of 39 SNPs include some marker pairs that are likely in high linkage disequilibrium (LD) from the same genic region. Additionally, the influence of the weighting scheme can clearly be seen in the table. Using weights in this way allows for strongly significant markers to sift to the top even when down-weighted. There are 6 SNP markers that were down-weighted, yet still achieved a genome-wide suggestive level. However, the remaining 30 SNP markers were up-weighted.

Therefore, this is largely a list of SNPs that have been pushed towards the top of the association signals as they are SNPs with prior information indicating evidence of association with BMI (GIANT consortium) and achieved at least nominal significance in the Add Health sample. A notable signal includes the highly replicable FTO gene region (chromosome 16) providing evidence that the Add Health sibling pairs sample is an informative genetic dataset for future use.

Discussion

The primary focus of this study was to introduce the Add Health sibling pairs subsample genome-wide association data and conduct initial analyses to demonstrate the scientific potential of the data as a resource to the Add Health community of researchers. Given the unconventional (among traditional genome-wide studies) Add Health pairs subsample, we adopted a relatively unconventional approach to carry out the genome-wide analysis. First, it is estimated that 96% of all genome-wide studies have been conducted on people of European descent (Bustamante et al., 2011). The reasons and explanations for focusing so exclusively on samples of European descent range from convenience and efficiency (using existing cohort studies that focus on subjects of European descent) to minimizing sources of genetic heterogeneity (Pulit et al., 2010; Bustamante et al., 2011). However, recently, there has been a series of studies that have empirically demonstrated a critical role of multiethnic studies in genome research of complex disease(Pulit et al., 2010; Masunuru et al., 2012; Carlson et al., 2013; Gong et al., 2013; Manichaikul et al., 2012; Manku et al., 2013; Marigorta et al., 2013; Sabater-Lleal et al., 2013). Often, these multiethnic studies will conduct genome-wide analyses within a relatively homogenous European descent sample and simultaneously conduct a genome-wide analysis among a more genetically diverse sample such as African-Americans before combining the association signals using meta-analysis. An alternative approach, and one that was chosen for the present study, conducts the genome-wide analysis on the entire sample across multiple ethnic backgrounds. This approach has been successfully conducted in other studies of complex disease (for examples, see Kurreeman et al., 2012 and Xu et al., 2013). However, rather than use the Add Health sample as a discovery sample (often requiring very large sample sizes) we adopted a weighting scheme based upon the GIANT consortium (Speliotes et al., 2010) that is comprised of a series of European descent samples. Therefore, the approach taken for the present study is one that explores the extent to which the variants discovered in European descent GIANT consortium may also be of relevance to the multiethnic Add Health pairs subsample.

We note here that there are a multitude of valid and reasonable approaches that investigators may take when conducting a genetic study of a multiethnic, family-based sample with longitudinal measures of phenotype, behaviors and the environment. For example, Add Health researchers may be interested in imputing genotypes for purposes of combining association signals across multiple data sources genotyped on different platforms. Add Health researchers may also be interested in incorporating Add Health sampling weights and/or taking into account phenotypic clustering particularly when studying genetic risk factors within an environmental context. Furthermore, the family-based structure of the data

would allow for more specific analyses using informative sibling pair family units through a variety of family-based association approaches (e.g. FBAT, Laird et al., 2000).

Through the weighted genome-wide association analysis, we observed association signals that align with previous and in some cases, established genetic variants associated with BMI. For example, we were able to identify the FTO region that has been previously identified and replicated (Frayling et al., 2007). Additionally, we were able to identify variants that are upweighted through the GIANT consortium results, but do not achieve genome-wide significance in either GIANT or the Add Health sibling pairs subsample. These variants may be of particular interest for researchers who wish to explore GxE interactions in the Add Health sample to further explain the variability of the effect of these variants on BMI over time (age and development), behaviors and under particular environmental contexts.

The Add Health study is unique because of the explicit emphasis on properly characterizing the multilevel and multidimensional aspects of adolescents' lives as they transition to adulthood. This design in conjunction with the related and unrelated pairs data (see Harris et al., 2013) has expanded the scope of the gene-environment interaction perspective to a multilevel perspective in which environmental influences are measured at the level of the state (Boardman 2009), neighborhood (Cleveland et al., 2003; Beaver et al., 2012), and schools (Boardman et al., 2012). Most importantly, the research design enables the measurement of factors such as social norms (Boardman et al., 2008) that are otherwise difficult to assess. The assessment of these contextual factors has been highlighted as a critical area for future research in gene-environment interplay (Spittel et al., 2013) and the utilization of genome-wide data in conjunction with this social environmental backdrop may provide important insights in the etiology of complex morbidities such as obesity (Boone-Heinonen and Gordon-Larsen, 2012).

As described elsewhere (Boardman et al., 2013), the existing gene-environment interaction typology includes models in which genetic risk may be the most evident in the least risky, the most risky, or the typical environments. Depending on the anticipated GxE relationship and the specific phenotype, environments may either trigger or control genetic expression in a causal manner, or they may simply mask otherwise small genetic associations. Without a representation of the full range of environments, one may conclude that a specific polymorphism is either protective, risky, or not associated with a particular phenotype. Belksy and Pluess (2009) make a very strong case for the differential susceptibility hypothesis that argues that environmentally sensitive loci will be protective in the most enriching environments but deleterious in unhealthy environments. This cross-over association cannot be identified without a representative sample from the full continuum of environments that is, again, why the representativeness of the Add Health study is such an important resource in conjunction with the pairs data.

Finally, it is important to note that GWAS is but one use of genome wide data. For instance, the genome-wide relationship models discussed above (Yang et al., 2010) can be extended to incorporate these multilevel design features so that the contextual variation in the heritability of health behaviors can be examined using measured genetic similarity rather than assumed similarity from sibling-based models. Similarly, sibling fixed effects

approaches can take advantage of the "random assignment" of risk alleles to examine siblings residing and socializing in similar environments compared to those in very different social contexts (Fletcher et al., 2011). These methods provide unique and new possibilities to identify causal models and have thus far not been extended to the genome-wide level.

Add Health Sibling Pairs Subsample Data Access

The genome-wide data and phenotype measures used in this study will be made available to the scientific community through the NIH database of Genotypes and Phenotypes (dbGaP) by January 2015. Researchers interested in using the Add Health sibling pairs subsample genome-wide data will be required to access genotype data through the dbGaPauthorized access system. Once genotype data are available through the dbGaP and access has been granted, researchers who request other phenotypic data not in dbGaP will be able to apply for a Genome-wide Data Restricted Access Agreement through Add Health beginning in 2015 (http://www.cpc.unc.edu/projects/addhealth).This process will allow approved investigators access to the entire Add Health sibling pairs subsample longitudinal data in addition to the genome-wide data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). The genome-wide data generated for the Add Health sibling pairs subsample as well investigator effort for MBM, JDB, KMH, BD and AS was funded by grant R01-HD060726 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development.

References

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–1664. [PubMed: 19648217]

Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011; 12:246. [PubMed: 21682921]

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467(7311):52–58. [PubMed: 20811451]

Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.Computer software]. 2012:e0lme4. Retrieved from http://CRAN.R-project.org/packag.

Beaver KM, Wright JP, Delisi M, Daigle LE, Swatt ML, Gibson CL. Evidence of a gene × environment interaction in the creation of victimization: results from a longitudinal sample of adolescents. Int J Offender Ther Comp Criminol. 2007; 51(6):620–645. [PubMed: 17636204]

Belsky J, Pluess M. Beyond diathesis stress: differential susceptibility to environmental influences. Psychol Bull. 2009; 135(6):885–908. [PubMed: 19883141]

Boardman JD. State-level moderation of genetic tendencies to smoke. Am J Public Health. 2009; 99(3):480–486. [PubMed: 19150910]

Boardman JD, Domingue BW, Fletcher JM. How social and genetic factors predict friendship networks. ProcNatlAcadSci U S A. 2012; 109(43):17377–17381.

Boardman JD, Daw J, Freese J. Defining the environment in gene-environment research: lessons from social epidemiology. Am J Public Health. 2013; 103(Suppl 1):S64–S72. [PubMed: 23927514]

Boardman JD, Domingue BW, Blalock CL, Haberstick BC, Harris KM, McQueen MB. Is the Gene-Environment Interaction Paradigm Relevant to Genome-Wide Studies? The Case of Education and Body Mass Index. Demography. 2013

Boone-Heinonen J, Gordon-Larsen P. Obesogenic environments in youth: concepts and methods from a longitudinal national sample. Am J Prev Med. 2012; 42(5):e37–e46. [PubMed: 22516502]

Bustamante CD, Burchard EG, De la Vega FM. Genomics for the world. Nature. 2011; 475(7355): 163–165. [PubMed: 21753830]

Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, Schumacher FR, Peters U, Franceschini N, Ritchie MD, Duggan DJ, Spencer KL, Dumitrescu L, Eaton CB, Thomas F, Young A, Carty C, Heiss G, Le Marchand L, Crawford DC, Hindorff LA, Kooperberg CL, PAGE Consortium. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoSBiol. 2013; 11(9):e1001661.

Carty CL, Johnson NA, Hutter CM, Reiner AP, Peters U, Tang H, Kooperberg C. Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). Hum Mol Genet. 2012; 21(3):711–720. [PubMed: 22021425]

Cleveland HH. Disadvantaged neighborhoods and adolescent aggression: Behavioral genetic evidence of contextual effects. Journal of Research on Adolescence. 2003; 13(2):211–238.

Fletcher JM, Lehrer SF. Genetic lotteries within families. J Health Econ. 2011; 30(4):647–659. [PubMed: 21664708]

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007; 316(5826):889–894. [PubMed: 17434869]

Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, Nguyen EA, Drake KA, Huntsman S, Hu D, Sen S, Davis A, Farber HJ, Avila PC, Brigino-Buenaventura E, Lenoir MA, Meade K, Serebrisky D, Borrell LN, Rodríguez-Cintrón W, Estrada AM, Mendoza KS, Winkler CA, Klitz W, Romieu I, London SJ, Gilliland F, Martinez F, Bustamante C, Williams LK, Kumar R, Rodríguez-Santana JR, Burchard EG. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: The Genes-environments & Admixture in Latino Americans study. J Allergy ClinImmunol. 2014

Gong J, Schumacher F, Lim U, Hindorff LA, Haessler J, Buyske S, Carlson CS, Rosse S, Bůzková P, Fornage M, Gross M, Pankratz N, Pankow JS, Schreiner PJ, Cooper R, Ehret G, Gu CC, Houston D, Irvin MR, Jackson R, Kuller L, Henderson B, Cheng I, Wilkens L, Leppert M, Lewis CE, Li R, Nguyen KD, Goodloe R, Farber-Eger E, Boston J, Dilks HH, Ritchie MD, Fowke J, Pooler L, Graff M, Fernandez-Rhodes L, Cochrane B, Boerwinkle E, Kooperberg C, Matise TC, Le Marchand L, Crawford DC, Haiman CA, North KE, Peters U. Fine Mapping and Identification of BMI Loci in African Americans. Am J Hum Genet. 2013; 93(4):661–671. [PubMed: 24094743]

Harris KM, Halpern CT, Haberstick BC, Smolen A. The National Longitudinal Study of Adolescent Health (Add Health) sibling pairs data. Twin Res Hum Genet. 2013; 16(1):391–398. [PubMed: 23231780]

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008; 24(24): 2938–2939. [PubMed: 18974171]

Kurreeman FA, Stahl EA, Okada Y, Liao K, Diogo D, Raychaudhuri S, Freudenberg J, Kochi Y, Patsopoulos NA, Gupta N, Sandor C, Bang SY, Lee HS, Padyukov L, Suzuki A, Siminovitch K, Worthington J, Gregersen PK, Hughes LB, Reynolds RJ, Bridges SL, Bae SC, Yamamoto K, Plenge RM. Use of a multiethnic approach to identify rheumatoid- arthritis-susceptibility loci, 1p36 and 17q12. Am J Hum Genet. 2012; 90(3):524–532. [PubMed: 22365150]

Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000; 19(Suppl 1):S36–S42. [PubMed: 11055368]

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319(5866):1100–1104. [PubMed: 18292342]

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010; 26(22):2867–2873. [PubMed: 20926424]

Manichaikul A, Palmas W, Rodriguez CJ, Peralta CA, Divers J, Guo X, Chen WM, Wong Q, Williams K, Kerr KF, Taylor KD, Tsai MY, Goodarzi MO, Sale MM, Diez-Roux AV, Rich SS, Rotter JI, Mychaleckyj JC. Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. PLoS Genet. 2012; 8(4):e1002640. [PubMed: 22511882]

Manku H, Langefeld CD, Guerra SG, Malik TH, Alarcon-Riquelme M, Anaya JM, Bae SC, Boackle SA, Brown EE, Criswell LA, Freedman BI, Gaffney PM, Gregersen PA, Guthridge JM, Han SH, Harley JB, Jacob CO, James JA, Kamen DL, Kaufman KM, Kelly JA, Martin J, Merrill JT, Moser KL, Niewold TB, Park SY, Pons-Estel BA, Sawalha AH, Scofield RH, Shen N, Stevens AM, Sun C, Gilkeson GS, Edberg JC, Kimberly RP, Nath SK, Tsao BP, Vyse TJ. Trans-ancestral studies fine map the SLE-susceptibility locus TNFSF4. PLoS Genet. 2013; 9(7):e1003554. [PubMed: 23874208]

Marigorta UM, Navarro A. High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genet. 2013; 9(6):e1003566. [PubMed: 23785302]

Morrison J. Characterization and correction of error in genome-wide IBD estimation for samples with population structure. Genet Epidemiol. 2013; 37(6):635–641. [PubMed: 23740691]

Musunuru K, Romaine SP, Lettre G, Wilson JG, Volcik KA, Tsai MY, Taylor HA, Schreiner PJ, Rotter JI, Rich SS, Redline S, Psaty BM, Papanicolaou GJ, Ordovas JM, Liu K, Krauss RM, Glazer NL, Gabriel SB, Fornage M, Cupples LA, Buxbaum SG, Boerwinkle E, Ballantyne CM, Kathiresan S, Rader DJ. Multi-ethnic analysis of lipid-associated loci: the NHLBI CARe project. PLoS One. 2012; 7(5):e36473. [PubMed: 22629316]

Nelson SC, Doheny KF, Laurie CC, Mirel DB. Is 'forward' the same as 'plus'?...and other adventures in SNP allele nomenclature. Trends Genet. 2012; 28(8):361–363. [PubMed: 22658725]

Pulit SL, Voight BF, de Bakker PI. Multiethnic genetic association studies improve power for locus discovery. PLoS One. 2010; 5(9):e12600. [PubMed: 20838612]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–575. [PubMed: 17701901]

Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet. 2006; 78(2):243–252. [PubMed: 16400608]

Roeder K, Devlin B, Wasserman L. Improving power in genome-wide association studies: weights tip the scale. Genet Epidemiol. 2007; 31(7):741–747. [PubMed: 17549760]

Roeder K, Wasserman L. Genome-Wide Significance Levels and Weighted Hypothesis Testing. Stat Sci. 2009; 24(4):398–413. [PubMed: 20711421]

Sabater-Lleal M, Huang J, Chasman D, Naitza S, Dehghan A, Johnson AD, Teumer A, Reiner AP, Folkersen L, Basu S, Rudnicka AR, Trompet S, Mälarstig A, Baumert J, Bis JC, Guo X, Hottenga

JJ, Shin SY, Lopez LM, Lahti J, Tanaka T, Yanek LR, Oudot-Mellakh T, Wilson JF, Navarro P, Huffman JE, Zemunik T, Redline S, Mehra R, Pulanic D, Rudan I, Wright AF, Kolcic I, Polasek O, Wild SH, Campbell H, Curb JD, Wallace R, Liu S, Eaton CB, Becker DM, Becker LC, Bandinelli S, Räikkönen K, Widen E, Palotie A, Fornage M, Green D, Gross M, Davies G, Harris SE, Liewald DC, Starr JM, Williams FM, Grant PJ, Spector TD, Strawbridge RJ, Silveira A, Sennblad B, Rivadeneira F, Uitterlinden AG, Franco OH, Hofman A, van Dongen J, Willemsen G, Boomsma DI, Yao J, Swords Jenny N, Haritunians T, McKnight B, Lumley T, Taylor KD, Rotter JI, Psaty BM, Peters A, Gieger C, Illig T, Grotevendt A, Homuth G, Völzke H, Kocher T, Goel A, Franzosi MG, Seedorf U, Clarke R, Steri M, Tarasov KV, Sanna S, Schlessinger D, Stott DJ, Sattar N, Buckley BM, Rumley A, Lowe GD, McArdle WL, Chen MH, Tofler GH, Song J, Boerwinkle E, Folsom AR, Rose LM, Franco-Cereceda A, Teichert M, Ikram MA, Mosley TH, Bevan S, Dichgans M, Rothwell PM, Sudlow CL, Hopewell JC, Chambers JC, Saleheen D, Kooner JS, Danesh J, Nelson CP, Erdmann J, Reilly MP, Kathiresan S, Schunkert H, Morange PE, Ferrucci L, Eriksson JG, Jacobs D, Deary IJ, Soranzo N, Witteman JC, de Geus EJ, Tracy RP, Hayward C, Koenig W, Cucca F, Jukema JW, Eriksson P, Seshadri S, Markus HS, Watkins H, Samani NJ, Wallaschofski H, Smith NL, Tregouet D, Ridker PM, Tang W, Strachan DP, Hamsten A, O'Donnell CJ. 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):1310–1324. [PubMed: 23969696]

Shi G, Rice TK, Gu CC, Rao DC. Application of three-level linear mixed-effects model incorporating gene-age interactions for association analysis of longitudinal family data. BMC Proc. 2009; 3(Suppl 7):S89. [PubMed: 20018085]

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Mägi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, Weyant RJ, Segrè AV, Estrada K, Liang L, Nemesh J, Park JH, Gustafsson S, Kilpeläinen TO, Yang J, Bouatia-Naji N, Esko T, Feitosa MF, Kutalik Z, Mangino M, Raychaudhuri S, Scherag A, Smith AV, Welch R, Zhao JH, Aben KK, Absher DM, Amin N, Dixon AL, Fisher E, Glazer NL, Goddard ME, Heard-Costa NL, Hoesel V, Hottenga JJ, Johansson A, Johnson T, Ketkar S, Lamina C, Li S, Moffatt MF, Myers RH, Narisu N, Perry JR, Peters MJ, Preuss M, Ripatti S, Rivadeneira F, Sandholt C, Scott LJ, Timpson NJ, Tyrer JP, van Wingerden S, Watanabe RM, White CC, Wiklund F, Barlassina C, Chasman DI, Cooper MN, Jansson JO, Lawrence RW, Pellikka N, Prokopenko I, Shi J, Thiering E, Alavere H, Alibrandi MT, Almgren P, Arnold AM, Aspelund T, Atwood LD, Balkau B, Balmforth AJ, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Biebermann H, Blakemore AI, Boes T, Bonnycastle LL, Bornstein SR, Brown MJ, Buchanan TA, Busonero F, Campbell H, Cappuccio FP, Cavalcanti-Proença C, Chen YD, Chen CM, Chines PS, Clarke R, Coin L, Connell J, Day IN, den Heijer M, Duan J, Ebrahim S, Elliott P, Elosua R, Eiriksdottir G, Erdos MR, Eriksson JG, Facheris MF, Felix SB, Fischer-Posovszky P, Folsom AR, Friedrich N, Freimer NB, Fu M, Gaget S, Gejman PV, Geus EJ, Gieger C, Gjesing AP, Goel A, Goyette P, Grallert H, Grässler J, Greenawalt DM, Groves CJ, Gudnason V, Guiducci C, Hartikainen AL, Hassanali N, Hall AS, Havulinna AS, Hayward C, Heath AC, Hengstenberg C, Hicks AA, Hinney A, Hofman A, Homuth G, Hui J, Igl W, Iribarren C, Isomaa B, Jacobs KB, Jarick I, Jewell E, John U, Jørgensen T, Jousilahti P, Jula A, Kaakinen M, Kajantie E, Kaplan LM, Kathiresan S, Kettunen J, Kinnunen L, Knowles JW, Kolcic I, König IR, Koskinen S, Kovacs P, Kuusisto J, Kraft P, Kvaløy K, Laitinen J, Lantieri O, Lanzani C, Launer LJ, Lecoeur C, Lehtimäki T, Lettre G, Liu J, Lokki ML, Lorentzon M, Luben RN, Ludwig B, Manunta P, Marek D, Marre M, Martin NG, McArdle WL, McCarthy A, McKnight B, Meitinger T, Melander O, Meyre D, Midthjell K, Montgomery GW, Morken MA, Morris AP, Mulic R, Ngwa JS, Nelis M, Neville MJ, Nyholt DR, O'Donnell CJ, O'Rahilly S, Ong KK, Oostra B, Paré G, Parker AN, Perola M, Pichler I, Pietiläinen KH, Platou CG, Polasek O, Pouta A, Rafelt S, Raitakari O, Rayner NW, Ridderstråle M, Rief W, Ruokonen A, Robertson NR, Rzehak P, Salomaa V, Sanders AR, Sandhu MS, Sanna S, Saramies J, Savolainen MJ, Scherag S, Schipf S, Schreiber S, Schunkert H, Silander K, Sinisalo J, Siscovick DS, Smit JH, Soranzo N, Sovio U, Stephens J, Surakka I, Swift AJ, Tammesoo ML, Tardif JC, Teder-Laving M, Teslovich TM, Thompson JR, Thomson B, Tönjes A, Tuomi T, van Meurs JB, van Ommen GJ, Vatin V, Viikari J, Visvikis-Siest S, Vitart V, Vogel CI, Voight BF, Waite LL, Wallaschofski H, Walters GB, Widen E, Wiegand S, Wild SH, Willemsen G, Witte

DR, Witteman JC, Xu J, Zhang Q, Zgaga L, Ziegler A, Zitting P, Beilby JP, Farooqi IS, Hebebrand J, Huikuri HV, James AL, Kähönen M, Levinson DF, Macciardi F, Nieminen MS, Ohlsson C, Palmer LJ, Ridker PM, Stumvoll M, Beckmann JS, Boeing H, Boerwinkle E, Boomsma DI, Caulfield MJ, Chanock SJ, Collins FS, Cupples LA, Smith GD, Erdmann J, Froguel P, Grönberg H, Gyllensten U, Hall P, Hansen T, Harris TB, Hattersley AT, Hayes RB, Heinrich J, Hu FB, Hveem K, Illig T, Jarvelin MR, Kaprio J, Karpe F, Khaw KT, Kiemeney LA, Krude H, Laakso M, Lawlor DA, Metspalu A, Munroe PB, Ouwehand WH, Pedersen O, Penninx BW, Peters A, Pramstaller PP, Quertermous T, Reinehr T, Rissanen A, Rudan I, Samani NJ, Schwarz PE, Shuldiner AR, Spector TD, Tuomilehto J, Uda M, Uitterlinden A, Valle TT, Wabitsch M, Waeber G, Wareham NJ, Watkins H, Wilson JF, Wright AF, Zillikens MC, Chatterjee N, McCarroll SA, Purcell S, Schadt EE, Visscher PM, Assimes TL, Borecki IB, Deloukas P, Fox CS, Groop LC, Haritunians T, Hunter DJ, Kaplan RC, Mohlke KL, O'Connell JR, Peltonen L, Schlessinger D, Strachan DP, van Duijn CM, Wichmann HE, Frayling TM, Thorsteinsdottir U, Abecasis GR, Barroso I, Boehnke M, Stefansson K, North KE, McCarthy MI, Hirschhorn JN, Ingelsson E, Loos RJ. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42(11):937–948. [PubMed: 20935630]

Spittel ML, Spotts EL, Deeds BG. Integration of behavioral, social science and genetics research: exploring public health significance. Am J Public Health. 2013; 103(Suppl 1):S5–S7. [PubMed: 23927547]

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012; 91(1):122–138. [PubMed: 22748210]

Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, Pei D, Scheet P, Burchard EG, Eng C, Huntsman S, Torgerson DG, Dean M, Winick NJ, Martin PL, Camitta BM, Bowman WP, Willman CL, Carroll WL, Mullighan CG, Bhojwani D, Hunger SP, Pui CH, Evans WE, Relling MV, Loh ML, Yang JJ. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. J Natl Cancer Inst. 2013; 105(10):733–742. [PubMed: 23512250]

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42(7):565–569. [PubMed: 20562875]

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88(1):76–82. [PubMed: 21167468]

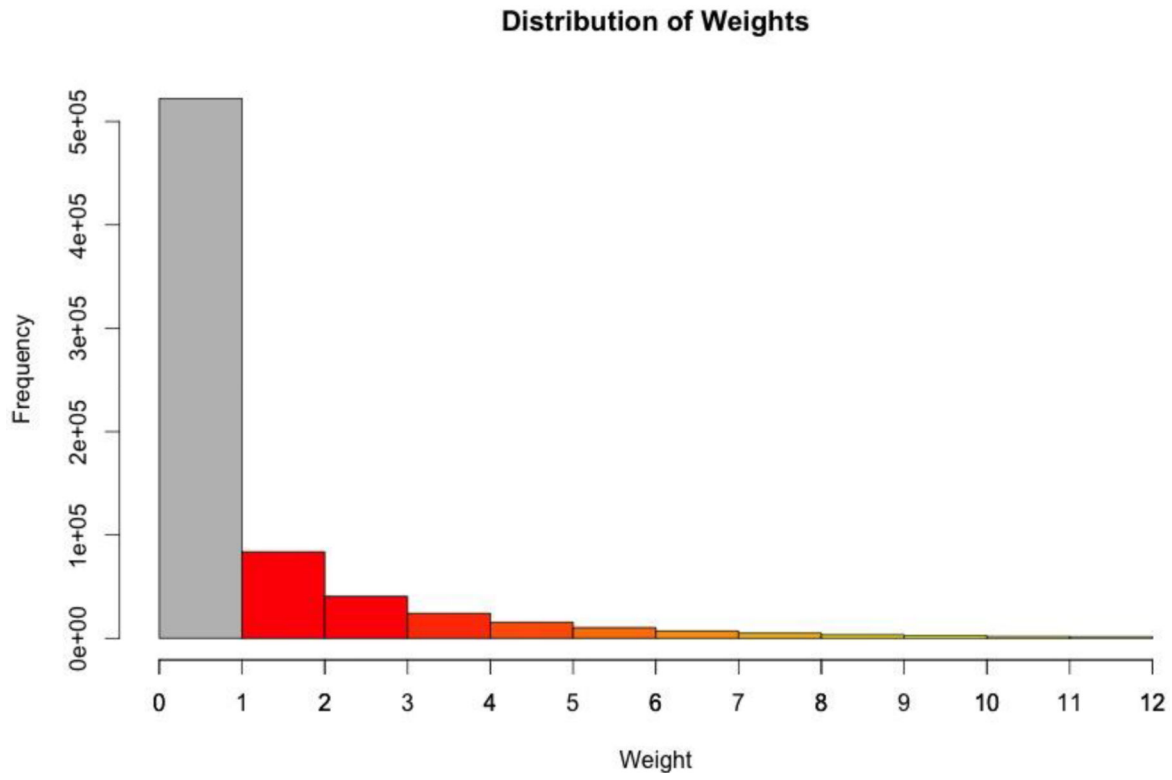

Figure 1.
Histogram of the weights derived from the GIANT consortium applied to the genome-wide association p-values.

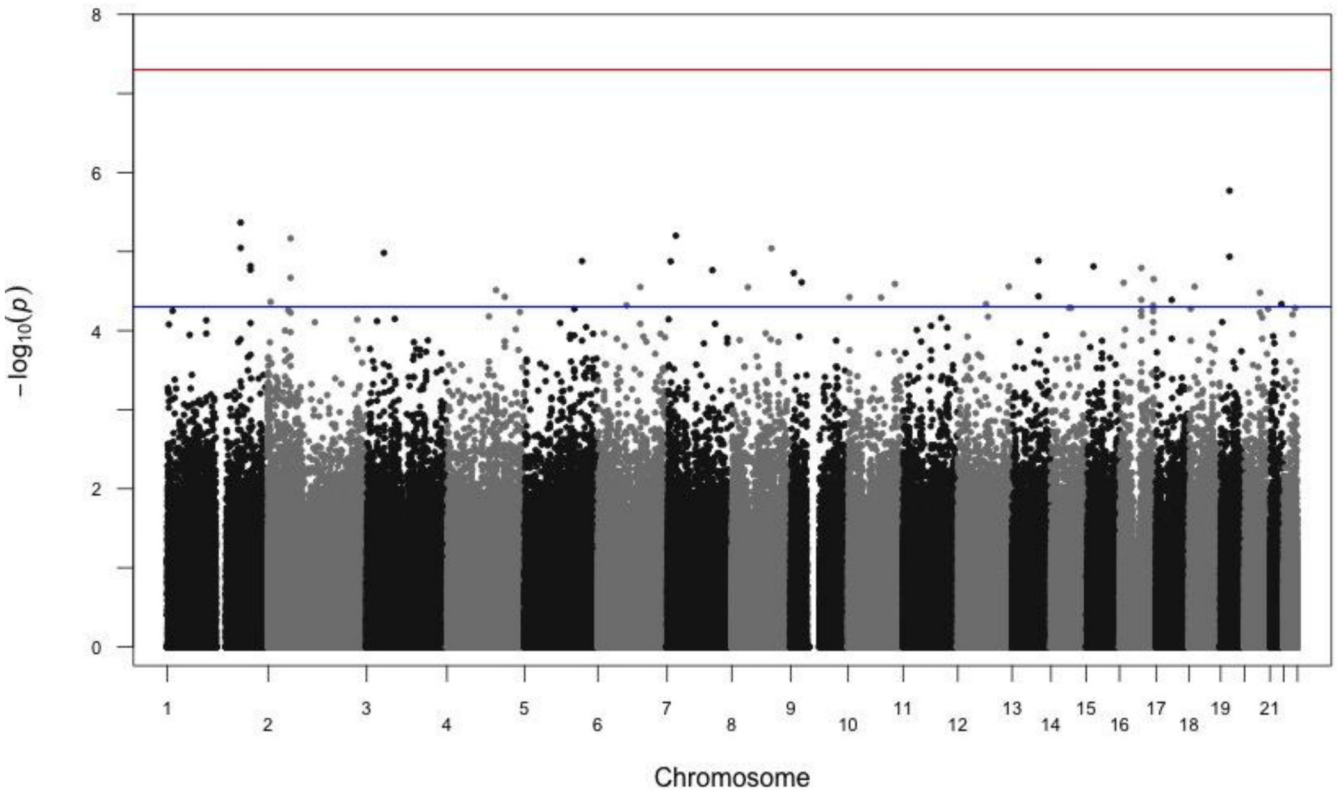

Figure 2.
Manhattan Plot of the Weighted GWAS Association Signals

Table 1

Observed versus Expected Relationship Status

		Observed Relationship (KING Kinship Coefficient)*					
		MZ	**1st Degree**	**2nd Degree**	**3rd Degree**	**NR**	***TOTAL***
Expected Relationship*	MZ	0	1	0	0	0	1
	DZ	1	173	2	0	1	177
	FS	0	480	33	0	6	519
	HS	0	6	38	2	5	51
	AV	0	0	0	2	1	3
	CO	0	1	0	19	18	38
	NR	0	1	1	264	1,780,270	1,780,536
	UD	1	2	0	0	0	3
	TOTAL	2	664	74	287	1,780,301	**1,781,328**

Note: MZ=monozygotic twin pair, DZ=dizygotic twin pair, FS=full sibling pair, HS=half-sibling pair, AV=avuncular pair, CO=cousin pair, NR=not related and UD=undetermined.

Table 2

Self-Identified Race and Proportion of Genetic Ancestry*

	Ancestral Population**				
Self-Identified Race* **	***Europe (CEU)***	***Africa (YRI)***	***Americas (AMR)***	***China (CHB)***	***Japan (JPT)***
White N=917 (48.6%)	0.983	0.007	0.004	0.003	0.003
Black N=677 (35.9%)	0.178	0.803	0.008	0.007	0.004
Hispanic N=209 (11.1%)	0.612	0.105	0.251	0.017	0.015
Asian N=73 (3.9%)	0.068	0.005	0.001	0.807	0.120
Native American N=8 (0.4%)	0.338	0.280	0.332	0.021	0.029

* Proportion of genetic ancestry as estimated by ADMIXTURE.

** Ancestral populations derived from the HGDP and HapMap reference populations.

*** As defined by the Add Health Race ("ah_race") variable (2 respondents coded as unknown).

Table 3

Characteristics of the Genome-Wide Association Sample

Characteristic	
Biological Sex *	
Male	905 (48%)
Female	981 (52%)
Age **	
Wave II (N=1761)	16.4 (1.7)
Wave III (N=1634)	22.4 (1.7)
Wave IV (N=1886)	28.9 (1.7)
Body Mass Index (BMI) **	
Wave II (N=1688)	23.4 (5.1)
Wave III (N=1562)	26.6 (6.4)
Wave IV (N=1859)	29.4 (7.9)

* Values expressed as sample size (relative percentage).

** N refers to the number of non-missing values for each wave and variable. Values expressed as mean (standard deviation).

Table 4

Top weighted genome-wide association signals ($p < 5 \times 10^5$) ordered by genomic location.

Marker Characteristics*				Allele Information**			Analysis Results***			
SNP	**Chr**	**Position**	**Gene (loc)**	**Risk**	**Oth**	**Freq**	**Coef (95% CI)**	***P-Value***	**Wght**	P_W***-Value***
rs199950	1	181590858	CACNA1E (I)	G	A	0.31	1.00 (0.58,1.42)	0.000003	0.65	0.000004
rs199939	1	181595979	CACNA1E (I)	T	C	0.29	0.96(0.53,1.40)	0.000013	1.45	0.000009
rs7415921	1	205910883	SLC26A9 (I)	T	G	0.49	0.77(0.37,1.16)	0.000135	8.86	0.000015
rs12047830	1	205916699	SLC26A9 FAM72A (IG)	A	G	0.45	0.77(0.38,1.15)	0.000103	6.08	0.000017
rs1358136	2	6183596	LOC400940 (IG)	A	G	0.54	0.76(1.19,0.34)	0.000436	10.05	0.000043
rs11686766	2	55385259	RTN4 C2orf63 (IG)	T	C	0.06	1.66(0.89,2.43)	0.000021	3.13	0.000007
rs13032294	2	55404883	C2orf63 (C-N)	C	T	0.06	1.60 (0.83,2.37)	0.000050	2.34	0.000021
rs33500	3	42427191	LOC100287105 LYZL4 (IG)	C	T	0.13	1.09(0.55,1.64)	0.000086	8.32	0.000010
rs1587734	4	120809455	LOC730456 MAD2L1 (IG)	T	C	0.75	0.82(1.25,0.38)	0.000223	7.22	0.000031
rs1879685	4	142484038	ZNF330 LOC100286983	T	C	0.22	0.96(0.49,1.42)	0.000054	1.45	0.000037
rs58644	5	141791827	SPRY4 FGF1 (IG)	T	C	0.26	0.93(0.48,1.39)	0.000065	4.96	0.000013
rs652462	6	71024741	COL9A1 FAM135A (IG)	A	C	0.40	0.80 (0.40,1.21)	0.000103	2.16	0.000048
rs156192	6	104917557	FLJ10088 HACE1 (IG)	T	C	0.62	0.81(1.21,0.42)	0.000058	2.06	0.000028
rs1468298	7	8800995	NXPH1 PER4 (IG)	G	A	0.13	1.14(0.60,1.68)	0.000036	2.73	0.000013
rs4722037	7	21586595	DNAH11 (I)	T	C	0.62	0.96(1.40,0.52)	0.000017	2.71	0.000006
rs634010	7	111698108	DOCK4 (I)	T	C	0.24	1.04(0.59,1.48)	0.000005	0.28	0.000017
rs7013912	8	40200810	C8orf4 ZMAT4 (IG)	C	A	0.88	1.15(1.72,0.59)	0.000061	2.14	0.000028
rs10090663	8	97979470	PGCP (I)	C	T	0.08	1.44(0.77,2.10)	0.000025	2.77	0.000009
rs10976070	9	7150975	KDM4C (I)	C	T	0.51	0.84(1.22,0.46)	0.000017	0.89	0.000019

Marker Characteristics*				Allele Information**			Analysis Results***			
SNP	**Chr**	**Position**	**Gene (loc)**	**Risk**	**Oth**	**Freq**	**Coef (95% CI)**	***P-Value***	**Wght**	***P_W-Value***
rs6475914	9	26553400	TUSC1 C9orf82 (IG)	G	T	0.07	1.43(0.69,2.17)	0.000155	6.32	0.000025
rs7898695	10	3184507	PITRM1 (I)	G	A	0.57	0.72(1.09,0.35)	0.000149	3.96	0.000038
rs11002893	10	81012071	ZMIZ1 (I)	A	G	0.81	0.86(1.32,0.40)	0.000236	6.19	0.000038
rs885567	10	115357179	NRAP (I)	C	T	0.73	0.92(1.33,0.51)	0.000013	0.50	0.000026
rs686407	12	69744104	LYZ (I)	C	T	0.62	0.85(1.30,0.40)	0.000209	4.53	0.000046
rs3803155	12	125834395	TMEM132B (C-N)	A	G	0.35	0.81(0.40,1.22)	0.000104	3.73	0.000028
rs2210727	13	84492665	SLITRK1 SLITRK6 (IG)	C	T	0.52	0.81(1.20,0.41)	0.000061	4.66	0.000013
rs7327886	13	84518116	SLITRK1 SLITRK6 (IG)	C	T	0.53	0.74(1.14,0.33)	0.000356	9.66	0.000037
rs3752854	15	38158906	MEIS2 TMCO5A (IG)	G	T	0.07	1.60 (0.86,2.34)	0.000021	1.34	0.000015
rs2243716	16	10088149	GRIN2A (I)	G	A	0.15	1.14(0.62,1.65)	0.000014	0.55	0.000025
rs1421085	16	53800954	FTO (I)	C	T	0.27	0.84(0.40,1.29)	0.000190	11.83	0.000016
rs12149832	16	53842908	FTO (I)	A	G	0.28	0.79(0.35,1.23)	0.000482	11.83	0.000041
rs16958392	16	82770272	CDH13 (I)	G	A	0.10	1.25(0.62,1.88)	0.000107	2.25	0.000048
rs12926503	16	84031482	NECAB2 (I)	A	G	0.85	1.23(1.77,0.69)	0.000009	0.41	0.000022
rs2269457	17	38254689	NR1D1 (I)	G	A	0.34	0.79(0.38,1.20)	0.000156	3.80	0.000041
rs877128	18	13911628	MC2R (I)	A	G	0.18	0.98(0.49,1.46)	0.000076	2.71	0.000028
rs8105895[P]	19	22215457	ZNF208 ZNF257 (IG)	T	C	0.14	1.26(0.73,1.79)	0.000003	1.86	0.000002
rs17450788	19	22224755	ZNF208 ZNF257 (IG)	C	T	0.13	1.22(0.67,1.77)	0.000015	1.33	0.000012
rs6016086	20	38142337	LOC33956 MAFB (IG)	G	A	0.18	0.99(0.49,1.48)	0.000093	2.82	0.000033
rs460976[P]	21	42835494	MX1 TMPRSS2 (IG)	G	A	0.96	2.22(3.20,1.23)	0.000011	0.23	0.000046

*SNP markers denoted with a P were derived from proxy SNPs in high LD with SNPs from the GIANT consortium. Gene codes: I=intron, IG=intergenetic, C-N=Coding, nonsynonymous.

**Risk refers to the allele that corresponds to an increase in BMI units while Oth refers to the other (non-risk) allele. Freq corresponds to the allele frequency of the risk allele.

***Coeff (95% CI) is the beta coefficient and 95% confidence interval as estimated through a linear mixed effects model. The *P-Value* is the unweighted nominal p-value, the Wght is the GIANT-derived weights and *P_W-value* is the weighted p-value