Allelic differentiation of complex trait loci across human populations

Emily S Wong[1,2ψ] and Joseph E Powell[3,4]

[1] School of Biological Sciences, The University of Queensland, St Lucia, 4072, Australia

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[3] Institute for Molecular Biosciences, The University of Queensland, St Lucia, 4072, Australia

[4] Queensland Brain Institute, The University of Queensland, St Lucia, 4072, Australia

[ψ] Corresponding author: emily.wong@uq.edu.au

Abstract

The genetic basis for differences between humans lies at those DNA bases that vary between individuals. Comparative genomics is a powerful tool for dissecting the functions of genomes by comparing the genomes of divergent organisms to identify functional regions. However, much of the functional signals in mammalian genomes are non-coding and evolving rapidly leading to population-specific differences. Here, we use a sensitive genome-wide measure of human genetic differentiation between global populations to detect differences associated with complex traits. Highly differentiated genomic regions were associated with regulatory elements and morphological features. We observed variation in allelic differentiation between populations at tissue-specific expression quantitative trait loci (eQTL), with greatest effects found for genes expressed in a region of the brain that has been linked to schizophrenia and bipolar disorder. Consistent with this, genome-wide association study regions also showed high levels of population differentiation for these diseases suggesting that loci linked to neurological function evolve rapidly. Clear differences for genetic structure in populations were observed for closely related complex human phenotypes. We show that the evolutionary forces acting at pleiotropic loci are often neutral by comparing directional effects of traits under selection. Our results illustrate the value of within species comparisons to understanding complex trait evolution.

Author Summary

Differences in genome sequence, both within and between populations, are linked to a wide spectrum of morphological, physiological and behavioral differences between humans. Many of these observed differences are considered 'complex' – meaning that they are encoded in the genome at many locations. By comparing variation in DNA bases between populations, to genomic regions for complex traits, we can gain insights into differences in the evolutionary trajectory of different types of complex traits. Here, we use a sensitive method to quantitate genetic differences between the major human populations. Although certain diseases are closely related, they can show distinct patterns of genetic differentiation between populations. Certain regions of the genome are more distinct between populations, including regulatory regions, particularly those linked to changes in the number of genes expressed in the brain, and those involved in types of neurological disease.

Introduction

An overarching goal of genomics is to understand how the genome sequence encodes phenotypes. Comparative genomics, based on examining evolutionarily constrained sequences, have been instrumental in identifying those genic features of mammalian genomes conserved over millions of years. In recent years, the comparative approach has allowed for a broadened understanding of the genetic and molecular basis for phenotypic differences amongst populations [1–3].

In recent years, it has become increasing clear that the vast majority of genetic signals causal for phenotypic variance locate to non-protein coding regions [4]. However, these regions are challenging to study due to their rapid evolution, which hampers the unambiguous alignment of sequences from different species for inferring common function. For instance, only 10-22% of lineage-specific transcription factor (TF) binding events are conserved between human and mouse and only ~20% of mouse TF binding sites have orthologs in humans [5,6]. Here, we use a population genetics approach to understanding phenotypic differences in humans at complex traits. By comparing between human populations, it is possible to examine human-specific genetic variants and capture rapid evolutionary changes between populations.

There has long been an interest in understanding selection and adaptation in modern human populations [7,8]. Many disease, anthropomorphic, and behavioral phenotypes that vary within human populations are complex, in that they are determined by a combination of alleles at a large number of independent loci, as well as non-genetic factors. The process of global colonization by humans has played an important role in shaping current patterns of genetic diversity, both within and between population groups. Therefore, it is reasonable to assume that differences between populations in such phenotypes will partially be explained by geographic variation in trait loci and allele frequency [9–11]. The recent availability of high-density genetic information allows us to infer relationships between human populations and, through this, gain an understanding of past demographic events [12].

The classical selective sweep model in which a new advantageous mutation arises, and spreads quickly to fixation due to natural selection is often termed as ‘hard-sweep’ [13]. Under this model, neutral variation near a selective site ‘hitch-hikes’ along with the favored allele. This impacts patterns of variation around the selected site in ways that can be detected using a variety of tests of selection [7]. However, there are two slightly different scenarios

that both contrast with the standard hard sweep model. In one scenario, due to a change in selection standing variation becomes selectively favored, resulting in changes in allele frequency. In the second scenario, multiple independent mutations at a single locus are all favored, and all increase in frequency simultaneously until the sum of the frequencies is 1. If the favored alleles were all similarly advantageous, then typically none of the favored mutations would fix during the selective event. Both scenarios tend to be more difficult than hard sweeps to detect using standard tests of selection, and are typically termed 'soft sweeps' [13].

Allelic flux not due to non-adaptive forces also plays a key role in shaping phenotypic diversity within and between populations and species [14,15]. Herein, we use a sensitive method based on principal components that accounts for differences in allele frequencies at genetic variants. Using this metric, we interrogated complex phenotypes such as gene expression, anthropomorphic, behavioral and disease traits, where observed phenotypic differences could be due to genetic drift, population demography or natural selection.

Highly differentiated genomic regions were associated with regulatory elements and morphological features. Variation in allelic differentiation was observed between populations at tissue-specific eQTLs. The highest levels of differentiation were observed at brain-related eQTLs in genomic regions associated with psychiatric diseases. Indeed, increased genetic differentiation between populations were also observed at bipolar disorder- and schizophrenia-associated loci compared to other diseases suggesting that loci linked to neurological traits have evolved rapidly among human populations. Interestingly, clear differences in the polygenic signatures of trait evolution can be observed between related complex human phenotypes after accounting for differences in the evolutionary rates of different genomic regions. Finally, genetic differences at loci shared between selectively constrained complex traits are often indistinguishable from genome-wide expectation suggesting that evolutionary forces acting at pleiotropic loci are mostly neutral.

Results and Discussion

We used principal component analysis (PCA) to devise a measure that reflects allelic differentiation between major extant human populations using the 1000 Genomes Phase 3 dataset (**Methods**). This method is most similar to those described by [16,17]. Each variant is scored based the orthogonally transformed genotype using the dot product of the PC loading

and the standardized genotypes (termed the 'PC score')(n=~12 million SNPs) (**Supplementary Figure 1A, 1B, 2**). Our final measure, which we term *d* (**Supplementary Figure 1C**), corresponds well to the common measure of genetic differentiation – the F_{ST} index (**Supplementary Figure 1D**). The method differs from the popular F_{ST} measure in several ways. *i)* Population differences are measured concurrently in a single framework reflecting evolutionary radiation rather between two groups. *ii)* Population structure was defined using over 800,000 SNPs. *iii)* The method avoids biases in the pre-specification of population groups. *iv)* The measure covered the genome at over 12 million SNPs, allowing quantitative assessment of evolutionary change at complex trait loci.

Using this metric, we observed strong correspondence at previously reported regions that are highly differentiated between populations [10,18](**FIG 1A-B**). These regions include skin and hair pigmentation genes in *SLC24A5* and *SLC45A2* responsible for light skin in Europeans, as well as immune and metabolic genes.

We first performed high-level comparisons between functionally annotated regions and between loci associated to Mendelian traits and diseases. We then focused on quantifying genetic differentiation at complex traits.

Coding regions are the least genetically differentiated between populations while regulatory elements show the most structure

We partitioned the genome by functional annotation categories which include protein-coding regions, introns, 5' and 3' untranslated regions (UTRs), repeats-containing regions as well as enhancer and promoter locations inferred by eRNAs and ChIP-seq of transcription factor binding sites (TFBS) and chromatin marks in a number of cell types [19]. Compared to the genome-wide median, population differentiation was significantly enriched at loci marking weak enhancers, defined by histone modification patterns [20] and, as expected, significantly reduced near protein-coding regions (permutation $p<0.05$) (**Supplementary Figure 3**). A high level of genetic structure in regulatory elements is consistent with the fast evolutionary turnover of TF binding sites across species [5,6].

Highly differentiated region are enriched at morphological features

We ranked all variants by the degree of genetic differentiation and examined the top 10,000 most differentiated variants (0.08%) for functional enrichment of human phenotypes.

Genotype-phenotypes associations from the Online Mendelian Inheritance in Man (OMIM) database were used to predict the function of putative cis-regulatory regions underlying our variant test set. Variants were coupled to genes based on genomic distance (**Methods**). We observed enrichment of OMIM annotations for skeletal structure, hair patterning, facial morphology and glucose metabolism (binomial and hypergeometric tests, FDR<0.05)(**Supplementary Figure 4**).

Brain and immune-related eQTLs show greater population differentiation compared to the eQTLs of other tissues

Changes in gene expression underlie many phenotypic differences between organisms [21]. Across mammalian species, the rate of gene expression evolution varies between organs and lineages [22]. Genetic variants that effect expression levels are known as expression quantitative trait loci (eQTLs). Using eQTLs from the GTEx dataset (V6) comprising of individuals from several distinct populations, where most individuals were of European Americans descent (84%), with smaller proportions of African Americans (13%) and Asians (1%), we compared the mean *d* value at eQTLs taken from the GTEx database across 44 tissues derived from 449 donors and 7,051 samples in total. 1000 Genome variants were overlay with eQTL SNPs and an average *d* value was calculated on a tissue-specific basis. Permutations were used to assess the significance of the mean d value by sampling an equivalent number of eQTLs across the genome matched for MAF and distance to TSS. This procedure was repeated 1,000 times to calculated an empirical two-tailed *p*-value.

The extent of overall population differentiation at eQTLs varies between tissues with the highest level of genetic differentiation observed at brain- and immune-associated eQTLs, with the most difference observed at the anterior cingulate cortex (**FIG 1C**). By contrast, lung associated eQTLs were the least differentiated.

Across mammals, changes to the expression of conserved genes are slower in nervous tissues compared to other tissue types [22]. However, in humans, conserved non-coding regions show accelerated evolution near neuronal genes [23], and recently-acquired brain development genes also display accelerated sequence changes [24]. To compare brain eQTLs versus eQTLs identified in other tissues, we calculated the median *d* values at eQTLs for genes expressed in each tissue and compared between grouped brain and non-brain-associated tissues. Our results revealed increased allelic differentiation at brain eQTLs (Mann-Whitney-

U test p=2.2×10^{-4}). Similar findings were observed at human-specific brain eQTLs (Mann-Whitney-U test p=8.5e-3). These were determined by sequence alignment of the human and chimpanzee genomes (**Methods**).

We also sought to determine whether more genetically differentiated eQTLs were associated with greater variability in gene expression levels. We measured median variability in gene expression of genes expressed in each tissue using the coefficient of variation (CV) for normalized read counts form individuals in the GTEx study (after removing genes with less than 10 mapped reads across at least 20 individuals)(genes n=56,318). Consistently greater CV was observed across brain tissues (12 brain regions) compared to other tissues (excluding whole blood, n=41) (Mann-Whitney-U test of median cv for genes expressed in each tissue, p=0.001). However, at the individual gene level, a linear relationship between genetic differentiation at QTLs and gene expression was not found.

Our results newly revealed that regulatory regions of genes expressed in brain are more genetically differentiated between populations and this observation is consistent with reported rapid sequence evolution at human neuronal genes.

Loci for neurological traits show an increase in genetic differentiation

The anterior cingulate cortex has been associated to mood disorders including bipolar disorder and schizophrenia [25,26]. Thus, we asked whether increased genetic differences could also be observed at GWA loci associated with these diseases.

We compared GWA associated variants using summary statistics from 49 disease, anthropomorphic and behavioral phenotypes for differences in levels of allelic differentiation between populations. We extended our analysis to include all GWA variants below a p-value <1×10^{-4}, as the majority of polygenic trait heritability appears due to SNPs that do not reach genome-wide significance [27] (**FIG 2A**, **Supplementary Table1, Supplementary Figure 5**). We set as null a value of allelic differentiation for each set of GWA variants on a per trait-basis and tested for significant departure of allelic differentiation from the null distribution. We constructed the null distributions in two ways. In the first instance, the null value was generated based on genomic functional categories to account for differences in the underlying evolutionary rates. Our empirical null distribution is generated per functional category per trait based on resampling an equivalent number of SNPs to the number of pruned SNPs per

iteration. (**Methods)(FIG 2B**). To control for type I errors, we used permutation strategies to test for increased or reduced levels of genetic differentiation. We performed 1000 permutations to estimate the significance of our results. We also tested against a null model that was estimated by matching variants based on minor allele frequency (MAF) and the distance to the nearest transcriptional start site (TSS) (**Methods)**. By sampling from matched values, we generated the background distribution for each variant and this was used to calculate an empirical *p*-value.

We detected some of the highest levels of overall genetic differentiation at loci associated with bipolar disorder and schizophrenia (**FIG 2A,B**). Results were consistent between alternate methods of estimating null values (**Methods**), and the majority of the tested GWA loci did not directly overlap with eQTLs. This observation is in contrast to the reduced levels of genetic differentiation measured across all disease-associated loci, including, for example, those loci associated with metabolic and skeletal diseases (**FIG 2A,B**). The most plausible explanation for this general reduction in genetic differentiation at disease loci is the widespread selection against harmful mutations in a healthy cohort such as 1000 Genomes [28].

In a pairwise manner, we further assessed whether shared loci between traits display a degree of allelic distributions significantly different from the trait-specific background for each trait. For each pair of traits, we overlapped GWA variants (*p*-value cut-off=$1x10^{-4}$) based on genomic location and removed SNPs determined to be in LD with one another based on the parameters described above. Significant departure of the *d* score from null was calculated by performing 1000 randomizations of all SNPs meeting the GWA cut-off for each trait in turn. We found a higher degree of allelic differentiation at loci shared between educational attainment and height (permutation $p<0.01$)(**Methods**). Modest enrichment in *d* was also observed at shared loci for schizophrenia and educational attainment (permutation $p\leq0.1$, n=7). Closer examination of these regions revealed increased genetic differentiation at regions on chromosomes 3 and 12 as well as the MHC. Interestingly, these loci are some of the most pleiotropic in the genome and have been linked to a number of diseases [29].

We note that population stratification in the mapping panel of discovery GWA studies can confound GWA results by producing false positives with high genetic divergence. Recent methods such as mixed models control for stratification was used for the major depressive

study analyzed here [30]. The use of sib-pairs also avoids confounding association from stratification [31]. For other cases, we performed two simulations to access the potential impact of population stratification on our results. First, we simulated a polygenic trait using a non-ascertained cohort (GERA) and show that standard correction for population stratification in a GWA study is sufficient control for bias in our study (*45,* **Methods, Supplementary figure 7, Supplementary Table 4**). Second, given that many GWA studies comprised of European cohorts only, we demonstrated that loci with high population structure in European populations typically does not correspond with highly structured loci in 1000 Genomes (**Methods, Supplementary figure 8**). This suggests that false positives discovered using European cohorts are unlikely to exert a systematic bias to our analyses. We also replicated our analyses across GWA studies using an independent cohort and obtained highly concordant results (**Method, Supplementary Figure 9**)

In summary, increased genetic differentiation at GWA loci associated with neurological traits is consistent with differences at eQTLs of brain-expressed genes. That loci linked to neurological traits appear show increased differences between populations suggest that they have evolved rapidly. Increase incidence of deleterious alleles for neurological diseases may be a consequence of selection for other, non-disease phenotypes at pleiotropic loci.

Disparities in population differentiation between related complex diseases

Next, we examined genomic patterns of population differentiation between related traits. We directly compared the degree of genetic differentiation amongst populations for type I (T1D) versus type II diabetes (T2D) and between ulcerative colitis and inflammatory bowel disease (IBD).

T1D loci compared to T2D loci showed remarkably different signatures of genetic differentiation (**FIG 3A**). T1D associated loci displayed much reduced levels of genetic differentiation across populations compared to T2D loci (**FIG 3B**, Mann Whitney U $p=1.0\times10^{-5}$). Decreased population differentiation is suggestive of extensive purifying selection while signatures of increased genetic differentiation at genetic loci indicate differences in allele frequencies across different populations. T2D loci with significantly increased population differentiation were proximal to four genes that have been robustly implicated in the disease. Earlier studies using HapMap variants have reported high genetic

differentiation between populations at *TCF7L2* [32,33], and we newly identified enrichment of genetic differentiation at variants associated with *PPARG*, *WFS1,* and *IGF2BP2*.

We further compared LD scores at loci associated with both diseases. LD scores provide a summary of linkage disequilibrium in a local region [34]. Measures of LD, such as long-range haplotypes in a population, are commonly used to characterize the increased levels of LD expected in a region undergoing positive selection [35]. We observed that LD scores were also more highly elevated for T1D versus T2D indicating stronger ongoing selection against deleterious T1D alleles (**FIG 3C**, Mann-Whitney U p=2.5×10^{-19}).

The high prevalence of T2D in human population is likely due to a combination of genetic drift and positive selection where increased genetic differentiation at some T2D loci could indicate soft selective sweep. However, it is unknown whether these variants were directly selected, for example, for metabolic sequestration during periods of food shortages (i.e. 'thrifty gene hypothesis'), or indirectly so through pleiotropy and hitchhiking, where an allele is linked to the sweep of a beneficial allele.

Large disparities in evolutionary histories were also observed between different forms of inflammatory bowel disease (IBD) We compared population differentiation at genetic loci implicated for ulcerative colitis to those at associated with IBD. We observed much reduced population structure at ulcerative colitis loci (Mann-Whitney U test p<3.2 × 10^{-31}) (**Supplementary Figure 6**). As both ulcerative colitis and Crohn's disease are forms of IBD, we removed Crohn's disease associated loci from the IBD variants to determine whether this difference is primarily due to Crohn's disease variants. The difference remained significant suggesting that ulcerative colitis alleles are under stronger purifying selection than other forms of IBS, and not only at loci implicated for Crohn's disease. We also observed significantly greater LD at variants associated with ulcerative colitis compared to IBD (Mann-Whitney U test p<3.2×10^{-31}), again indicative of strong negative selection acting upon ulcerative colitis associated loci.

Taken together, our results reveal surprisingly distinct patterns of population structure for loci of highly related diseases. These large differences in genetic architecture are suggestive of differences in selective pressure between the diseases.

Opposing effects are common at pleiotropic loci

Theoretical studies have corroborated that most quantitative traits are due to many loci of small effects [13,36,37]. Thus, it stands to reason that many loci are shared between phenotypes [38]. To examine evolutionary histories at shared loci, we tested for departures of observed *d* from the null at 341 shared loci obtained from analyses of 44 GWA studies, excluding loci containing pairs of highly related traits and loci in the MHC region using a published dataset [39]. We considered resampled datasets of SNPs randomly drawn from the *d* distribution that is matched for MAF and distance to nearest TSS. Across loci, we observed an excess of regions with a lower than expected *d* score (binomial test p=1.9x10^{-11}) suggestive of negative or stabilizing selection.

To gain insights into the directional contribution of GWAS traits at pleiotropic loci, we considered how alleles shared between traits contribute to each shared loci. At half of those instances where multiple disease traits collocate to a genomic region, a risk allele, that is, an allele associated with an increased risk at a locus for a disease, was associated with the non-risk allele of another disease. A similar proportion of cases where height-increasing alleles, a well-studied highly polygenic trait widely accepted to be evolving under positive selection, were coupled to a disease risk allele were found.

We performed similar analyses using a different strategy to classify shared loci, whereby GWA variants for all traits (p-value cut-off of 1x10^{-4}) were clustered by genomic proximity with a flanking region of 5kb added to each side. Overlapping regions were merged and the mean *d* of all variants falling within each region was calculated. For each clustered region (where there are more that two traits associated; n=2,649), we tested for departures of the mean *d* against a background of variants at tiled genomic regions the same genomic length as that of the clustered region. Two-tailed significance was calculated by comparing the mean *d* value for each cluster by sampling from its null background for 1,000 permutations.

Several highly pleiotropic genomic clusters with elevated levels of genetic differentiation among populations were identified (two-tailed permutation p<0.05; **FIG 4A, 4B**). Consistently, height loci were equally as likely to be located at loci of increase disease risk compared to one of decrease risk (n=243). However, risk alleles for disease traits at shared loci were significantly more concordant than expected by chance indicative of purifying selection at shared disease loci (binomial test p=9.0x10^{-11}, n=204).

Taken together, our results suggest that opposing directional effects are widespread and that despite instances of elevated population differences at shared regions, small effects at shared regions often act under effectively neutral forces.

Conclusion

Understanding the genetic basis of traits or phenotypes remains a central goal of evolutionary genomics. To this end, 'reverse-genetics' approaches have used genome scans to infer evolutionary histories by focusing on detecting the reduction of genetic diversity associated with selective sweeps [33,40,41]. However, bottlenecks and recent growth, common to human populations, can cause much of the genome to resemble selective sweeps, which can bias the identification of adaptive loci [42]. We have focused on quantitating genetic differences in allele frequencies across a range of complex traits using eQTLs and GWA summary statistics, accepting that these differences may be caused by genetic drift, population demography or natural selection.

Although GWA SNPs discovered in one genetic background can often describe the same phenotype in another population of a different genetic background [43], we note that loci identified for a GWA phenotype in one population, for example, Europeans, specifically describes differences among individuals in that population and the same loci in other populations may not reflect the same trait, evolutionary forces or demographic events. For instance, educational attainment [44] and height [45] GWA studies were both based on mapping individuals of European descent and the interpretation of their associated genetic loci in other populations should be interpreted with caution.

In conclusion, we measured allelic differentiation at complex trait loci using a sensitive measure of allelic differentiation across human populations. Allelic differentiation between populations was highly dependent upon the phenotype measured reflecting recent evolutionary history. Brain-associated regulatory elements tend to show a higher level of population differentiation. Disease states typically show less genetic structure between populations but the degree of this varies even between closely related traits. Shared loci commonly show antagonistic allelic effects with neutral outcomes consistent with small effects of contributing alleles for each phenotype, each with likely minimal influence on overall fitness.

Methods

Quantitation of genetic differentiation amongst populations by principal component analysis

Principal Component Analysis (PCA) was used to orthogonally transform genotype information from the 1000 Genomes Phase 3 dataset [28](human genome version hg19). PCA is a common technique to visualize population structure amongst a cohort of individuals using genotype data. The first principal components, which account for the majority of the variability within the data typically reflects population structure [46].

If a centered genotype matrix **X** (where means are subtracted from each column) is of $n \times p$ size, where n is the number of individuals and p is the number of SNPs.

Then the $p \times p$ covariance matrix $\mathbf{C}$ is given by $\mathbf{C} = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n-1}$. This is a symmetric matrix and can be diagonalized:

$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathrm{T}},$$

where **V** is a matrix of eigenvectors (each column is an eigenvector), **L** is a diagonal matrix with eigenvalues λ_i in decreasing order along the diagonal.

PCA can then be performed using singular value decomposition (SVD) of the genotype matrix **X**.

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}},$$

where **S** is the diagonal matrix of singular values s_i. Hence,

$$\mathbf{C} = \frac{\mathbf{VSU}^{\mathrm{T}}\mathbf{USV}^{\mathrm{T}}}{n-1} = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^{\mathrm{T}}$$

it is evident from this that **V** are the eigenvectors and the PC scores, **XV,** are equal to $\mathbf{USV}^{\mathrm{T}}\mathbf{V} = \mathbf{US}$.

We used FastPCA (https://github.com/ajverster/FastPCA) to perform PCA. FastPCA uses an approximation procedure of performing SVD by first randomly projecting the data into subspace based on [47]. The dataset was preprocessed to remove variants that may bias the estimation of population structure. We used PLINK 1.9 [48,49] to remove SNPs with a minor allele frequency (MAF) less than 1%. SNPs that were not in Hardy-Weinberg Equilibrium were also removed (p<1x10^{-6}) and the data was iteratively LD-pruned with an r^2 cutoff of 0.2 [17]. SNPs located on sex chromosomes were also removed. The data was then mean centered and standardized to unit variance prior to PCA and PCA was run on 818,280 SNPs.

We define a score to reflect the extent of genetic differentiation between populations. To do this, we first derive the PC scores for each SNP using the first three PCs (n=12,056,538 including those SNPs which were filtered out previously for the accurate estimation of population structure). For each SNP, this is calculated as the dot product of the PC loading and the standardized genotypes. The PC score is thus a composite variable that provides information on how each SNP is placed with respect to a PC. Finally, we combined the squared factor scores for each of the three PCs into one score by summing the scores for each PC after weighting each by the amount of variance proportionally explained by the first three PCs:

$$d = \sum_i p_i \times s^2$$

where p_i is the proportion of total variance explained by the first three PCs explained by the i th PC where $i(1,2,3)$, and s is the PC score.

***Cis*-regulatory functional enrichment using OMIM annotation**

The 10,000 most highly differentiated variants across the genome were associated with the *cis*-regulatory regions of nearby genes using the GREAT tool [50] and tested for functional enrichment using OMIM annotation using both binomial and hypergeometric tests with FDR adjustment. Variants were associated with a gene if it is located within the basal regulatory domain of a gene, which is defined as up to 1000kb extended from a gene's TSS in both directions until within 5kb upstream or 1kb downstream of the TSS of the next gene. The test set of 10,000 genomic regions was associated with 2,459 (14%) of all 18,041 genes in GREAT's OMIM database.

Estimation of allelic differentiation at eQTLs

eQTL datasets were downloaded from the Genotype-Tissue Expression (GTEx) consortium website (www.gtexportal.org). The dataset comprised for eQTLs for 44 tissues, derived from 449 donors and 7,051 samples in total (V6). The GTEx dataset (version V6p) comprises of individuals of different self-reported ethnicities. Most individuals were of European Americans descent (84%), with smaller proportions of African Americans (13%) and Asians (1%). 1000 Genome variants were overlay with eQTL SNPs and an average *d* value was calculated on a tissue-specific basis. Permutations were used to assess the significance of the mean d value by sampling an equivalent number of eQTLs across the genome matched for MAF and distance to TSS. This procedure was repeated 1,000 times to calculated an empirical two-tailed *p*-value.

We defined putative human-specific variants as those variants that did not map to the chimpanzee genome (panTro4) using the UCSC liftOver tool. Alignments (chain files) used in the liftOver tool were created based on joining the longest syntenic or best-conserved aligned regions.

Significance testing of selection at GWA traits and partitioning of selection coefficient to functional categories

We partitioned our genome-wide selection score by functional annotations [19] taken from a range of published studies including assays of transcription factor binding sites, histone marks, DNaseI hypersensitivity regions across numerous cell types (**Table S2**). This strategy has the advantage of removing bias due to differences in evolutionary rates in different genomic regions.

Due to the polygenic nature of complex phenotypes, we considered all SNPs with GWA *p*-values below $1x10^{-4}$. At each functional annotation region for each trait, we LD pruned SNPs (r^2 cutoff=0.2) and calculated the mean *d* score for variants in linkage equilibrium. To control of type I error, we used permutation strategies to assess for increased or reduced levels of genetic differentiation. We performed 1000 permutations to estimate the significance of our results by resampling SNPs from the same functional category. Our empirical null distribution is therefore generated per functional category per trait based on resampling an equivalent number of SNPs to the number of pruned SNPs per iteration.

We also performed the same analyses using only the top 100 variants from each GWA study, ranked by *p*-value conditional on $p<1x10^{-4}$, and obtained similar results (data not shown).

For each SNP where GWA $p<1x10^{-4}$, we also calculated an empirical *p*-value based on generating a null distribution for each variant by subsampling based on matched MAF and distance to TSS. We first binned all 1000 Genome variants into 10 equally sized bins for MAF and 20 equally sized bins for distance to TSS. By sampling from matched bin sizes, we generated the background distribution for each variant and used this to calculate an empirical *p*-value.

Randomizations to assess the significance of genetic differentiation at shared loci

We clustered proximal GWA variants (*p*-value cut-off of $1x10^{-4}$) based on genomic location with a flanking region of 5kb added to each side. Overlapping regions were merged and the mean *d* of all variants falling within each region was calculated. We excluded variants of the same trait within a cluster and IBD loci which contain both Crohn's disease and ulcerative colitis loci For each clustered region (where there are more that two traits associated; n=2,649), we tested for departures of the averaged *d* against a background of averaged variants at tiled genomic regions the same genomic length as that of the clustered region. Significance (two-tailed) was calculated by comparing the mean *d* value for each cluster by sampling from its null background for 1,000 permutations.

Pairwise estimation of allelic differentiation at shared loci

In a pairwise manner, we assessed whether shared loci between traits possessed a *d* score that was significantly different from the trait-specific background from each trait. For each pair of traits, we overlapped GWA variants (*p*-value cut-off=$1x10^{-4}$) based on genomic location and

removed SNPs determined to be in LD with one another based on the parameters described above. Significant departure of the *d* score from null was calculated by performing 1000 randomizations of all SNPs meeting the GWA cut-off for each trait in turn.

We found significantly greater population differentiation at independent loci shared between educational attainment and height. All comparisons (i.e. in both directions) between measures of educational attainment (college and number of years) and height were significant at $p<0.05$. Ten independent loci were observed between educational attainment college and height with sixteen independent loci associated with both number of educational years and height.

Assessing the impact of population stratification at GWA loci

Population stratification in GWA samples poses a concern if there is a match in the discovery GWA study between the direction of phenotypic differentiation among populations and genetic stratification in allele frequency in the 1000 Genomes dataset. Although this confounding effect is typically controlled for in discovery GWA studies, there is no guarantee that all stratification has been removed [31]. In such cases, inadequate control of stratification could lead to GWA variants with high levels of allelic differentiation between populations.

We performed simulations to assess the potential impact of population stratification in the discovery GWA study on our overall results. First, we simulated a polygenic trait using genotype data and use standard protocols for correcting population stratification to examine whether this is sufficient to remove false positives in our results. Next, given that some studies comprised only of European cohorts, we tested whether stratification in the discovery GWA study could systematically bias our results when stratified variants in the European population were projected to the 1000 Genomes population.

Simulation of a polygenic trait using GERA data with and without correction of population stratification

First, we randomly sampled 500 individuals from each of the four self-reported ethnic groups (European, East-Asian, Latin Americans and Africans) in the GERA (Genetic Epidemiology Research on Aging) cohort. The GERA cohort is a subsample of the longitudinal cohort of the Kaiser Permanente Research Program in the Northern California region. High-density genotyping was done with custom arrays for each of the four major ethnic groups. Next, we

randomly assigned 5,000 independent loci across the genome to be causal variants. Their effects were sampled from a normal distribution with a mean of 0 and a variance of 1. The heritability of the trait (h^2) was set at 50%. A phenotype was created as $y = \Sigma\, x_k b + e$, where x_k is the genotype copy at locus k (0,1,2), b the effect size $\frac{h^2}{n}$, where $n = 5{,}000$ and e is the residual variance where $e = N(0, 1 - h^2)$.

We used PLINK 1.9 [48,49] to estimate SNP effects using ordinary least-squares regression without any control for population stratification. We then repeated the GWAS estimation and controlled for stratification using the first three principal components as covariates in the regression. We found that correction for stratification reduced *d* values across the genome (**FIG S7**).

If the predictors are created from SNPs that were unbiased by population stratification, we expect no significant difference in overall measure of phenotype differentiation versus null expectation. This was observed to be the case (**Table S4**).

Studies where population stratification in the mapping panel of discovery GWA study was corrected for using PCs are listed in the Supplemental Materials.

Simulations with highly stratified loci in European populations

We note that the GWA metadata we used are from studies conducted in predominately or exclusively European cohorts. Stratification is potentially confounding if a GWA variant, incorrectly assigned due to stratification in the discovery set, is also highly genetically differentiated in the 1000 Genomes population. We can explicitly test for such a correspondence – between the top PCs from 1000 Genomes and PCs from a European population. If such a relationship exists, our results may indeed be confounded.

To assess the impact of stratification within European populations on 1000 Genomes result, we reran our analyses using highly stratified variants between European populations as our background. Population structure was inferred from European individuals only using a separate PCA. To accurately measure population structure, PCA was performed after removing signals of long-range LD (iterative pruning with r^2 cut-off=0.2), rare alleles (MAF<0.01) and variants that fail Hardy-Weinberg equilibrium test ($p<1\text{x}10^{-6}$). Next, those

SNPs displaying the strongest signals of population stratification between Europeans subpopulations were extracted. To do this, we calculated the dot product of the PC loading and SNP genotype for the first three PCs and selected variants that in the top 200,000 (~2%) of the squared dot product for each of the first three PCs. Our original selection scores from 1000 Genomes were calculated at locations of these variants. Using these stratified SNPs as our background, we reran our analysis by partitioning GWA SNPs into functional regions followed by permutations (as described above) to assess statistical significance.

We find little correspondence between variants explaining the most variation amongst European populations with variants explaining the most variation amongst the main 1000 Genome populations (r=-0.02, first three PCs). Changing the background set of variants to those showing high levels of population differentiation amongst Europeans, while keeping to same set of 'observed' variants did not significantly impact our results (**FIG S8**). Notably, we also obtained highly concordant results between two height GWA studies where in one study [31] effect size estimates were inferred from sib-pairs and is thus robust against stratification (data not shown).

Replication in GERA dataset

We replicated our study on the GERA (Genetic Epidemiology Research on Aging) cohort. We sampled 500 individuals from each of the four ethnic groups: European, East-Asian, Latin Americans and Africans. Analyses were repeated as described above for the 1000 Genomes dataset. Variants not genotyped amongst all selected individuals were filtered out. 6,690,254 variants remained after filtering (MAF>0.01). Following iterative pruning, 366,178 variants were used in PCA to estimate population structure.

We compared the log(observed mean d/expected mean d) for variants classed by GWA phenotype and annotation category between GERA and 1000 Genomes and observed good correspondence between results r=0.63 (Pearson's correlation p-value<2.2×10^{-16}) (**Fig S9**).

References

1. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484:55–61.

2. Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, et al. Natural Selection and Genetic Diversity in the Butterfly Heliconius melpomene. Genetics. 2016;203:525–41.

3. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, et al. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat. Commun. 2016;7:13195.

4. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum. Mol. Genet. 2015;24:R102-110.

5. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010;328:1036–40.

6. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature. 2014;515:365–70.

7. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312:1614–20.

8. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. Annu. Rev. Genet. 2013;47:97–120.

9. Novembre J, Di Rienzo A. Spatial patterns of variation due to natural selection in humans. Nat. Rev. Genet. 2009;10:745–55.

10. Myles S, Tang K, Somel M, Green RE, Kelso J, Stoneking M. Identification and analysis of genomic regions with large between-population differentiation in humans. Ann. Hum. Genet. 2008;72:99–110.

11. Weir BS. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 2005;15:1468–76.

12. McEvoy BP, Powell JE, Goddard ME, Visscher PM. Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 2011;21:821–9.

13. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. CB. 2010;20:R208-215.

14. Ackermann RR, Cheverud JM. Detecting genetic drift versus selection in human evolution. Proc. Natl. Acad. Sci. 2004;101:17946–51.

15. Smith HF. The Role of Genetic Drift in Shaping Modern Human Cranial Evolution: A Test Using Microevolutionary Modeling. Int. J. Evol. Biol. 2011;2011:1–11.

16. Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. Mol. Biol. Evol. 2015;msv334.

17. Galinsky KJ, Bhatia G, Loh P-R, Georgiev S, Mukherjee S, Patterson NJ, et al. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet. 2016;98:456–72.

18. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307:1072–9.

19. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 2015;47:1228–35.

20. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41:827–41.

21. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat. Rev. Genet. 2012;13:505–16.

22. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011;478:343–8.

23. Prabhakar S, Noonan JP, Paabo S, Rubin EM. Accelerated Evolution of Conserved Noncoding Sequences in Humans. Science. 2006;314:786–786.

24. Zhang YE, Landback P, Vibranovski MD, Long M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. Wolfe KH, editor. PLoS Biol. 2011;9:e1001179.

25. Drevets WC, Savitz J, Trimble M. The Subgenual Anterior Cingulate Cortex in Mood Disorders. CNS Spectr. 2008;13:663–81.

26. Ivleva EI, Clementz BA, Dutcher AM, Arnold SJM, Jeon-Slaughter H, Aslan S, et al. Brain Structure Biomarkers in the Psychosis Biotypes: Findings From the Bipolar-Schizophrenia Network for Intermediate Phenotypes. Biol. Psychiatry. 2016;

27. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 2010;42:565–9.

28. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

29. Trowsdale J. The MHC, disease and selection. Immunol. Lett. 2011;137:1–8.

30. CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523:588–91.

31. Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, et al. Population genetic differentiation of height and body mass index across Europe. Nat. Genet. 2015;47:1357–62.

32. Klimentidis YC, Abrams M, Wang J, Fernandez JR, Allison DB. Natural selection at genomic regions associated with obesity and type-2 diabetes: East Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. Hum. Genet. 2011;129:407–18.

33. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009;19:826–37.

34. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 2015;47:291–5.

35. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419:832–7.

36. Rockman MV. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evol. Int. J. Org. Evol. 2012;66:1–17.

37. Pritchard JK, Di Rienzo A. Adaptation - not by sweeps alone. Nat. Rev. Genet. 2010;11:665–7.

38. Gratten J, Visscher PM. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome Med. [Internet]. 2016 [cited 2016 Sep 22];8. Available from: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0332-x

39. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. 2016;48:709–17.

40. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528:499–503.

41. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005;3:e170.

42. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15:1566–75.

43. de Candia TR, Lee SH, Yang J, Browning BL, Gejman PV, Levinson DF, et al. Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent. Am. J. Hum. Genet. 2013;93:463–70.

44. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016;533:539–42.

45. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 2014;46:1173–86.

46. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 2010;11:459–63.

47. Halko N, Martinsson PG, Tropp JA. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Rev. 2011;53:217–88.

48. Purcell S, Chang C. PLINK [Internet]. Available from: https://www.cog-genomics.org/plink2

49. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience [Internet]. 2015 [cited 2017 Jan 15];4. Available from: http://gigascience.biomedcentral.com/articles/10.1186/s13742-015-0047-8

50. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 2010;28:495–501.

Data accession

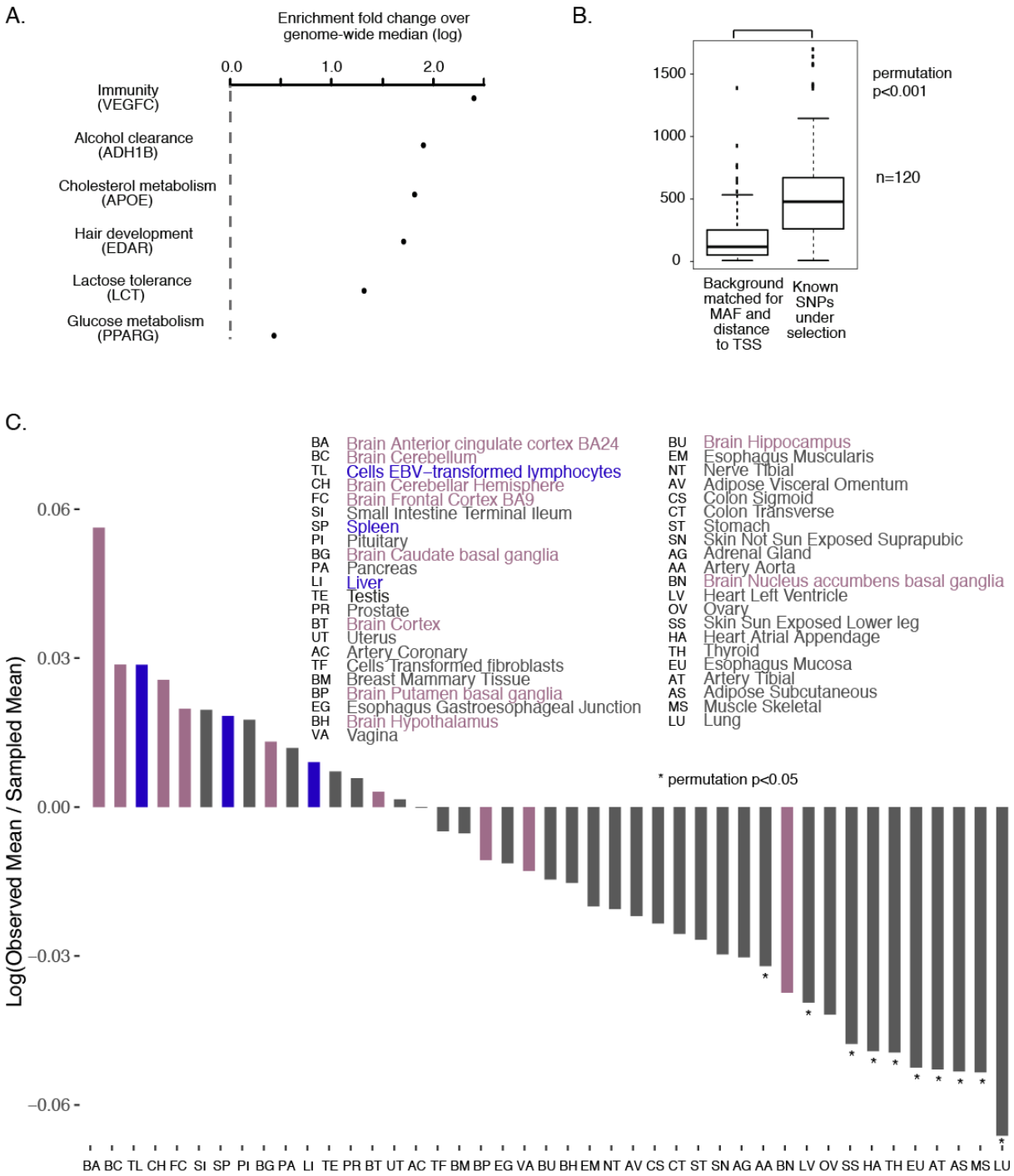
The GERA dataset was obtained from the database of Genotypes and Phenotypes (dbGaP) through accession phs000674.v1.p1.

Contributions

ESW and JEP designed experiments. ESW performed analyses and wrote the manuscript. JEP edited the manuscript and supplied data.

Acknowledgements

We would like to thank Matthew Robinson for his advice and input on use of the differentiation score. We would also like to thank Paul Flicek and Greg Gibson for valuable discussion and feedback on the manuscript. ESW is supported by EMBO Advanced fellowship (1672-2014) and Australian Research Council Discovery Early Career Award (DE160100755). JEP is supported by National Health and Medical Research Council Career Development Fellowship (1107599) and grant (1083405). The authors have no completing interests.



5

Figure 1. Allelic differentiation between populations at tissue-specific expression quantitative trait loci is highly variable between tissues

(A) High levels of population differentiation can suggest positive selection driving changes in allele frequencies. Enrichment of *d* over genome-wide median is observed at well-known

examples of loci that have undergone positive selection. **(B)** Increased *d* was observed at 120 loci previously reported to show high F_{ST} between populations [10] **(C)** 1000 Genome variants were overlaid with eQTL SNPs from the Genotype-Tissue Expression dataset (GTEx) comprising for eQTLs from 44 tissues. An average *d* value was calculated on a tissue-specific basis. Permutations were used to assess the significance of the mean d value by sampling an equivalent number of eQTLs across the genome matched for MAF and distance to TSS. This procedure was repeated 1000 times to calculated an empirical two-tailed *p*-value. Brain tissues are shown in purple with immune related tissues and cells in in blue.

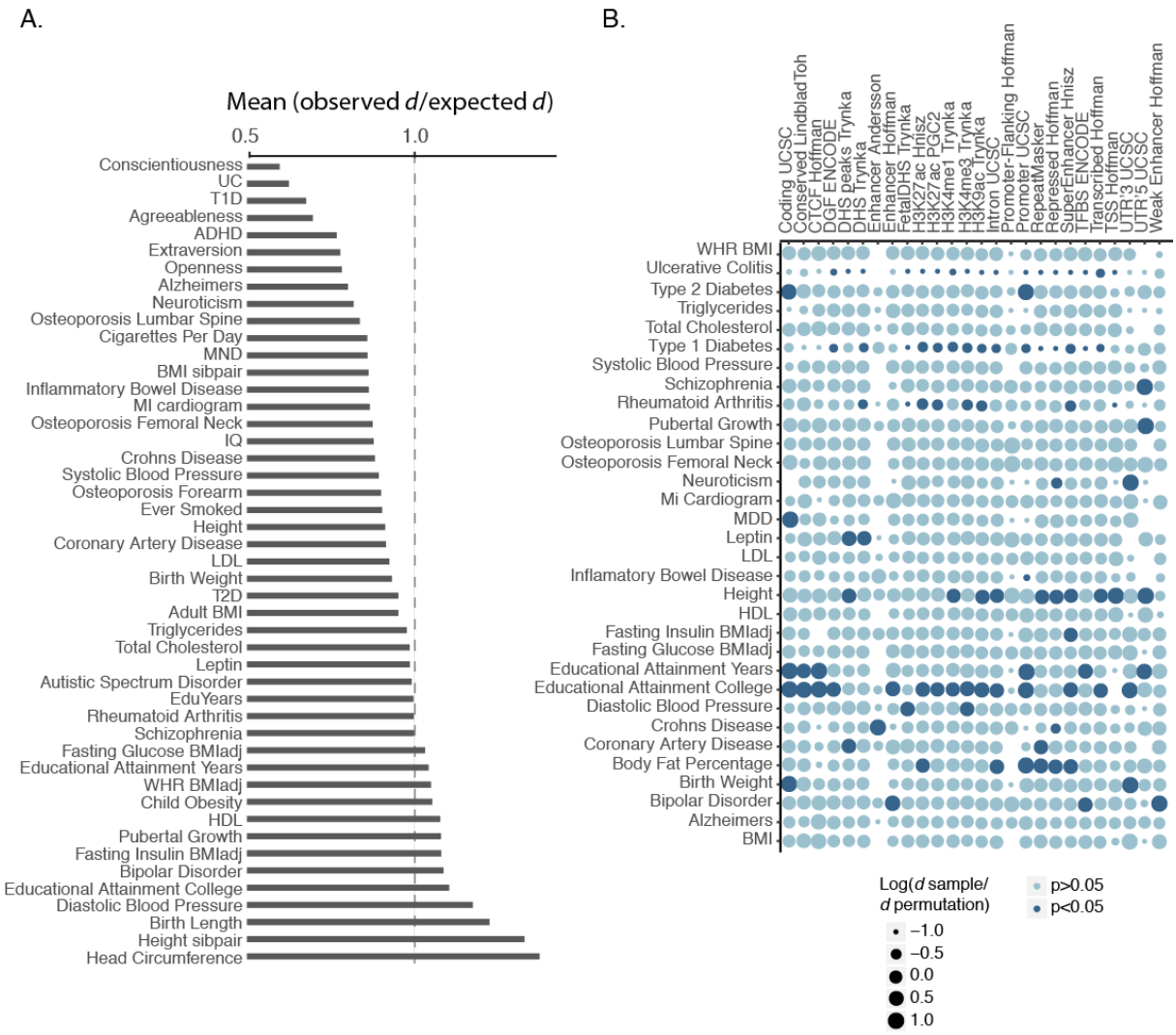


Figure 2. Striking disparities in population differentiation at loci for various complex traits indicating distinct evolutionary trajectories

(A) Due to the polygenic nature of complex phenotypes, we quantitated allelic differentiation for all SNPs with GWA *p*-values below $1x10^{-4}$. The mean(expected d/observed d) was calculated for each phenotype as follows. 1) We calculated the genetic differentiation at SNPs GWA $p<1x10^{-4}$ where n>100 (observed *d*). 2) For each SNP in Step 1, we sampled 1000 SNPs with replacement from a pool of SNPs matched for MAF and distance to TSS bins. We took the average d of these SNPs (expected *d*). 3) For each SNP in Step 1 with take the observed *d* divided with its matched expected *d* 4) We obtain the average across SNPs i.e. all values in Step 3. **(B)** We partitioned our genome-wide selection score by functional annotations values to account for differences in evolutionary rates among different genomic regions [19]. At each functional annotation region for each trait, we LD pruned SNPs (*p*-values<$1x10^{-}4$, *r2* cutoff=0.2) and calculated the mean *d* score for variants in linkage

equilibrium. We performed 1000 permutations to estimate the significance of our results by resampling SNPs from the same functional category. Our empirical null distribution is therefore generated per functional category per trait based on resampling an equivalent number of SNPs to the number of pruned SNPs per iteration. The visualizations here depict the log of the median *d* across all SNPs for a functional region divided by the median of medians of sampled values. The size of the dots shows enrichment (>0) or depletion (<0) of this ratio. Dot colour denotes statistical significance based on the permutation.

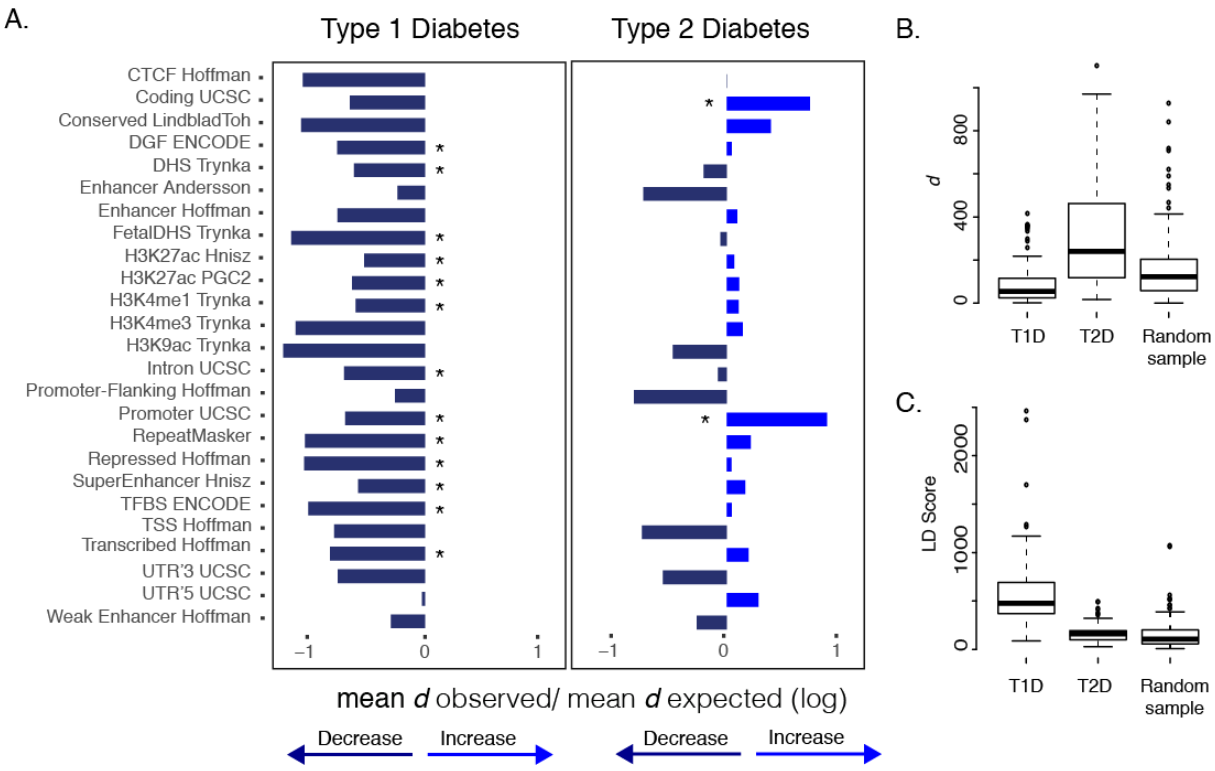


Figure 3. Distinct patterns of allelic differentiation at loci associated with type I versus type 2 diabetes

(A) T1D loci show reduced population differentiation compared to T2D. The x-axis depicts the log of the mean *d* across all SNPs for a functional region (mean *d* observed) divided by the mean of median sampled null values (mean *d* expected). The calculation of mean *d* expected is detailed below. We calculated a two-tailed empirical *p*-value by LD-pruning (r^2 cutoff=0.2) those SNPs where GWA $p<1\text{x}10^{-4}$ and n>100. Next, for each SNP, we performed 1000 permutations to estimate statistical significance by resampling matched number of SNPs from the same functional category. Specifically, mean *d* expected = $\sum_i(\text{median}(\text{null values across 1000 permutations}))\,/n$, where i = each observed SNP and n = number of observed SNPs. Starred functional categories denote a permutation $p<0.05$. **(B)** Boxplot depicting *d* values for individual variants associated with T1D, T2D and a random sample of variants from across the genome **(C)** Boxplot depicting LD scores for individual variants associated with T1D, T2D and a random sample of variants from across the genome.

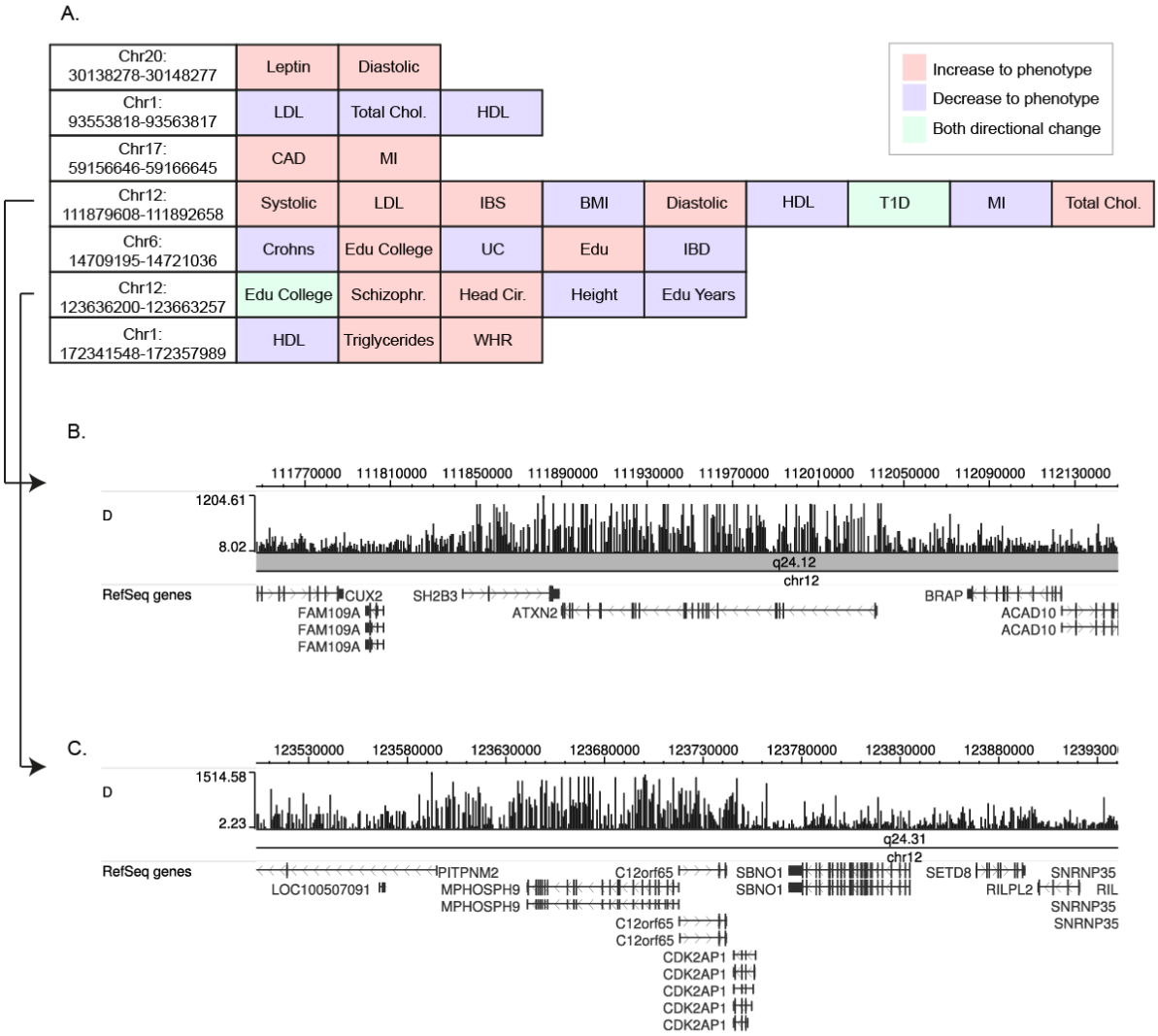


Figure 4. Opposing allelic directions are common at shared loci and a small proportion of these show evidence of increased populations differentiation

(A) Regions of shared loci were clustered and departure of the mean *d* from background was calculated**.** We clustered proximal GWA variants (*p*-value cut-off of $1x10^{-4}$) based on genomic location with a flanking region of 5kb added to each side. Overlapping regions were merged and the mean *d* of all variants falling within each region was calculated. For each clustered region (where there are more that two traits associated; n=2649), we tested for departures of the averaged *d* against a background of averaged variants at tiled genomic regions the same genomic length as that of the clustered region. Significance (two-tailed) was calculated by comparing the mean *d* value for each cluster by sampling from its null background for 1000 permutations. Shown loci possess a permutation p-value below 0.05. Shared loci between GWA phenotypes often show opposing allelic direction between disease

traits. **(B)** A highly pleiotropic loci located on chromosome 12 associated to many lipid-related shows increased genetic differentiation amongst populations and harbors two tightly linked genes – *SH2B3*, involved in signaling and immune function, and *ATXN2*, associated with multiple neurodegenerative and neuromuscular disorders **(C)** High differentiation was also observed at a second highly pleiotropic loci located on chromosome 12 linked to educational attainment.