Ultraconserved Elements in the Human Genome: Association and Transmission Analyses of Highly Constrained Single-Nucleotide Polymorphisms

Charleston W. K. Chiang,[*,†,‡,1] Ching-Ti Liu,[§] Guillaume Lettre,[**,††] Leslie A. Lange,[‡‡] Neal W. Jorgensen,[§§] Brendan J. Keating,[***] Sailaja Vedantam,[†,‡] Nora L. Nock,[†††] Nora Franceschini,[‡‡‡] Alex P. Reiner,[§§§] Ellen W. Demerath,[****] Eric Boerwinkle,[††††] Jerome I. Rotter,[‡‡‡‡] James G. Wilson,[§§§§] Kari E. North,[‡‡‡] George J. Papanicolaou,[*****] L. Adrienne Cupples,[§,*****] Genetic Investigation of ANthropometric Traits (GIANT) Consortium,[2] Joanne M. Murabito,[†††††,‡‡‡‡‡] and Joel N. Hirschhorn[*,†,‡,3]

[*]Department of Genetics, Harvard Medical School, Boston, Massachusetts, [†]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, [‡]Program in Genomics and Divisions of Genetics and Endocrinology, Children's Hospital, Boston, Massachusetts, [§]Department of Biostatistics, School of Public Health, Boston University, Boston, Massachusetts, [**]Faculté de Médecine, Université de Montréal, Montréal, Québec H3C 3J7, Canada, [††]Institut de Cardiologie de Montréal, Montréal, Québec H1T 1C8, Canada, [‡‡]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, [§§]Department of Biostatistics, School of Public Health, University of Washington, Seattle, Washington, [***]Cardiovascular Institute and the Institute for Translational Medicine and Therapeutics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, [†††]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, [‡‡‡]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, [§§§]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, [****]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota, [††††]Human Genetics Center, University of Texas Health Science Center, Houston, Texas, [‡‡‡‡]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, [§§§§]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, [*****]Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, Bethesda, Maryland, [†††††]Framingham Heart Study, Framingham, Massachusetts, and [‡‡‡‡‡]Section of General Internal Medicine, Department of Medicine, Boston University School of Medicine, Boston, Massachusetts

ABSTRACT Ultraconserved elements in the human genome likely harbor important biological functions as they are dosage sensitive and are able to direct tissue-specific expression. Because they are under purifying selection, variants in these elements may have a lower frequency in the population but a higher likelihood of association with complex traits. We tested a set of highly constrained SNPs (hcSNPs) distributed genome-wide among ultraconserved and nearly ultraconserved elements for association with seven traits related to reproductive (age at natural menopause, number of children, age at first child, and age at last child) and overall [longevity, body mass index (BMI), and height] fitness. Using up to 24,047 European-American samples from the National Heart, Lung, and Blood Institute Candidate Gene Association Resource (CARe), we observed an excess of associations with BMI and height. In an independent replication panel the most strongly associated SNPs showed an 8.4-fold enrichment of associations at the nominal level, including three variants in previously identified loci and one in a locus (*DENND1A*) previously shown to be associated with polycystic ovary syndrome. Finally, using 1430 family trios, we showed that the transmissions from heterozygous parents to offspring of the derived alleles of rare (frequency ≤0.5%) hcSNPs are not biased, particularly after adjusting for the rates of genotype missingness and error in the data. The lack of transmission bias ruled out an immediately and strongly deleterious effect due to the rare derived alleles, consistent with the observation that mice homozygous for the deletion of ultraconserved elements showed no overt phenotype. Our study also illustrated the importance of carefully modeling potential technical confounders when analyzing genotype data of rare variants.

REFERENCE genome alignments of multiple distantly related species have identified thousands of elements that are 98–100% identical (Bejerano *et al.* 2004; Derti *et al.* 2006; Chiang *et al.* 2008; Stephen *et al.* 2008). These ultraconserved or nearly ultraconserved elements are expected to harbor important biological functions as they are dosage sensitive (Derti *et al.* 2006; Chiang *et al.* 2008), are enriched in or near specific classes of genes (Bejerano *et al.*

2004), and are under purifying selection (Drake *et al.* 2006; Chen *et al.* 2007; Katzman *et al.* 2007; Sakuraba *et al.* 2008). Indeed, a number of reports have demonstrated the ability of intergenic and intronic highly conserved elements to direct tissue-specific expression (Woolfe *et al.* 2005; McEwen *et al.* 2006; Pennacchio *et al.* 2006; Ahituv *et al.* 2007; Paparidis *et al.* 2007; Visel *et al.* 2008). As such, DNA sequence variants within highly conserved elements may have a higher likelihood of being associated with phenotypic traits, particularly if examined directly within suspected disease genomic regions (such as under linkage peaks, near candidate genes, or within genetic association loci defined by linkage disequilibrium).

Although comparative genomic approaches comparing human sequences to orthologous sequences have successfully identified variants in noncoding enhancer regions that contribute to the pathogenesis of complex human diseases [*e.g.*, for Hirschsprung disease (Emison *et al.* 2005)], such investigations have mostly been limited to candidate gene studies. A recent evaluation of the types of variants reported by genome-wide association (GWA) studies showed some enrichment of trait-associated variants in conserved sequences (Hindorff *et al.* 2009), suggesting the potential utility of systematically screening variants in conserved sequences. Moreover, functionally important variants among conserved elements have not been as well assayed in recent GWA studies as they are expected to be of lower frequency in the population. Indeed, prioritizing variants on the basis of evolutionary constraint has been shown to narrow down the potential causal variant to Mendelian disease in resequenced exome data (Cooper *et al.* 2010). Therefore, it may be useful to directly genotype variants within extremely constrained regions in large cohorts to assess genotype–phenotype correlations.

In such genotype–phenotype correlations, however, no *a priori* hypothesis predicts which specific phenotypes might be affected. Despite the assumed importance of these extremely highly conserved elements, mice heterozygous or homozygous for a deletion of any of four nonexonic ultraconserved enhancers appeared to be phenotypically normal, with normal viability and fertility (Ahituv *et al.* 2007). This observation suggests that the effects of altering these highly conserved elements may affect fitness across generations (*i.e.*, reproductive fitness) but that effects on individual organisms may be subtle enough that they are not apparent in a laboratory setting (Derti *et al.* 2006; Ahituv *et al.* 2007). Because of this possibility, we decided to examine the association of SNPs in ultraconserved and nearly ultraconserved elements (collectively denoted as nUCEs) genome-wide with various phenotypes related to long-term reproductive and overall fitness in human populations. Specifically, we examined the number of children, age at birth of first child, age at birth of last child, and age at natural menopause as phenotypes representative of reproductive fitness. In addition, we also examined longevity and anthropometric traits such as body mass index (BMI) and height as examples of overall fitness, as they may reflect the general nutritional status and well-being during childhood or adulthood. Furthermore, the availability of family data in our cohort provided an alternative approach to test the hypothesis that nUCEs, collectively, are currently under strong purifying (negative) selection. Thus, we also examined the likelihood of the transmission from heterozygous parents to offspring of the derived allele of nUCE SNPs, as an additional measure of the purifying selective pressure reported for these highly conserved regions.

We used the Candidate Gene Association Resource (CARe) established by the National Heart, Lung, and Blood Institute (NHLBI) (Musunuru *et al.* 2010; Lettre *et al.* 2011) as the study population to test our hypothesis. CARe is a consortium involving planned genetic analysis across nine NHLBI cohorts and encompasses cohorts with population-, community-, family-, and hospital-based designs. Each CARe cohort contains an extensive catalog of phenotypic data, and together >32,000 self-identified European-American subjects were genotyped with a cardiovascular gene-centric 50K Illumina Human CVD array (the ITMAT-Broad-CARe, or "IBC", chip) designed to densely map ~2000 candidate gene loci (Keating *et al.* 2008; Musunuru *et al.* 2010). Additionally, 816 SNPs within all genomic elements with ≥98% conservation genome-wide were included on the IBC chip to screen variants in highly conserved regions for their association with a variety of phenotypes (Derti *et al.* 2006; Chiang *et al.* 2008; Keating *et al.* 2008).

In the present study, we first identified a subset of the 816 SNPs in nUCEs that appear to be under strong evolutionary constraint [$N = 533$, which we refer to as highly constrained SNPs (hcSNPs)]. We then examined the association of hcSNPs with phenotypes of reproductive fitness and overall fitness described above and followed up the best associated SNPs in an additional >100,000 replication samples from the GIANT consortium (Lango Allen *et al.* 2010; Speliotes *et al.* 2010). Of the 19 hcSNPs followed up, eight variants replicated at the nominal level, including three variants that lie within BMI and height loci previously identified in GWA studies (Lango Allen *et al.* 2010; Speliotes *et al.* 2010). The strongest association outside of previously known loci is an intronic SNP in *DENND1A*. Although this variant's association to BMI did not survive Bonferroni correction for the number of SNPs/trait pairs followed up, it has been

doi: 10.1534/genetics.112.141945
Manuscript received May 10, 2012; accepted for publication May 28, 2012
Supporting information is available online at http://www.genetics.org/content/suppl/2012/06/19/genetics.112.141945.DC1.
The primary data sets used in this article have been deposited with dbGAP under accession nos. ARIC, dbGap ID phs000280.v1.p1; CARDIA, dbGap ID phs000285.v1.p1; CFS, dbGap ID phs000284.v1.p1; CHS, dbGap ID phs000287.v1.p1; FHS, dbGap ID phs000282.v2.p5; and MESA, dbGap ID phs000283.v1.p1.

[1] Present address: Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA.

[2] A full list of members is provided as Supporting Information.

[3] Corresponding author: Children's Hospital Boston, CLS 16065, 300 Longwood Ave., Boston, MA 02115. E-mail: joelh@broadinstitute.org

previously shown to be associated with polycystic ovary syndrome (Chen *et al.* 2010).

Finally, as a class, we found an apparent trend toward undertransmission of the derived allele of rare (frequency ≤0.5%) hcSNPs in 1430 family trios. However, through simulations we demonstrated that proper modeling of the rates of genotype missingness and error can account for the observed undertransmission. Our results suggest that rare derived alleles of hcSNPs assayed here do not exert an immediate or strong deleterious effect on survival and illustrate the importance of considering technical artifacts in genetic analysis of rare variants.

Materials and Methods

Identification of hcSNPs and their derived alleles

A total of 816 SNPs within genomic elements with ≥98% conservation genome-wide were included on the IBC chip (Keating *et al.* 2008). These regions were identified as sequences at least ≥200 bp with at least 98% sequence identity within human–mouse–rat, human–mouse–dog, and human–chicken alignments (Derti *et al.* 2006;Chiang *et al.* 2008; Keating *et al.* 2008). For each of these SNPs, we obtained the corresponding allele from chimpanzee (panTro2), orangutan (ponAbe2), and macaque (rheMac2) from the University of Califronia at Santa Cruz (UCSC) Genome Bioinformatics site (http://genome.ucsc.edu). We required that alleles from at least two of the three primate species be available and consistent to designate the primate allele as the ancestral allele and the other human allele as the derived allele. SNPs representing potential multiple mutational events and SNPs with potential strand issues were removed. In total, the derived allele could be definitively assigned for 782 of 816 SNPs in nUCEs.

To narrow our focus to SNPs truly under strong evolutionary constraint, we used the chimpanzee allele as the reference and for each SNP calculated the rate of substitution when the allele was aligned to other species available from the vertebrate 44-way multialignment and conservation track from the UCSC Genome Bioinformatics site (Rhead *et al.* 2010). After removing alignments from species that were used to bioinformatically identify the nUCEs [*i.e.*, human, hg18; mouse, mm9; rat, rn4; dog, canFam2; and chicken, galGal3 (Derti *et al.* 2006; Chiang *et al.* 2008)], a conservation score was assigned to each SNP as 1 – ($N_{\mathrm{SUB}}/N_{\mathrm{TOT}}$), where N_{SUB} is the number of substitution events (relative to the chimpanzee allele) observed from the alignment of all remaining species, while N_{TOT} is the number of species aligned. SNPs with conservation scores >0.9 were defined as hcSNPs, with 577 of 782 SNPs meeting this definition. We note that this approach is not the only approach for estimating the level of evolutionary constraint on each SNP; alternative approaches such as using the phyloP program (Pollard *et al.* 2010) could be just as reasonable. We chose to use a straightforward substitution scheme to be consistent with the definition of the UCE regions we used (Derti *et al.* 2006; Chiang *et al.* 2008), but our conservation scores here are highly correlated with scores produced by phyloP (r = 0.72, data not shown).

Of the 577 hcSNPs, 533 hcSNPs were successfully genotyped and passed quality-control filters in at least one of our study populations, 511 of which were polymorphic in at least one of our study populations (Supporting Information, Table S1) and formed the basis of most of our analyses. Because of the stringent conservation cutoff, the minimum lower bound of the 95% confidence interval around the conservation score for any of these SNPs is still >0.7 (Figure S1).

Association analysis

For all cohorts, cryptically related individuals (for the non-family–based cohorts) and population outliers as determined by principal components analysis were removed from analysis. Unrelated panels [Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk Development in Young Adults (CARDIA), Cardiovascular Health Study (CHS), and Multi-Ethnic Study of Atherosclerosis (MESA)] were analyzed in the following manners. For number of children, we analyzed the discrete trait using Poisson regression, controlling for overdispersion using the generalized linear model in R. Longevity was analyzed as a dichotomous trait by logistic regression using PLINK v.1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/). The remaining traits were analyzed as continuous traits with the linear regression function using PLINK. In all cases, we assumed an additive genetic model and included the top 10 principal components in the analysis as covariates, to control for potential spurious associations due to population stratification (Price *et al.* 2006; McCarthy *et al.* 2008) or other systematic biases that may be associated with the individual genotypes. The principal components were distributed by the CARe consortium. Briefly, the HapMap CEU, YRI, and CHB+JPT populations, also genotyped on the IBC chip, were used as reference populations with the CARe cohorts. One cohort at a time, the principal components were calculated using the smartpca utility implemented in EIGENSTRAT v3.0 (http://genepath.med.harvard.edu/~reich/Software.htm).

In addition, variation among traits related to reproductive fitness is likely to be influenced by additional nongenetic factors such as sociocultural or economic status (Pluzhnikov *et al.* 2007). Therefore, to begin to account for including individuals who may not have realized their true reproductive potential, we have included additional nongenetic factors either as inclusion criteria for the study or as covariates in the association model, where appropriate. These include gender, age, study site, marital status, birth control use, education level, family income, history of hysterectomy or ovariectomy, etc. Please refer to File S1 and Table S2 for more detailed descriptions of phenotype and covariate definitions.

All traits other than the number of children were also analyzed in the related panels [Cleveland Family Study (CFS) and Framingham Heart Study (FHS)], using the GWAF package (Chen and Yang 2010). Longevity, as a dichotomous trait, was analyzed using the generalized estimating equations (GEE) routine in GWAF; the remaining continuous traits (age at natural menopause, age at first child, age at last child, BMI, and height) were analyzed using the linear mixed-effects (LME) routine in GWAF.

We focused our association analyses on the hcSNPs as defined above to specifically test the hypothesis that these highly constrained variants are associated with traits related to reproductive and overall fitness. The genome-wide screening of all ~50K variants on the IBC chip for traits such as age at menopause, BMI, or height will be the subject of other investigations stemming from the CARe consortium. For each trait, association testing was conducted within each cohort before meta-analysis, using the inverse variance weighting (fixed-effect) method implemented in Metal (February 2009 release) (http://www.sph.umich.edu/csg/abecasis/Metal/index.html).

Transmission distortion analysis

We used perfect trio data available in the FHS to test for transmission distortion of autosomal hcSNPs. For each hcSNP, we counted the number of times the derived allele was transmitted to the offspring from heterozygous parents, while keeping track of the parent of transmission origin, and summed the counts over all available trios. In the case of two heterozygous parents and a heterozygous offspring, the single derived allele transmitted was counted as half a transmission each to the paternal and the maternal lineage, as we could not be certain of the parental origin in this case. Multiple siblings were treated as separate trios. To describe the degree of under- or overtransmission, we calculated the average proportion of derived allele transmission as the number of derived allele transmissions divided by the total number of transmissions, averaged over all SNPs with at least one informative transmission. To test for distortion in transmission, the number of derived allele transmissions was compared to that of the ancestral allele transmissions by a paired *t*-test, testing the null hypothesis that the number of derived allele transmissions should equal that of the ancestral allele transmissions. Paternal and maternal transmissions were analyzed separately for parent-of-origin effects and summed for overall transmission effects.

Estimate of the empirical genotype missing rate for homozygous and heterozygous individuals

Differential rates of genotype missingness between homozygous and heterozygous individuals can lead to the appearance of transmission distortions (Hirschhorn and Daly 2005). For instance, a comparatively greater rate of missing genotypes among individuals heterozygous for a rare SNP would preferentially remove trios transmitting the minor allele from the analysis, resulting in an apparent undertransmission of the minor allele.

We sought to estimate and test whether the rate of genotype missingness is the same between the two classes of individuals. We first define T_1 and T_2 as the total number of offspring genotypes from the mating types AA × AA (*i.e.*, both parents are homozygous for the common allele) and AA × AB (*i.e.*, one of the parents is heterozygous), respectively. We also define M_1 and M_2 as the number of observed missing genotypes in the offspring from mating types AA × AA and AA × AB, respectively. Then, the number of missing genotypes among AA offspring is $M_1 + M_1T_2/2T_1$, and the number of missing genotypes among AB offspring is $M_2 - M_1T_2/2T_1$, as half of the offspring from mating types AA × AB will result in an AA offspring. When tabulated over genotypes at 5180 SNPs with minor allele frequency (MAF) ≤ 0.005 from the 1430 trios from the FHS, there were 7,370,925 called genotypes and 3746 missing genotypes for the AA offspring, for a missing genotype rate of 5.1×10^{-4}; there were 18,067 called genotypes and 243 missing genotypes for the AB offspring, for a missing genotype rate of 0.013. The difference is significant ($P < 2 \times 10^{-16}$ by χ^2-test with 1 d.f.). Genotypes at which either or both of the parents are missing were not analyzed. Genotypes for BB offspring were also not analyzed; the estimated missing rate of 0.013 for heterozygous individuals was also applied for individuals homozygous for the rare allele in the simulation (see below).

Simulation to estimate γ and η

In addition to differential rates of genotype missingness, differential rates of genotyping error for homozygous and heterozygous individuals could also lead to the appearance of transmission distortion (Gordon *et al.* 2001; Mitchell *et al.* 2003). For example, a parent with genotype AB erroneously called as AA could result in removing the trio from the analysis due to Mendelian inconsistencies while an offspring with genotype AB erroneously called as AA would result in the apparent overtransmission of the common allele.

To estimate the error rate most consistent with our observed data in the FHS, we adopted a two-parameter error model (Douglas *et al.* 2002): γ, the probability of a homozygous genotype incorrectly called as a heterozygote, and η, the probability of a heterozygous genotype incorrectly called as a homozygote. For each pair of γ- and η-values, we simulated genotypes at 5180 rare variants with the same allele-frequency spectrum as that observed in the FHS, while maintaining the nuclear family structure. After parental genotypes were simulated, each offspring inherits either of the parental alleles with equal probability (*i.e.*, no transmission distortion). After all genotypes were simulated, each genotype was randomly assigned as missing with probability equal to the estimated rate of missing genotype, depending on the simulated genotype. If the genotype was not missing, then the genotype was recoded with probabilities γ and η, depending on the simulated genotype. Under our error model, if an AB genotype was recoded, it was recoded as AA or BB with equal probability.

For each pair of γ and η, 100 simulations were conducted. The distribution of the number of Mendelian errors for each round of simulation was compared to the observed distribution (5117 SNPs with zero Mendelian error, 35 with one error, 18 with two errors, 5 with three errors, and 5 with four or more errors), using Pearson's goodness-of-fit test. The pair of values that produced the lowest median χ^2-test statistics among the 100 simulations was deemed as the pair most consistent with the observed data.

The effect of genotype missingness, genotyping errors, and selection on transmission distortion

To determine the effect of genotype missingness, genotyping errors, and selection on transmission distortion, we simulated genotype data for a perfect trio in the same manner as described above. The simulated genotype data were then subjected to the same transmission distortion analysis also described above. The process was then iterated to obtain a distribution of the average proportion of the derived allele or the minor allele transmission.

To determine the effect on transmission due to technical artifact, we simulated genotype data, assuming the empirically estimated rates of genotype missingness and genotyping error, and assumed no transmission distortion. Genotypes at 5180 SNPs were simulated 100 times each.

To estimate the strength of selection experienced by the rare hcSNPs in our data, we used the same simulation scheme but assumed transmission distortion depending on the selection coefficient, *s*. We assumed AA individuals to have a relative fitness of 1 and AB individuals to have a relative fitness of $1 + s$. Then, the probability of transmitting the B allele is $(1 + s)/(2 + s)$. We tested a range of *s*, from 0.02 to −0.04; for each *s*, we simulated 1000 sets of genotype data at 49 SNPs [the number of hcSNPs with derived allele frequency (DAF) ≤ 0.005 in the FHS] and calculated the likelihood of observing the degree of transmission distortion in the FHS (0.494) given the distribution of the expected proportion of derived allele transmission.

Results

Identifying hcSNPs among variants found in highly conserved regions

A total of 816 SNPs within genomic elements with ≥98% conservation genome-wide [∼94% of all polymorphic SNPs in HapMap phase 2 (International HapMap Consortium et al. 2007)] were included on the IBC chip (Keating *et al.* 2008). Among them, we were able to definitively determine the derived allele on the basis of sequence alignment to the chimpanzee, orangutan, and macaque genomes for 782 SNPs (*Materials and Methods*). As some SNP sites may represent regions of relaxed evolutionary constraint within highly conserved elements, we sought to identify a subset of truly highly constrained SNPs for further analysis. Thus, for each SNP we calculated a conservation score across species, which is defined as the proportion of up to 44 vertebrate orthologous sequences in agreement with the chimpanzee reference allele (*Materials and Methods*). Most of the SNPs found in nUCEs are indeed highly constrained with a conservation score >0.9, although as a class the constraint appears slightly relaxed when compared to the immediately adjacent nucleotide ($P = 1.1 \times 10^{-7}$ by Wilcoxon's signed rank sum test, Figure S2). In total, 533 SNPs had a conservation score >0.9, hereafter referred to as hcSNPs. Notably, 9 hcSNPs have a derived allele frequency >90%, perhaps reflecting variants that offer some form of adaptive advantage (Figure S3); for the purposes of this study, these SNPs were neither singled out nor removed from analysis. In total, 511 hcSNPs were polymorphic in at least one of the CARe cohorts and formed the basis of all subsequent analyses.

Association of hcSNPs with traits related to reproductive and overall fitness

We examined a total of seven phenotypes related to reproductive and overall fitness (number of children, age at first child, age at last child, age at natural menopause, longevity, BMI, and height) in European-American individuals from the CARe cohorts. Number of children was analyzed as a discrete trait, using a Poisson regression. Longevity was analyzed as a dichotomous trait, with cases designated as individuals surviving past the age of 85 and controls designated as individuals not surviving past the age of 75. The remaining traits were analyzed as quantitative traits, using linear regression (see File S1). For each phenotype, up to six of the CARe cohorts were used in our analyses, with the largest sample sizes (and therefore the greatest power) available for height and BMI (Table 1). Each cohort was analyzed separately, assuming the additive genetic model, correcting for nongenetic covariates (*e.g.*, age, sex, oral contraceptive use, education level, etc.) when appropriate. Moreover, since spurious associations could result from population stratification, where the phenotype is correlated with genetic ancestry [such as height in Europe (Campbell *et al.* 2005)], we also included the top 10 principal components as covariates in our analysis to help account for this potential confounding (Price *et al.* 2006; McCarthy *et al.* 2008). Results were then combined via fixed-effect meta-analysis.

Across all phenotypes analyzed, the best evidence of association reached a meta-analysis *P*-value of 2.54×10^{-5} for height (Table 2). However, given that ∼500 SNPs were analyzed for each of seven phenotypes, the Bonferroni threshold for significance would be $\sim 1.4 \times 10^{-5}$. For each trait, we also calculated the proportion of nominally associated SNPs as a measure of excess association. A collective excess of nominal associations could be due to signals individually not strong enough to overcome multiple-hypothesis correction or to uncorrected population substructure or other sources of systematic bias. Under the null hypothesis of no association, ∼5% of the SNPs analyzed should be associated with the phenotype with $P \leq 0.05$. This was true for most of the traits analyzed here, with the exception of height and BMI, where ∼8–10% of the SNPs reached a *P*-value ≤ 0.05 (Table 2). This excess of lower *P*-values is consistent with the

Table 1 Sample sizes for each phenotype analyzed

	ARIC	CARDIA	CFS[a]	CHS	FHS[a]	MESA
Age at natural menopause	2951	NA	NA	1048	1418	841
Number of children	2308	NA	NA	879	NA	702
Age at first child	NA	370	193	907	1304	887
Age at last child	NA	NA	65	595	409	NA
Longevity	NA	NA	NA	1754	804	128
(No. cases)				(1484)	(686)	(97)
BMI	9115	1348	541	3880	7116	2047
Height	9116	1217	504	3707	6934	2286

Shown is the maximum number of samples analyzed for each phenotype in each of the CARe cohorts. The largest sample sizes were available for BMI and height. For age at natural menopause, number of children, age at first child, and age at last child only female participants were analyzed. NA, not available or not analyzed.

[a] Family-based cohorts.

shape of the quantile–quantile plot for each trait as well. For BMI and height only, there are obvious departures from the expectation under the null hypothesis of no association (Figure 1). The lack of obvious associations with the reproductive traits and with longevity most likely reflects the lack of power in our panel and/or the lack of strong effects due to these hcSNPs (Table 1). Additional panels that are many times larger than that analyzed here will be necessary to determine whether hcSNPs are more associated with these traits than random SNPs.

Since one would expect variants with large effects to be rare in the population, especially variants in nUCEs, we also analyzed our data set focusing solely on the 122 variants with an overall DAF ≤ 0.005. To increase power to analyze these extremely rare variants, we scored each individual as either 0 (no derived alleles at any of the 122 variants) or 1 (having at least one rare derived allele). When the presence of one or more rare derived alleles was tested for association with normally distributed traits (BMI, height, age at natural menopause, age at first child, and age at last child), we observed no association (Table S3). We also individually examined samples homozygous for at least one of these rare derived alleles, but none of these samples had extreme phenotypes with respect to any of the normally distributed quantitative traits (Figure S4).

Replication of the top associated hcSNPs for BMI and height

Because of the excess of hcSNPs nominally associated with the anthropometric traits, we examined in more detail the 10 and 9 hcSNPs most strongly associated ($P < 0.01$) with BMI and height, respectively. To distinguish reproducible associations from chance associations among these hcSNPs, we attempted to replicate these findings by looking up their association P-values in data from the GIANT consortium, where a subset of the hcSNPs was either directly genotyped or imputed in >100,000 individuals from >40 cohorts for both BMI and height (we removed cohorts overlapping with CARe). The association results from the GIANT consortium had been corrected for the cohort-specific genomic inflation factor (Devlin and Roeder 1999), λ, before meta-analysis, and then again after meta-analysis to account for any residual stratification (Lango Allen *et al.* 2010; Speliotes *et al.* 2010).

Among the 10 BMI hcSNPs, 4 SNPs reached a one-tailed $P < 0.05$ (Table 3), including 2 SNPs near previously identified loci, *FANCL* and *TFAP2B* (Lindgren *et al.* 2009; Speliotes *et al.* 2010). An additional SNP, rs9463175, just missed the nominal significance threshold (Table 3). Outside of previously identified loci, the hcSNP with the most significant P-value in the replication data set was rs10818872 (one-tailed $P = 0.00518$), which lies in the gene *DENND1A*.

Table 2 Summary of the distribution of association statistics for each phenotype analyzed

	Best two-tailed P	$N_{P<0.05}$	N_{TOT}	$\pi_{P<0.05}$
Age at natural menopause	0.002144	30	436	0.069
Number of children	0.000818	22	365	0.060
Age at first child	0.000973	17	421	0.040
Age at last child	0.001455	20	380	0.053
Longevity	7.32×10^{-5}	31	406	0.076
BMI	0.00035	47	472	0.100
Height	2.54×10^{-5}	39	472	0.083

For each phenotype analyzed, the best two-tailed meta-analyzed P-value, the number of SNPs reaching $P < 0.05$ ($N_{P<0.05}$), total number of hcSNPs analyzed (N_{TOT}), and the proportion of SNPs reaching the nominal significance level ($\pi_{P<0.05}$) are listed. Under the null, ~5% of SNPs analyzed should reach nominal significance. BMI and height showed an excess in the number of hcSNPs nominally associated with the phenotype. For number of children, age at last child, and longevity, the hcSNP must have been analyzed in at least two of the three cohorts to be included in the meta-analysis; for the rest of the traits, the hcSNP must have been analyzed in at least three cohorts to be included in the meta-analysis. For age at natural menopause, number of children, age at first child, and age at last child, we have a strong prior hypothesis that the derived allele of hcSNPs should lower reproductive fitness. Analysis of one-tailed P-values for these traits to reflect the direction of our hypothesis showed no excess of nominally associated SNPs (data not shown).

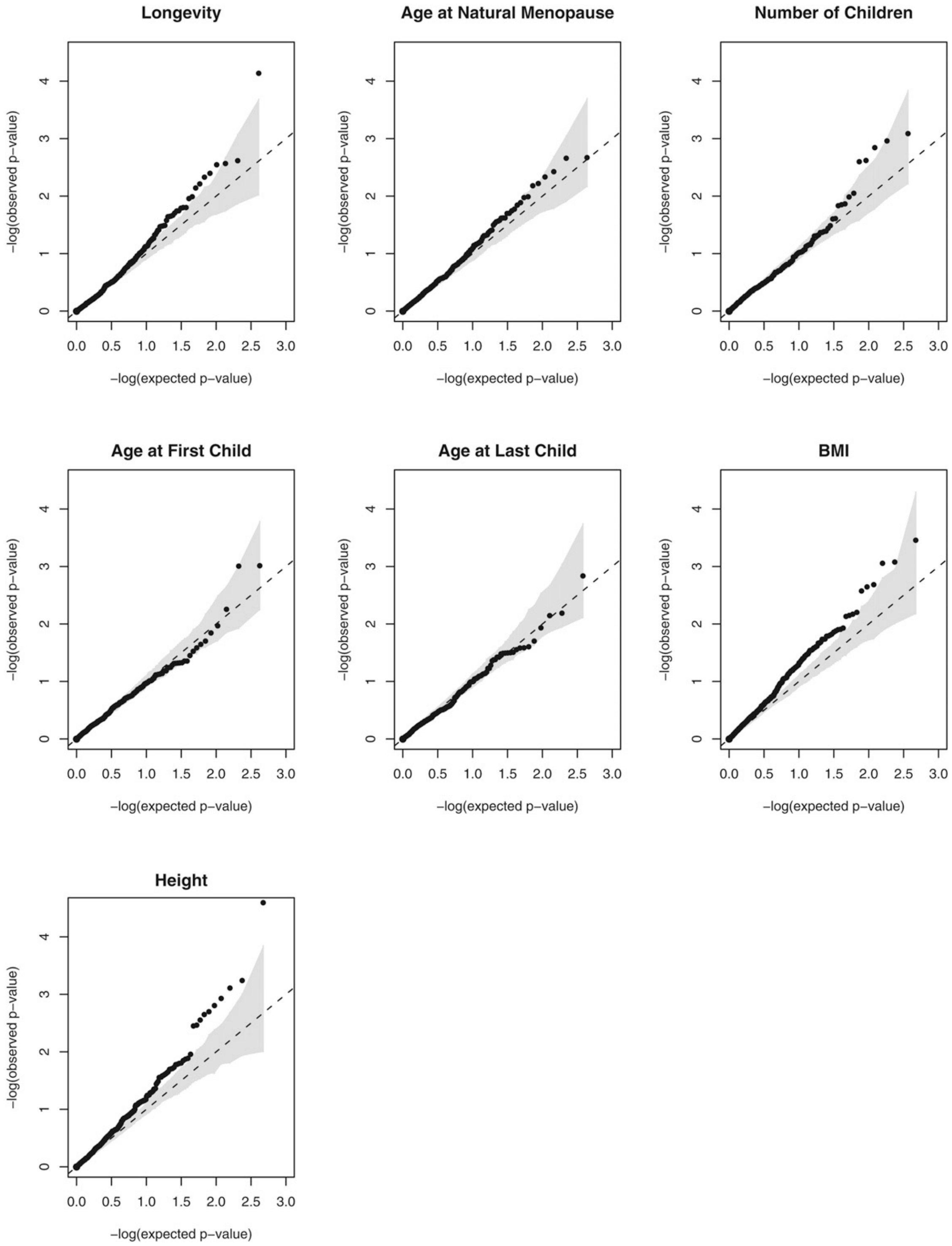

Figure 1 Quantile–quantile plots for all traits analyzed. In all cases, the *P*-values used in the plot are meta-analyzed two-tailed *P*-values. Only hcSNPs found in at least three populations (or two for longevity, number of children, and age at last child) were included. Observed association statistics along the *y*-axis are plotted against the expected association statistics under the null hypothesis of no association along the *x*-axis. Under the null, the points should fall within the shaded area (representing the 95% confidence interval). In general, the evidence of association for most hcSNPs analyzed and for most traits is distributed randomly, as seen by their falling within the 95% confidence interval around the dashed line representing the null distribution. The sole exceptions are the few hcSNPs associated with BMI and with height. Note, however, that the lack of association for most reproductive traits likely reflects the lack of alleles with strong effects and lack of power in our analysis (Table 1).

Although the association to BMI for rs10818872 did not survive Bonferroni correction for testing a total of 19 SNPs between the two traits in replication, a SNP in the same locus and in high linkage disequilibrium (LD) with rs10818872 ($r^2 = 0.826$) was recently shown to be associated with polycystic ovary syndrome (Chen *et al.* 2010), suggesting that variants within nUCEs may indeed affect reproductive fitness.

Of the nine height hcSNPs, four SNPs reached a one-tailed $P < 0.05$ (Table 3), including one SNP near a previously identified locus, *NPPC* (Gudbjartsson *et al.* 2008; Estrada *et al.* 2009; Lango Allen *et al.* 2010). One SNP near another previously identified locus, *PXMP3* (Gudbjartsson *et al.* 2008; Lango Allen *et al.* 2010), just missed the nominal significance threshold. Outside of previously identified loci, the hcSNP with the most significant *P*-value in the replication data set was rs3846984 (one-tailed $P = 0.0117$), an intronic SNP in *AUTS2*.

Given the number of SNPs we examined in replication, we observed an 8.4-fold enrichment of SNPs associated with our traits (8 of 19 SNPs replicated with $P < 0.05$ when 0.95 SNP was expected). These results are unlikely to be affected by population stratification, as the top 10 principal components of ancestry were used as covariates in our association analysis, and the data from the GIANT consortium were conservatively corrected twice using genomic control (Devlin and Roeder 1999). Therefore, the enrichment of nominal replication suggests that the excess of association we observed with BMI and height may be an enrichment of true associations rather than residual uncorrected population stratification. The joint evidence of association for these SNPs is given in Table 3.

Testing for transmission distortion of hcSNPs in family trios

Multiple lines of evidence have shown that ultraconserved elements are under negative selection (Drake *et al.* 2006; Chen *et al.* 2007; Katzman *et al.* 2007; Sakuraba *et al.* 2008), although the strength of the negative selection may be relatively weak and the consequence not immediate, as mice homozygous for a deletion of the UCE showed no overt phenotype (Ahituv *et al.* 2007). We approached the same question from a human genetics perspective: If the effects of the derived alleles of hcSNPs are strongly and immediately deleterious and affecting the viability of offspring, then one should expect to observe an undertransmission of the derived allele from parents to offspring. Conversely, if the effects are weakly deleterious or neutral, then no apparent distortion in transmission should be apparent. However, differences in the rates of genotype missingness and error in individuals with homozygous or heterozygous genotypes are known to cause apparent transmission distortions if not properly modeled (Gordon *et al.* 2001; Mitchell *et al.* 2003; Hirschhorn and Daly 2005). Therefore, such technical confounders must be carefully considered when analyzing transmission distortion using genotype data.

Using the 1430 family trios from one of the CARe cohorts, the FHS, we tested for possible distortion in transmission from heterozygous parents to offspring. A total of 440 autosomal hcSNPs showed at least one informative transmission among the trios examined. Overall, the average difference in the transmission of the derived allele to the ancestral allele was consistent with undertransmission of the derived allele, but the difference was not significant (proportion of derived allele transmission = 49.8%; average derived allele to ancestral allele transmission difference = -0.70 per SNP, $P = 0.295$ by one-tailed paired *t*-test; Figure 2A). As the rarest hcSNPs are more likely to be functionally important, we also examined the 43 hcSNPs with DAF $\leq$ 0.005 in transmission analysis. As a class, these hcSNPs did appear to be more undertransmitted than their common counterparts, although the difference was still not significant (proportion of derived allele transmission = 49.4%; average difference = -0.37 per SNP, $P = 0.214$; Figure 2B).

When we stratified our analysis by parent of origin (Figure S5) or by the degree of conservation of the genomic elements (Figure S6), the overall pattern remained the same: The derived alleles display a nonsignificant trend toward undertransmission, although hcSNPs found in elements conserved at 100% showed a nominally significant undertransmission (Figure S6). The general lack of statistical significance could be a result of small sample size. However, we show below that the apparent transmission distortion could arise due to technical artifacts and that the transmission of rare derived alleles of hcSNPs is consistent with no transmission distortion, as would be predicted if the derived alleles of the hcSNPs do not currently exert strong deleterious effects.

Technical explanation of the apparent undertransmissions of rare minor alleles

We noted that the minor alleles of all other rare SNPs (MAF $\leq$ 0.005) on the IBC chip are also significantly undertransmitted as a class, with a similar degree of undertransmission ($N = 5180$, of which 4267 had at least one informative transmission; proportion of minor allele transmission = 49.1%; average difference = -0.13 per SNP, $P = 0.002$). These results potentially suggest that the rare minor alleles as a whole are enriched for alleles of functional consequences. However, apparent transmission distortion could also occur if the rates of genotype missingness and/or the rates of genotyping error are different between homozygous and heterozygous genotypes in our data (Gordon *et al.* 2001; Mitchell *et al.* 2003; Hirschhorn and Daly 2005). Since these technical issues are likely more pronounced for SNPs with very low minor allele frequency (Korn *et al.* 2008), they could potentially be responsible for the undertransmission of the rare derived or minor alleles in our analysis. Therefore, to determine whether the apparent undertransmission of the rare minor alleles is reflective of true functional consequences or is due to technical artifacts, we estimated, empirically and through simulation, the

Table 3 ***In silico*** **replication of hcSNPs associated with BMI and height, using data from the GIANT consortium**

			CARe			GIANT				
rs ID	Chr	DA	DAF	Effect	*P*	DAF	Effect	One-tailed *P*[b]	Joint *P*	Genes nearby
BMI										
rs6575421	14	C	0.519	−0.0331	3.50×10^{-4}	0.514	0.0012	0.592	0.198	NOVA1, C14orf23
rs1159609[a]	2	G	0.427	−0.0312	8.38×10^{-4}	0.4336	−0.0096	0.0294	8.42×10^{-4}	FANCL, VRK2, BCL11A, PAPOLG, REL
rs9463175	6	T	0.358	−0.0322	8.83×10^{-4}	0.349	−0.0084	0.0554	0.00259	SLC35B3, OFCC1
rs2669892	3	C	0.219	−0.0348	0.00207	0.212	−0.0063	0.148	0.0193	ZBTB20, TIGIT, GAP43, LSAMP
rs10818872	9	C	0.041	0.0734	0.00226	0.042	0.0331	0.00518	9.66×10^{-5}	DENND1A, CRB2, LHX2
rs2807310	9	A	0.186	−0.0357	0.00266	0.188	−0.0024	0.356	0.102	TLE4, CHCHD9, TLE1
rs12469063	2	G	0.242	0.0295	0.00629	0.236	−0.002	0.632	0.411	MEIS1, SPRED2, ACTR2, ETAA1
rs1430331	10	T	0.250	0.0289	0.00675	0.262	0.0111	0.0284	0.00197	C10orf11, ZNF503, SPA17P1, KCNMA1
rs2272903[a]	6	A	0.109	−0.0403	0.00710	0.119	−0.0232	0.00206	5.39×10^{-5}	TFAP2B, RPS17P5, TFAP2D, PKHD1
rs9675567	18	T	0.595	−0.0253	0.00739	0.607	−0.0008	0.439	0.194	TCF4, TXNL1, WDR7
Height										
rs2248913[a]	8	C	0.266	0.0434	2.54×10^{-5}	0.2649	0.0078	0.0718	0.00349	ZFHX4, HNF4G, PXMP3
rs17409319	2	T	0.089	0.0562	5.76×10^{-4}	0.102	0.004	0.272	0.116	ZEB2, ACVR2A, ORC4L, MBD5
rs9853631	3	C	0.471	0.0306	7.77×10^{-4}	0.456	0.005	0.109	0.0270	IL20RB, SOX14
rs10496382	2	T	0.062	−0.0612	0.00118	0.061	0.0091	0.848	0.904	TMEM182, MFSD9, SLC9A2, SLC9A4, POU3F3
rs4237770	11	C	0.481	−0.0287	0.00157	0.511	−0.0086	0.0255	0.00418	LMO1, RIC3, STK33
rs7648568	3	T	0.171	−0.0373	0.00200	0.175	−0.0096	0.0356	0.00813	IL20RB, SOX14
rs3103295[a]	2	C	0.633	0.0286	0.00225	0.65	0.0169	0.000241	2.49×10^{-6}	DIS3L2, NPPC, ALPP, ECEL1P2
rs3846984	7	C	0.733	0.0297	0.00344	0.733	0.0102	0.0117	0.00239	AUTS2, STAG3L4, TYW1, SBDS, C7orf42, RABGEF1, KCTD7, WBSCR17
rs763853	12	G	0.112	0.0421	0.00356	0.113	−0.0023	0.636	0.537	ETNK1, SOX5

The top associated hcSNPs ($P < 0.01$ in CARe), representing potential novel loci for BMI and height, were examined *in silico* using data from the GIANT consortium. For each SNP, the derived allele (DA), the derived allele frequency (DAF) in CARe and in GIANT, and the effect of the derived allele (Effect) in both studies are given. The *P*-value in GIANT is given as a one-tailed *P*-value to incorporate the direction of effect. The joint-analysis *P*-value (Joint *P*) reports the *P*-value after meta-analyzing CARe and GIANT. Meta-analysis was performed using the standard error method correcting for genomic control inflation factor, λ (λ = 1.039 and 1.323 for BMI and height, respectively, in CARe, as determined by testing ~5100 non-hcSNPs matched by allele frequency). Chr, chromosome.

[a] SNPs near previously identified loci.

[b] *P*-value after applying two rounds of genomic control.

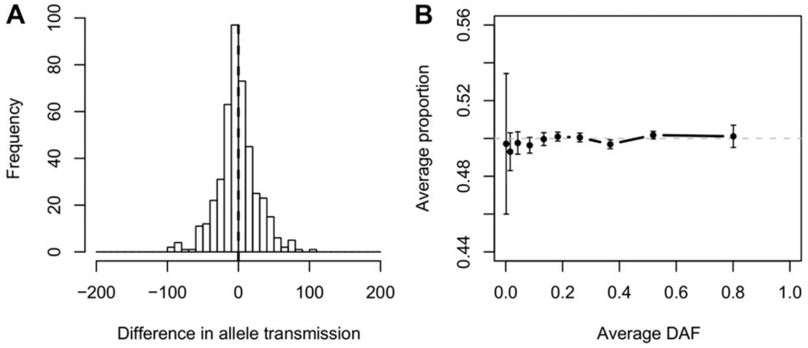

Figure 2 Transmission distortion analysis using trio data from the FHS. (A) Distortion in the transmission of the derived alleles of hcSNPs from heterozygous parents to offspring was measured by subtracting the ancestral allele transmissions from the derived allele transmission for each SNP. Under the null hypothesis of no distortion in transmission, the mean difference should be 0. In the FHS, we observed that the derived allele on average trends nonsignificantly toward undertransmission: Average proportion of derived allele transmission (derived allele transmissions/total transmissions) = 49.8%, mean difference = −0.70 per SNP, $P = 0.295$. (B) When SNPs were stratified into deciles by derived allele frequency (DAF), the average proportion of transmission appears to be <50% for SNPs with lower DAFs, but remains insignificant (mean difference = −0.386 per SNP, $P = 0.201$ for SNPs with DAF ≤ 0.005). Error bars reflect the standard error in the proportion of derived allele transmissions within the decile.

rate of missing genotypes and genotyping error. We then examined whether the observed rate of missingness and errors could produce transmission data consistent with the pattern we observed in the FHS.

To test whether differential rates of genotype missingness are evident in our data, we compared rates of missingness in offspring from different parental mating types (for example, if AB heterozygotes are harder to genotype than AA homozygotes, then AA × AB parents will have higher rates of missingness in their offspring than will AA × AA parents; *Materials and Methods*). We found that for SNPs with MAF ≤ 0.005, offspring homozygous for the major allele are significantly less likely to be called missing than heterozygous offspring on average (5.1×10^{-4} *vs.* 0.013, $P < 2 \times 10^{-16}$), suggesting that the apparent transmission distortion in our data could at least be partly explained by different rates of genotype missingness between homozygous and heterozygous individuals.

To estimate the rate of genotyping error in the FHS pedigrees, we simulated nuclear family pedigrees with different error models while maintaining the FHS pedigree structure, the allele frequency spectrum, and the estimated rates of missing genotypes (*Materials and Methods*). We adopted an error model with two parameters (Douglas *et al.* 2002): γ, the probability of a homozygous genotype incorrectly called as a heterozygote, and η, the probability of a heterozygous genotype incorrectly called as a homozygote. In a grid-like search over possible values of γ and η, we conducted 100 simulations for each pair of values and computed Pearson's goodness-of-fit test statistics, comparing the distribution of the simulated number of Mendelian errors to that observed in the FHS (*Materials and Methods*). For SNPs with MAF ≤ 0.005, the error model that best fitted our observed data is $\gamma = 1 \times 10^{-6}$ and $\eta = 1.5 \times 10^{-3}$ (Figure 3), suggesting that the apparent transmission distortion in our data could at least partly be explained by differential rates of genotyping errors between homozygous and heterozygous genotypes.

To assess the degree to which the estimated parameters of genotype missingness and genotyping error could explain the transmission distortion we have observed, we used the same scheme to simulate family trio data, but assumed no transmission distortion for either allele. Assuming no genotyping errors, simulation showed that our estimated differential rates of missing genotype alone would account for ~48% of the observed undertransmission of SNPs with MAF ≤ 0.005. Our estimated differential rates of genotyping error would then account for another ~10% of the observed undertransmissions (Table 4). While a small excess of undertransmission of the minor allele remains after accounting for genotype missingness and error, the remaining undertransmission is not significantly different from the null hypothesis of no transmission distortion, as in all models tested

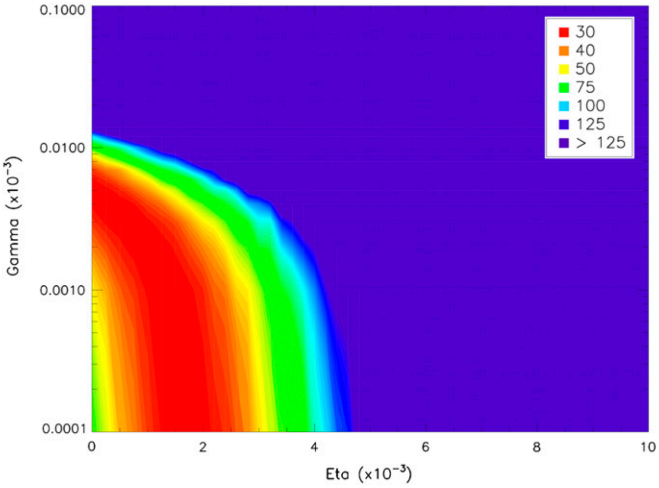

Figure 3 Contour map summarizing the simulation results at different values of γ and η. At each pair of values of γ (the probability of a homozygous genotype incorrectly called as a heterozygote) and η (the probability of a heterozygous genotype incorrectly called as a homozygote), genotypes at 5180 sites were simulated for 1430 trios while incorporating the genotyping error rate and the estimated rate of missing genotype and maintaining the nuclear family structure and allele-frequency spectrum as observed in the FHS (*Materials and Methods*). One hundred simulations were conducted at each pair of γ and η, and the fit was assessed by Pearson's goodness-of-fit test according to the distribution of Mendelian errors among simulated SNPs and that of the observed data in the FHS. The median of Pearson's χ^2-statistics among the 100 simulations is plotted as the contour; the minimum (representing the best fit) occurred at $(\gamma, \eta) = (1 \times 10^{-6}, 1.5 \times 10^{-3})$ and $(0, 1.75 \times 10^{-3})$.

>5% of the simulated pedigrees would show an equal or greater extent of undertransmission than that observed in the FHS (Table 4).

Given the estimated rates of missing genotypes and of genotyping error for rare SNPs on the IBC chip, we also estimated the strength of selection on the rare hcSNPs most consistent with our observed data. We simulated family trio data for genotypes at 49 rare hcSNPs (DAF ≤ 0.005), but assumed differing degrees of transmission distortion corresponding to different levels of the strength of selection experienced by the rare derived alleles (see *Materials and Methods*). Using this approach, the maximum-likelihood estimate of the selection coefficient was −0.004 (Figure S7). However, this estimate should be considered preliminary as the range of the estimated selection coefficients was not significantly different from zero due to the small number of SNPs available for the analysis.

In summary, we showed that the apparent undertransmission of derived alleles of rare hcSNPs and minor alleles of rare SNPs in general can be largely explained by the difficulties in genotyping rare SNPs. Consistent with the observation that mice homozygous for deletion of ultraconserved elements appeared to be phenotypically normal (Ahituv *et al.* 2007), our observation suggests that any effects of the derived allele may not be immediately and largely deleterious in the next generation to cause strong transmission distortion. Moreover, our analysis demonstrates the need to carefully model the differential rates of genotype missingness and error when analyzing genotyping data of SNPs with extremely low minor allele frequency.

Discussion

The origin and evolution of the ultra- and near-ultraconserved elements have presented a mystery. Their extreme conservation reflects functional constraints, rather than mutational coldspots (Drake *et al.* 2006; Chen *et al.* 2007; Katzman *et al.* 2007; Sakuraba *et al.* 2008). Yet, paradoxically, these elements have produced no detectable phenotypes when deleted in mice (Ahituv *et al.* 2007). In the present study, we conducted, to our knowledge, the first large-scale genome-wide screen involving SNPs in extremely highly conserved elements spanning across the genome. We observed an excess of hcSNPs nominally associated with BMI and height. Using data based on >100,000 additional individuals, we replicated 8 of the top 19 SNPs at a nominal significance level, when <1 such SNP is expected. The top novel association was an intronic hcSNP in *DENND1A* for association with BMI (joint $P = 9.66 \times 10^{-5}$, Table 3). Interestingly, a SNP in the same locus and in high LD with the hcSNP assayed here was recently shown to be associated with polycystic ovary syndrome (Chen *et al.* 2010). *DENND1A* encodes a domain that can bind to ERAP1, whose serum level has been reported to be associated with polycystic ovary syndrome accompanied by obesity (Del Villar and Miller 2004; Olszanecka-Glinianowicz *et al.* 2007). This result suggests that we may be detecting a signal due to a trait that is correlated with BMI and that variants within nUCEs may indeed result in phenotypes with reproductive consequences.

Table 4 Simulated results of the effect of differential rate of genotype missingness and genotyping error on the transmission of the minor allele

γ ($\times 10^{-3}$)	η ($\times 10^{-3}$)	χ^2	Mean	P_5	P_{95}
0	0	NA	0.4957	0.4893	0.5019
0	1.75	28.4	0.4946	0.4870	0.5012
0.001	1.5	28.4	0.4948	0.4894	0.5005
0.002	1.25	28.6	0.4952	0.4886	0.5020
0.003	0.75	28.6	0.4944	0.4866	0.5017
0.004	0.25	28.6	0.4938	0.4876	0.4999
0.005	0	28.8	0.4936	0.4880	0.4999

At each pair of γ and η, 100 simulations were conducted in which the genotypes for 5180 rare SNPs (DAF ≤ 0.005) were simulated for 1430 trios, assuming 50% transmission probability for the minor allele. In all cases, a background genotype missing rate of 5.1×10^{-4} for individuals homozygous for the major allele and of 0.013 otherwise was set. χ^2, Pearson's goodness-of-fit statistics for the 100 simulations when compared to the observed data in the FHS; Mean, the average proportion of transmission of the minor allele across 100 simulations; P_5 and P_{95}, the 5th and 95th percentiles of the transmission proportion among the 100 simulations, respectively. NA, not applicable. When $\gamma = \eta = 0$, only the effect of differential missing genotypes on transmission is considered. For comparison, the proportion of minor allele transmission observed in the FHS is 0.4908. A number of other pairs of γ- and η-values near the best-fit values are also given for comparison.

Aside from BMI and height, the other traits examined in our analysis generally showed a null distribution of the association statistics (Figure 1, Table 2). These traits include reproductive traits, which have generally been shown to be heritable in human populations (Peccei 2001; Pluzhnikov *et al.* 2007; Kosova *et al.* 2010a). A combination of a lack of large (or any) effects due to the hcSNPs and reduced power due to smaller sample size or uncontrolled nongenetic factors for these traits may explain our null results. Potentially, one could include measures of male fertility to increase sample size [*e.g.*, measure birth rate as number of births per year of marriage (Kosova *et al.* 2010b)], although in our case we do not have detailed information regarding the number of years of marriage. Note that our results should not be interpreted as variants in highly conserved elements having no effect on reproductive fitness. In fact, the hcSNP in *DENND1A* reported to be associated with polycystic ovary syndrome is consistent with hcSNPs affecting reproductive fitness. Therefore, as larger data sets for a variety of reproductive phenotypes become available, consideration of conservation as presented here could become a valuable approach for data mining.

Interestingly, if we relaxed the conservation score threshold slightly for identifying hcSNPs, then two unrelated individuals homozygous for a rare SNP (rs10493810, conservation score = 0.82) in a highly conserved region would now be 1.6 and 4.6 standard deviations above the population mean for age at natural menopause, respectively. Another individual homozygous for a different rare SNP (rs4664775, conservation score = 0.72) would now be 5.4 standard deviations above the population mean for the same trait. Because of the increased juvenile dependence specific to humans, it has been suggested that prolonged postmenopausal life span increases reproductive fitness in women by allowing their children or grandchildren to reproduce more frequently and successfully (Peccei 2001; Hawkes 2004; Lahdenpera *et al.* 2004). Therefore, our observations here could be consistent with the expectation of rare derived alleles of hcSNPs to lower reproductive fitness by increasing age of natural menopause.

Our analyses have demonstrated the potential utility of incorporating information from evolutionary conservation into association studies. However, only if the hcSNPs were themselves causal would it be fruitful to systematically screen highly constrained SNPs for association with phenotypes. Therefore, it is important to demonstrate that the hcSNPs tested here (or another highly constrained SNP in LD) are causal for the association signal. Preliminarily, conditional analysis using ~100,000 European-derived individuals in the GIANT consortium (Lango Allen *et al.* 2010; Speliotes *et al.* 2010) showed that three of the four hcSNP associations in previously identified loci (footnote *a* in Table 3) remain nominally significant after controlling for the effects of known variants influencing BMI and height (data not shown). This finding supports the notion that hcSNPs are independent from known signals and might be themselves causal.

One potential drawback of our study is the focus on relatively common variation within the highly conserved regions. The derived allele-frequency spectrum for hcSNPs is likely biased toward low frequencies (Drake *et al.* 2006; Katzman *et al.* 2007) and one would expect variants with important functional consequences, particularly within nearly ultraconserved regions, to be extremely rare in the population. The selection of variants within nUCEs in our analysis was based on polymorphisms in phase 2 of HapMap (Keating *et al.* 2008), which is biased toward common variation. Therefore, even though ~25% of the hcSNPs examined here have a derived allele frequency <0.005 in the CARe cohorts, resequencing of the highly conserved regions to generate a complete catalog of rare variations could present a different picture.

Finally, through the analysis of transmission distortion using the trio data in our cohort, we ruled out a large deleterious effect for the derived alleles of the hcSNPs as a class (Figure 2A). Our preliminary maximum-likelihood estimate of the average selective coefficient, s, is -0.004 for these rare hcSNPs. The range of our estimate is large and not significantly different from zero (Figure S7); more robust analysis will be needed when additional rare variants in the nUCEs are studied in aggregate. This estimate is significantly larger than previous estimates of the strength of selection [on the order of 10^{-4} (Chen *et al.* 2007; Katzman *et al.* 2007)], although previous studies focused on all variation within the UCEs and we focused only on the rarest hcSNPs detected among thousands of individuals, for which the strength of selection is expected to be greater. Our observation that carriers of rare hcSNPs showed no significant immediate survival disadvantage is thus consistent with mouse knockout models for homozygous deletion of UCEs (Ahituv *et al.* 2007).

Just like the association analysis, our transmission analysis here may be influenced by not having sampled all of the rare variants in the highly conserved regions. Moreover, while the genotype data from CARe are of extremely high quality in general (Lettre *et al.* 2011; Lo *et al.* 2011), genotype calls at rare SNPs displayed differences in the rates of genotype missingness and error between homozygous and heterozygous genotypes (Figure 3, Table 4). Therefore, our observation illustrates that the potential for confounding due to technical artifacts when analyzing extremely rare SNPs is important to consider. With genotyping by sequencing, we may be able to avoid differential genotype calling errors and missingness, at least for regions of high depth of coverage, and draw more robust conclusions concerning the undertransmission of the rare derived alleles and the strength of selection acting on them.

Our results have implications for both studying the function of ultraconserved elements and conducting human association studies. We have shown that the rare derived alleles of highly constrained SNPs are not immediately or strongly deleterious, consistent with previous mouse knockout models (Ahituv *et al.* 2007). Moreover, because current genome-wide association studies are burdened with the large number of hypotheses being tested simultaneously, analytical methods to integrate additional historical, demographic, evolutionary, biological, or functional information when mining through GWA data could be important in interpreting association results while maintaining a stringent standard. Indeed, attempts to incorporate expression data have been successful in identifying novel candidate obesity loci (Naukkarinen *et al.* 2010). Here, we present an initial attempt to leverage information from an evolutionary standpoint, but it remains to be seen whether focusing on variants in highly conserved regions in a more sophisticated statistical framework will be a generally fruitful approach.

Acknowledgments

The authors thank J. Aylor, N. Dermond, R. Goodloe, C. Johnson,, E. Larkin, A. Newman, W. Post, S. Redline, P. Schreiner, N. Zakai, and X. Zhu for assistance with phenotype definition within different CARe cohorts; members of the Candidate Gene Association Resource (CARe) Anthropometric and Aging working groups, past and present members of the Hirschhorn laboratory, and C.-t. Wu for ideas and comments; Vivian U. for assistance in the graphical representation in Figure 3; and D. Farlow, Y. Meng, C. D. Palmer, and T. Young for assistance with analysis within the CARe framework. The authors thank the staff and participants of the Atherosclerosis Risk in Communities (ARIC) study for their important contributions. This work was supported by a graduate research fellowship from the National Science Foundation (to C.W.K. C.). The CARe acknowledges the support of the National Heart, Lung, and Blood Institute (NHLBI) and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research (NHLBI contract no. HHSN268200960009C). The following eight parent studies have contributed parent study data, ancillary study data, and DNA samples through the Broad Institute (N01-HC-65226) to create this genotype/phenotype database for wide dissemination to the biomedical research community: the ARIC study, the Cardiovascular Health Study (CHS), the Cleveland Family Study (CFS), the Coronary Artery Risk Development in Young Adults (CARDIA) study, the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Sleep Heart Health Study (SHHS). Individual study funding attributions can be obtained at http://public.nhlbi.nih.gov/GeneticsGenomics/home/care.aspx. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by NHLBI contracts HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C.

Literature Cited

Ahituv, N., Y. Zhu, A. Visel, A. Holt, V. Afzal *et al.*, 2007 Deletion of ultraconserved elements yields viable mice. PLoS Biol. 5: e234.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent *et al.*, 2004 Ultraconserved elements in the human genome. Science 304: 1321–1325.

Campbell, C. D., E. L. Ogburn, K. L. Lunetta, H. N. Lyon, M. L. Freedman *et al.*, 2005 Demonstrating stratification in a European American population. Nat. Genet. 37: 868–872.

Chen, C. T., J. C. Wang, and B. A. Cohen, 2007 The strength of selection on ultraconserved elements in the human genome. Am. J. Hum. Genet. 80: 692–704.

Chen, M. H., and Q. Yang, 2010 GWAF: an R package for genome-wide association analyses with family data. Bioinformatics 26: 580–581.

Chen, Z. J., H. Zhao, L. He, Y. Shi, Y. Qin *et al.*, 2010 Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. Nat. Genet. 43: 55–59.

Chiang, C. W., A. Derti, D. Schwartz, M. F. Chou, J. N. Hirschhorn *et al.*, 2008 Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. Genetics 180: 2277–2293.

Cooper, G. M., D. L. Goode, S. B. Ng, A. Sidow, M. J. Bamshad *et al.*, 2010 Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat. Methods 7: 250–251.

Del Villar, K., and C. A. Miller, 2004 Down-regulation of DENN/MADD, a TNF receptor binding protein, correlates with neuronal cell death in Alzheimer's disease brain and hippocampal neurons. Proc. Natl. Acad. Sci. USA 101: 4210–4215.

Derti, A., F. P. Roth, G. M. Church, and C. T. Wu, 2006 Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. Nat. Genet. 38: 1216–1220.

Devlin, B., and K. Roeder, 1999 Genomic control for association studies. Biometrics 55: 997–1004.

Douglas, J. A., A. D. Skol, and M. Boehnke, 2002 Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. Am. J. Hum. Genet. 70: 487–495.

Drake, J. A., C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh *et al.*, 2006 Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. 38: 223–227.

Emison, E. S., A. S. McCallion, C. S. Kashuk, R. T. Bush, E. Grice *et al.*, 2005 A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434: 857–863.

Estrada, K., M. Krawczak, S. Schreiber, K. van Duijn, L. Stolk *et al.*, 2009 A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. Hum. Mol. Genet. 18: 3516–3524.

Gordon, D., S. C. Heath, X. Liu, and J. Ott, 2001 A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am. J. Hum. Genet. 69: 371–380.

Gudbjartsson, D. F., G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson *et al.*, 2008 Many sequence variants affecting diversity of adult human height. Nat. Genet. 40: 609–615.

Hawkes, K., 2004 Human longevity: the grandmother effect. Nature 428: 128–129.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106: 9362–9367.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95–108.

International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Katzman, S., A. D. Kern, G. Bejerano, G. Fewell, L. Fulton *et al.*, 2007 Human genome ultraconserved elements are ultraselected. Science 317: 915.

Keating, B. J., S. Tischfield, S. S. Murray, T. Bhangale, T. S. Price *et al.*, 2008 Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS ONE 3: e3583.

Korn, J. M., F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh *et al.*, 2008 Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. 40: 1253–1260.

Kosova, G., M. Abney, and C. Ober, 2010a Colloquium papers: heritability of reproductive fitness traits in a human population. Proc. Natl. Acad. Sci. USA 107(Suppl. 1): 1772–1778.

Kosova, G., J. K. Pickrell, J. L. Kelley, P. F. McArdle, A. R. Shuldiner *et al.*, 2010b The CFTR Met 470 allele is associated with lower birth rates in fertile men from a population isolate. PLoS Genet. 6: e1000974.

Lahdenpera, M., V. Lummaa, S. Helle, M. Tremblay, and A. F. Russell, 2004 Fitness benefits of prolonged post-reproductive lifespan in women. Nature 428: 178–181.

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.

Lettre, G., C. D. Palmer, T. Young, K. G. Ejebe, H. Allayee *et al.*, 2011 Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe project. PLoS Genet. 7: e1001300.

Lindgren, C. M., I. M. Heid, J. C. Randall, C. Lamina, V. Steinthorsdottir *et al.*, 2009 Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet. 5: e1000508.

Lo, K. S., J. G. Wilson, L. A. Lange, A. R. Folsom, G. Galarneau *et al.*, 2011 Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. Hum. Genet. 129: 307–317.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9: 356–369.

McEwen, G. K., A. Woolfe, D. Goode, T. Vavouri, H. Callaway *et al.*, 2006 Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res. 16: 451–465.

Mitchell, A. A., D. J. Cutler, and A. Chakravarti, 2003 Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am. J. Hum. Genet. 72: 598–610.

Musunuru, K., G. Lettre, T. Young, D. N. Farlow, J. P. Pirruccello *et al.*, 2010 Candidate gene association resource (CARe): design, methods, and proof of concept. Circ. Cardiovasc. Genet. 3: 267–275.

Naukkarinen, J., I. Surakka, K. H. Pietilainen, A. Rissanen, V. Salomaa *et al.*, 2010 Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes. PLoS Genet. 6: e1000976.

Olszanecka-Glinianowicz, M., M. Banas, B. Zahorska-Markiewicz, J. Janowska, P. Kocelak *et al.*, 2007 Is the polycystic ovary syndrome associated with chronic inflammation per se? Eur. J. Obstet. Gynecol. Reprod. Biol. 133: 197–202.

Paparidis, Z., A. A. Abbasi, S. Malik, D. K. Goode, H. Callaway *et al.*, 2007 Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. Dev. Growth Differ. 49: 543–553.

Peccei, J. S., 2001 Menopause: Adaptation or epiphenomenon? Evol. Anthropol. 10: 43–57.

Pennacchio, L. A., N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega *et al.*, 2006 In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502.

Pluzhnikov, A., D. K. Nolan, Z. Tan, M. S. McPeek, and C. Ober, 2007 Correlation of intergenerational family sizes suggests a genetic component of reproductive fitness. Am. J. Hum. Genet. 81: 165–169.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, 2010 Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20: 110–121.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Rhead, B., D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig *et al.*, 2010 The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 38: D613–D619.

Sakuraba, Y., T. Kimura, H. Masuya, H. Noguchi, H. Sezutsu *et al.*, 2008 Identification and characterization of new long conserved noncoding sequences in vertebrates. Mamm. Genome 19: 703–712.

Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson *et al.*, 2010 Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. 42: 937–948.

Stephen, S., M. Pheasant, I. V. Makunin, and J. S. Mattick, 2008 Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. 25: 402–408.

Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis *et al.*, 2008 Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat. Genet. 40: 158–160.

Woolfe, A., M. Goodson, D. K. Goode, P. Snell, G. K. McEwen *et al.*, 2005 Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3: e7.

Communicating editor: L. M. McIntyre

GENETICS

Supporting Information
http://www.genetics.org/content/suppl/2012/06/19/genetics.112.141945.DC1

Ultraconserved Elements in the Human Genome: Association and Transmission Analyses of Highly Constrained Single-Nucleotide Polymorphisms

Charleston W. K. Chiang, Ching-Ti Liu, Guillaume Lettre, Leslie A. Lange, Neal W. Jorgensen, Brendan J. Keating, Sailaja Vedantam, Nora L. Nock, Nora Franceschini, Alex P. Reiner, Ellen W. Demerath, Eric Boerwinkle, Jerome I. Rotter, James G. Wilson, Kari E. North, George J. Papanicolaou, L. Adrienne Cupples, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Joanne M. Murabito, and Joel N. Hirschhorn

DOI: 10.1534/genetics.112.141945

File S1

Supporting Methods

Study population

The Candidate Gene Association Resource (CARe) established by the National Heart, Lung, and Blood Institute (NHLBI) is composed of > 40,000 individuals representing 4 ethnic groups in 9 community-based cohorts: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults study (CARDIA), the Cleveland Family Study (CFS), the Cardiovascular Health Study (CHS), the Cooperative Study of Sickle Cell Disease (CSSCD), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Sleep Heart Health Study (SHHS) (Lettre *et al.* 2011; Musunuru *et al.* 2010). The six non-patient based cohorts with self-identified European American participants (ARIC, CARDIA, CFS, CHS, FHS, MESA) were used in the present study. Individuals were genotyped with the IBC chip and processed through quality control measures by CARe, as described elsewhere (Lo *et al.* 2011). In general, duplicate individuals, population outliers as determined by principal component analysis, cryptically related individuals in unrelated panels, and poorly genotyped individuals were removed from further analysis.

Phenotype definition

The reproductive traits examined here include age at natural menopause, number of children, age at first child, and age at last child. Descriptive statistics for each of the reproductive traits can be found in **Table S2**.

Age at natural menopause was treated as a continuous variable, defined as the age at the last menstrual period, after at least 12 consecutive months of amenorrhea. For the purpose of this study, we only included women with age at natural menopause between 40 and 60 years of age and excluded individuals whose menopause was induced by irradiation, hysterectomy and/or bilateral ovariectomy, or who used hormone replacement therapy before menopause. Residuals for each study were created separately by regressing age at natural menopause by cohort or study center where applicable.

For number of children, female participants who had not yet reached natural menopause or who had hysterectomy or oophrectomy before age 45 were removed. For the remaining women, the number of children was counted for each woman based on the self-reported number of live births. Number of children was analyzed as categorical variable using Poisson regression. Covariates included in the model were age, marital status, birth control use, education level, family income, study site, and the first 10 principal components. As the association statistics appeared to be unstable for low minor allele frequency (MAF) SNPs, we further restricted our analysis to SNPs with MAF > 0.02.

For age at last child, female participants who had not yet reached natural menopause or who had hysterectomy or oophrectomy before age 45 were also removed. If age at last term pregnancy was not reported, the age at last child was calculated based on the difference of the mother's age and the age of the youngest child.

For age at first child, no exclusion based on menopausal status or surgery was applied. If age at first term pregnancy was not reported, the age at first child was calculated based on the difference of the mother's age and the age of the oldest child.

Note that as it is difficult to differentiate biological children from step or adopted children in self-reported surveys, no such distinctions were stringently made but information on biological children were always used if available (FHS and MESA for age at first child; FHS for age at last child).

Age at first and last child were analyzed as continuous variables. Age at first child was log-transformed. Stratified by cohort, multivariate linear regression models were constructed for both phenotypes controlling for birth control pill use, education status, marital status, study site, income status, and age, where applicable. The residuals were normalized to a standard normal distribution. The associations of hcSNPs to these two traits were tested using linear regression, with the top 10 principal components as additional covariates.

The overall fitness traits examined here include longevity, BMI and height.

For longevity, cases (long-lived individuals) were defined as individuals surviving to ≥ 85 years of age at time of death or last contact, and controls (short-lived individuals) were defined as individuals who died ≤ 75 years of age. Longevity was analyzed as a dichotomous trait using logistic regression model, which included gender, study site, and the top 10 principal components as covariates.

For BMI, individuals < 20 years of age were excluded. Stratified by cohort and gender, raw BMI was regressed on age, age^2, and study site within cohort, if applicable. The residuals were then fit to a standard normal distribution before combining the gender-specific residuals within cohort. BMI was analyzed as continuous trait using linear regression, with the top 10 principal components included as covariates.

For height, we excluded men < 23 years of age and women < 21 years of age, as well as individuals > 85 years of age. Stratified by cohort and gender, we regressed height on age and study site, when available. Residuals were normalized to a standard normal distribution. Outliers (> 4 SD or < -4 SD) were excluded from the analyses. Height was analyzed as continuous trait using linear regression, with the top 10 principal components included as covariates.

The number of individuals analyzed for each phenotype can be found in **Table 1**, organized by each of the CARe cohorts.

REFERENCES

LETTRE, G., C. D. PALMER, T. YOUNG, K. G. EJEBE, H. ALLAYEE *et al.*, 2011 Genome-Wide Association Study of Coronary Heart Disease and Its Risk Factors in 8,090 African Americans: The NHLBI CARe Project. PLoS Genet **7:** e1001300.

LO, K. S., J. G. WILSON, L. A. LANGE, A. R. FOLSOM, G. GALARNEAU *et al.*, 2011 Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans. Hum Genet **129:** 307-317.

MUSUNURU, K., G. LETTRE, T. YOUNG, D. N. FARLOW, J. P. PIRRUCCELLO *et al.*, 2010 Candidate gene association resource (CARe): design, methods, and proof of concept. Circ Cardiovasc Genet **3:** 267-275.

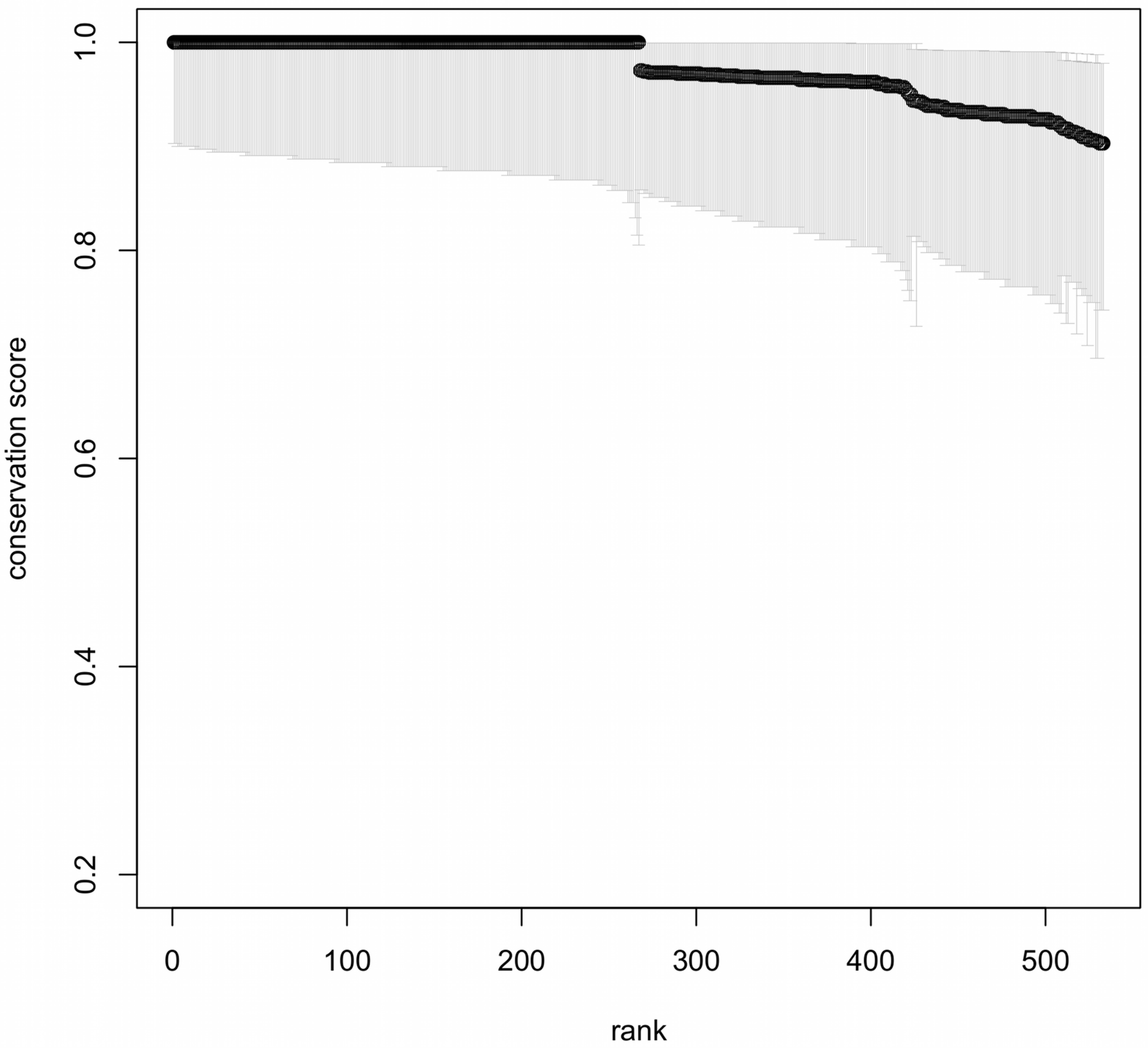

Figure S1 The conservation scores of hcSNPs. The set of hcSNPs were ranked by the conservation scores calculated as described in the **Materials and Methods**. For each SNP the exact 95% confidence interval of the conservation score was also calculated using the binom.test function in R, and are shown in the grey bars.

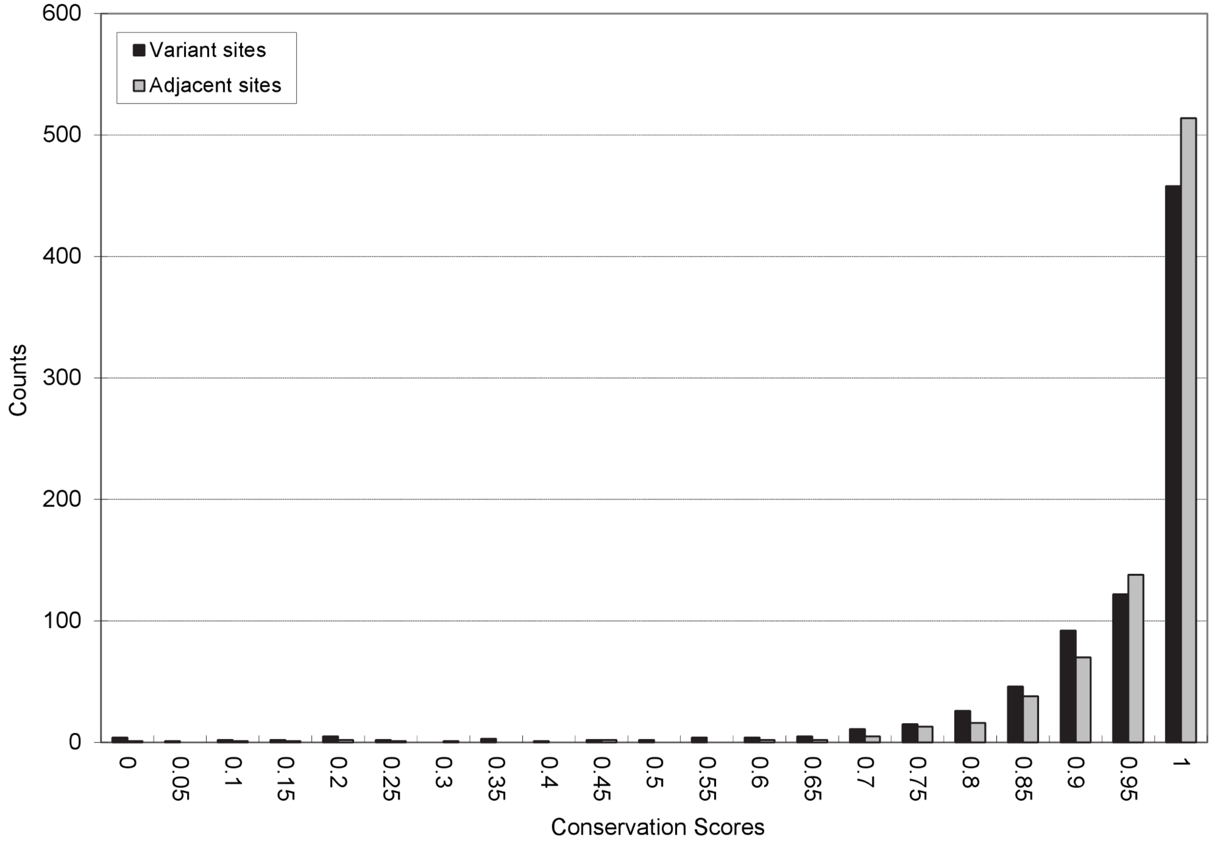

Figure S2 Distribution of conservation scores for SNPs and their immediate 3' or 5' neighbors found in the ultra- and near-ultraconserved elements. For each SNP found in ultra- or near-ultraconserved elements (nUCEs; black bars) we used the chimpanzee allele as the reference allele and calculated the conservation score as 1-proportion of bp substitutions found in alignments of 44 vertebrate species (see **Materials and Methods**). We also calculated the conservation score for the basepair randomly chosen to be immediately 3' or 5' of each SNP (grey bars). SNPs found in nUCEs are much more highly constrained than random SNPs from the genome, matched by allele frequency (data not shown), but the constraints are significantly relaxed relative to the adjacent basepair ($P = 1.1x10^{-7}$ by two-sided Wilcoxon signed rank sum test).

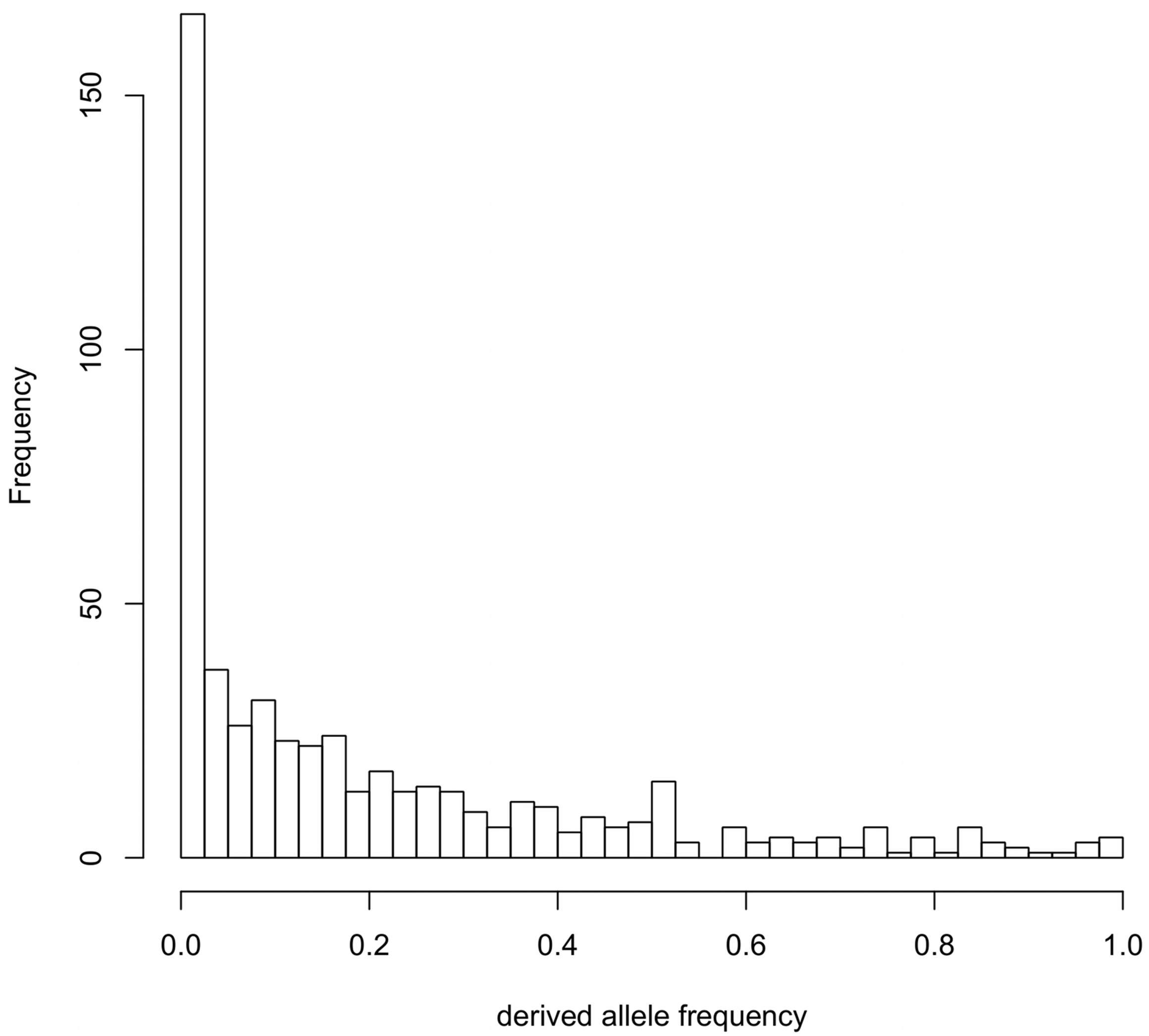

Figure S3 Derived allele frequency distribution of 533 hcSNPs in our study. The derived allele frequency (DAF) was computed by averaging the DAF across all European-American founders in all cohorts of CARe, weighted by sample size in each cohort. These cohorts include ARIC, CARDIA, CFS, CHS, FHS, and MESA. Note that 9 hcSNPs have a derived allele frequency > 0.9, perhaps reflecting variants that offer an adaptive advantage. However, for the purposes of the present study, these SNPs were neither singled out nor removed from analysis. Results here included SNPs that are monomorphic in all six cohorts examined.

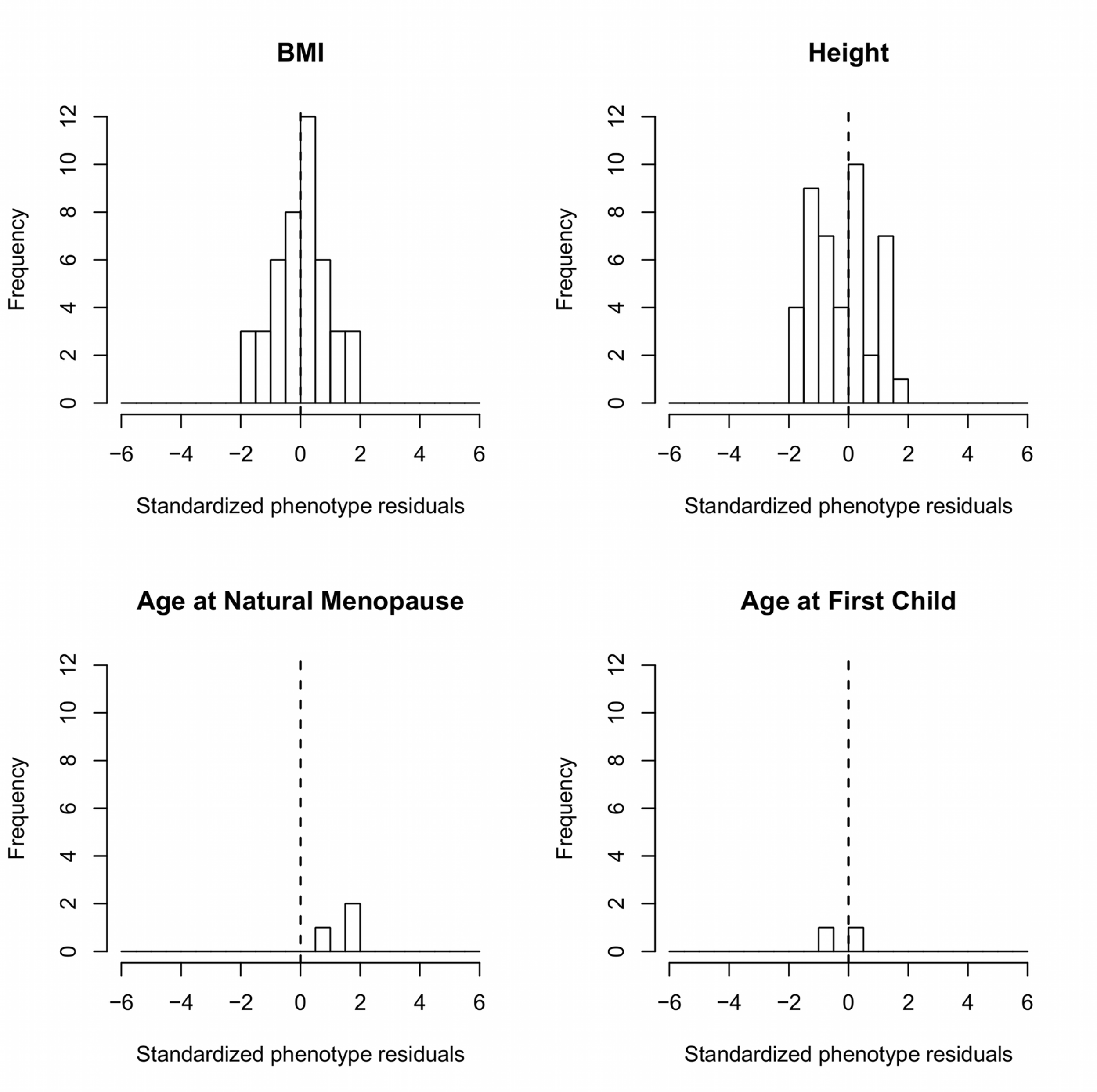

Figure S4 Distribution in standardized phenotype residuals for individuals homozygous for the rare derived allele of at least one hcSNP. Among the individuals homozygous for the derived allele (DAF ≤ 0.005) of at least one hcSNP, none exhibited extreme phenotype residuals relative to the population for any of the four quantitative normally-distributed traits. Interestingly, if all SNPs in ultra- or near-ultraconserved elements were examined, then two unrelated individuals each homozygous for rs10493810 were 1.6 and 4.6 standard deviations, respectively, above the population mean for age at natural menopause, while another individual homozygous for rs4664775 was 5.4 standard deviations above the population mean for the same trait. Since rare derived alleles are expected to lower reproductive fitness, it is surprising that multiple individuals homozygous for a rare variant would have a lengthened reproductive span.

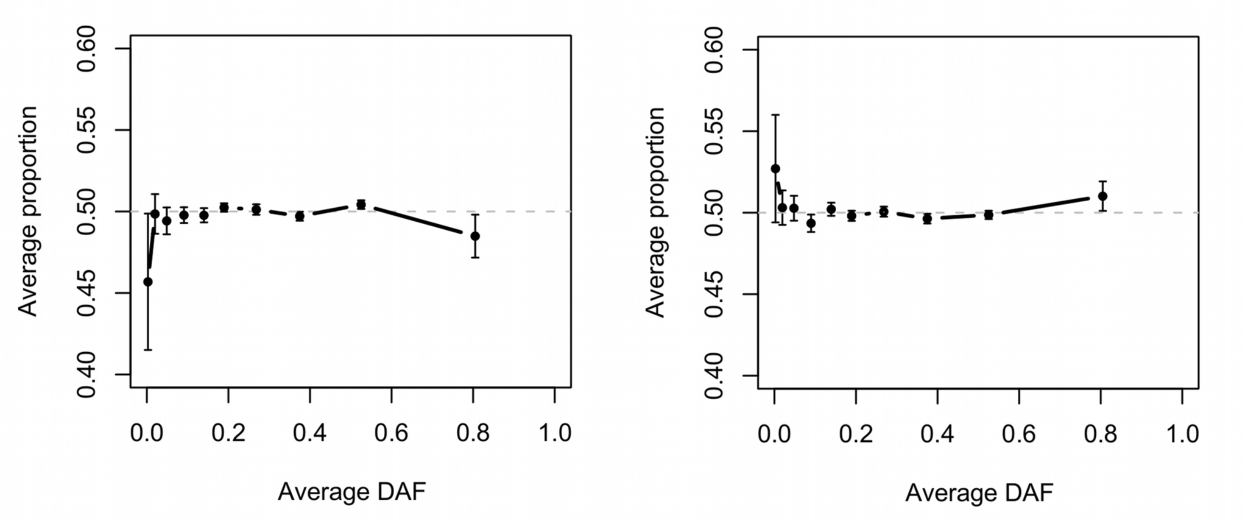

Figure S5 Transmission distortion analysis of hcSNPs stratified by paternal and maternal transmissions. SNPs were stratified into deciles by DAF, the average proportion of transmissions (derived allele transmissions / total transmissions). The derived alleles of hcSNPs are transmitted 49.3% of the time (average difference = -0.21, *P* = 0.40) when transmitted paternally (left) and 50.3% of the time when transmitted maternally (right, average difference = -0.50, *P* = 0.28). Among the lowest decile, the derived allele of hcSNPs appears to be under-transmitted when transmitted paternally, but over-transmitted when transmitted maternally. However, the difference is not significant (*P* = 0.09). Error bars reflect the standard error in the proportion of derived allele transmissions within the decile.

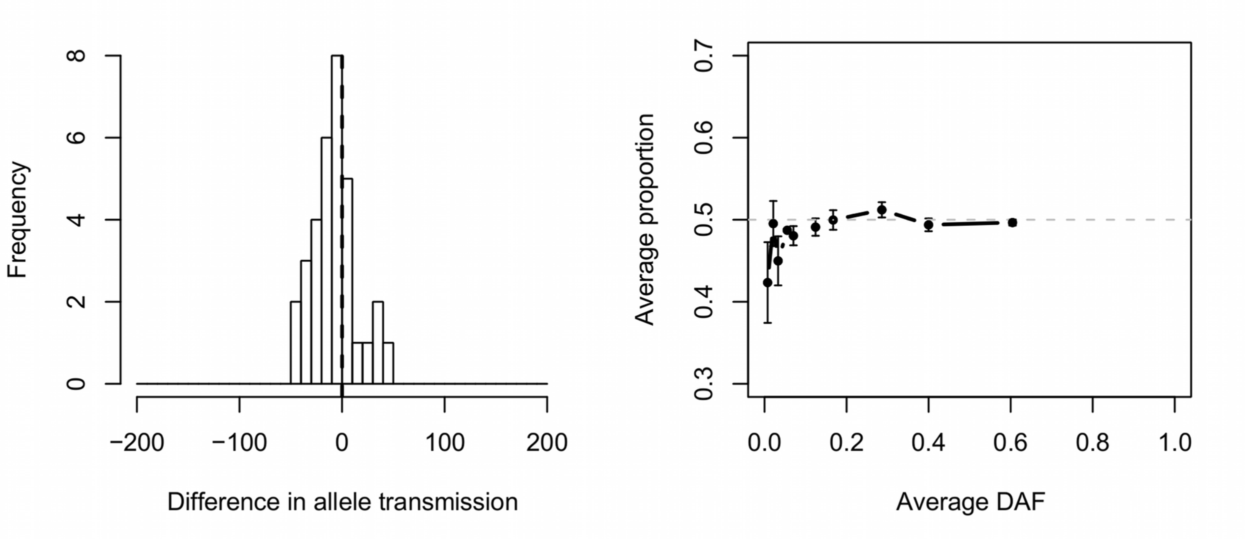

Figure S6 Transmission distortion analysis of only those hcSNPs found in ultraconserved elements (100% conservation) using trio data from FHS. Highly constrained SNPs found in elements conserved at 100% showed a nominally significant under-transmission (N = 33; proportion of derived allele transmission = 48.1%; average difference = -6.57, *P* = 0.048) when compared to those found in elements conserved at 99% and 98% (N = 177 and 229; proportion of derived allele transmission = 49.7% and 50.2%; average difference = -0.91 and 0.18, *P* = 0.34 and 0.54, respectively; data not shown). The less frequent transmission of the derived alleles at hcSNPs in 100% elements is unlikely to be explained by a stronger evolutionary constraint among SNPs in the 100% conserved elements, since we focused on the hcSNPs for this analysis and the distribution of the conservation scores for hcSNPs in the 100% conserved elements did not significantly differ from the distribution of conservation scores in the 98% and 99% conserved elements (mean SNP conservation score = 0.981 vs. 0.974, *P* = 0.166 by two-tailed Mann-Whitney test). (Left) Distortion in the transmission of the derived alleles of hcSNPs from heterozygous parents to offspring was measured by subtracting the ancestral allele transmissions from the derived allele transmission for each SNP. (Right) The average proportion of derived allele transmission when SNPs were stratified into deciles by DAF. Error bars reflect the standard error in the proportion of derived allele transmissions within the decile.

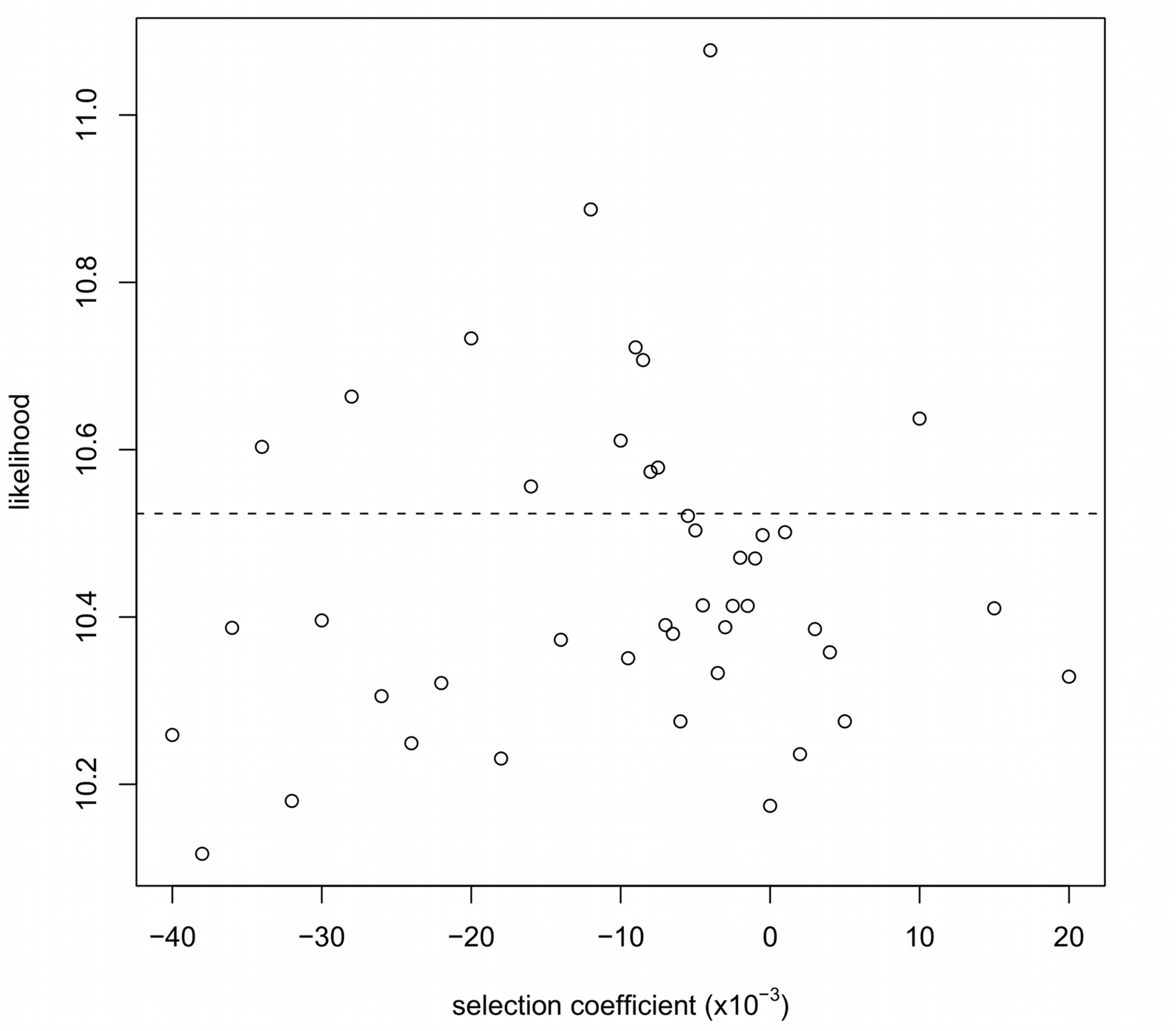

Figure S7 Estimating the selection coefficient most consistent with the data observed in FHS. At each value of selection coefficient tested, we simulated 1,000 sets of family trio data as described in **Materials and Methods**. Based on the distribution of the average proportion of derived allele transmission, we calculated the likelihood of observing a proportion of derived allele transmission of 0.494, as observed in the FHS data. The maximum likelihood estimate of the selection coefficient is -0.004. The dashed line denotes 95% of the maximum likelihood. The range of selection coefficients within 95% of the maximum likelihood is 0.01 to -0.034.

Table S1 The distribution of hcSNPs found to be polymorphic in at least one CARe cohort.

Number of cohorts	Number of hcSNPs
6	407
5	30
4	25
3	17
2	18
1	14
0	22

Of the 533 hcSNPs successfully genotyped in at least one of the six CARe cohorts (ARIC, CARDIA, CFS, CHS, FHS, and MESA), 22 were monomorphic in all cohorts for which it was successfully genotyped and passed quality control. The remaining 511 hcSNPs formed the basis of all of our analysis.

Table S2 Descriptive statistics of the reproductive traits by CARe cohorts.

Cohort		Age at Natural Menopause	Number of Children	Age at First Child	Age at Last Child
ARIC	mean	48.4	3	n.a.	n.a.
	s.d.	4.0	1.6	n.a.	n.a.
	range	[40,60]	[0,12]	n.a.	n.a.
CARDIA	mean	n.a.	n.a.	27.8	n.a.
	s.d.	n.a.	n.a.	5.2	n.a.
	range	n.a.	n.a.	[14,39]	n.a.
CFS	mean	n.a.	n.a.	24.7	31.2
	s.d.	n.a.	n.a.	4.5	4.9
	range	n.a.	n.a.	[15,36]	[16,41]
CHS	mean	49.2	2.7	25.0	31.7
	s.d.	4.4	1.7	5.0	5.6
	range	[40,60]	[0,13]	[15,42]	[16,49]
FHS	mean	49.6	n.a.	27.3	31
	s.d.	3.2	n.a.	5.1	5.4
	range	[40,56]	n.a.	[15,43]	[18,45]
MESA	mean	49.5	2.7	24.2	n.a.
	s.d.	4.2	1.7	4.8	n.a.
	range	[40,60]	[0,11]	[15,42]	n.a.

For each CARe cohort included in the study, the mean, standard deviation (s.d.), and the range for each of the four reproductive traits analyzed are listed. n.a. denotes not available. Descriptive statistics were calculated for all available individuals from each cohort, though a subset may be removed from analysis due to missing covariates or other factors.

Table S3 Association of carrier status of at least one rare derived allele of hcSNP to each phenotype analyzed.

Trait	Effect	StdErr	P-value	Direction
Age at Natural Menopause	-0.2995	0.1949	0.1244	-?+-
Age at First Child	-0.0271	0.0703	0.7002	?---
Age at Last Child	-0.1100	0.1132	0.3318	??-?
BMI	-0.0239	0.025	0.3398	-+--
Height	-0.0244	0.0247	0.3234	--+-

Individuals were scored as “0” if they did not carry the derived allele for any of the 122 hcSNPs with rare derived alleles (DAF ≤ 0.005) and scored as “1” if they carried at least one rare derived allele. When the presence of one or more rare derived alleles was tested against the continuous normally-distributed traits (BMI, height, age at natural menopause, age at first child, and age at last child), we observed no associations after meta-analysis, although for age at natural menopause and age at last child the presence of rare variants on average decreased reproductive fitness, albeit not significantly. The “Direction” column reflects, sequentially, the direction of effect for ARIC, CARDIA, CHS, and MESA. “?” indicates that phenotype data were not available in that particular cohort.