A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood

Jing Zhao,[1] Idowu Akinsanmi,[1] Dalia Arafat,[1] T.J. Cradick,[2] Ciaran M. Lee,[2,3] Samridhi Banskota,[1] Urko M. Marigorta,[1] Gang Bao,[2,3] and Greg Gibson[1,*]

In order to evaluate whether rare regulatory variants in the vicinity of promoters are likely to impact gene expression, we conducted a novel burden test for enrichment of rare variants at the extremes of expression. After sequencing 2-kb promoter regions of 472 genes in 410 healthy adults, we performed a quadratic regression of rare variant count on bins of peripheral blood transcript abundance from microarrays, summing over ranks of all genes. After adjusting for common eQTLs and the major axes of gene expression covariance, a highly significant excess of variants with minor allele frequency less than 0.05 at both high and low extremes across individuals was observed. Further enrichment was seen in sites annotated as potentially regulatory by RegulomeDB, but a deficit of effects was associated with known metabolic disease genes. The main result replicates in an independent sample of 75 individuals with RNA-seq and whole-genome sequence information. Three of four predicted large-effect sites were validated by CRISPR/Cas9 knockdown in K562 cells, but simulations indicate that effect sizes need not be unusually large to produce the observed burden. Unusually divergent low-frequency promoter haplotypes were observed at 31 loci, at least 9 of which appear to be derived from Neandertal admixture, but these were not associated with divergent gene expression in blood. The overall burden test results are consistent with rare and private regulatory variants driving high or low transcription at specific loci, potentially contributing to disease.

Introduction

In recent years, whole-exome sequencing has been used effectively to demonstrate that there is a burden of rare coding variants in individuals with a variety of neurological and developmental conditions.[1–4] Considering estimates that as many as 90% of disease-associated common variants are regulatory rather than structural,[5–7] it is reasonable to assume that rare regulatory variants influencing the expression of causal genes might also be enriched in individuals with congenital abnormalities or common chronic diseases. Here we demonstrate that there is a burden of rare variants with gene expression itself, focusing on just the promoter regions of a targeted set of genes whose expression was measured by microarray analysis of peripheral blood samples.

Our strategy, outlined in Figure 1, gains statistical power by pooling rare variant enrichments across the full range of expression of 472 genes measured in 410 individuals. This effectively generates almost 200,000 data points, but instead of focusing on just the most extreme individuals as required by burden tests designed for case-control comparisons,[8–11] we evaluate the shape of the distribution of cumulative counts of rare variants in equal sized bins of expression. For each gene in each individual, 2 kb of DNA sequence flanking the annotated transcription start site was sequenced after targeted capture of genomic DNA on custom beads.[12] The count of rare variants with minor allele frequency less than 5% (or 1%) was assessed after alignment to the HuRef19 reference human genome with the Unified Genotyper in GATK.[13] These counts were summed for 82 equal sized successive gene expression bins with 5 individuals each, and then tallied for all 472 genes.

Under the null hypothesis, there should be no relationship between rare variant count and gene expression and a plot of rare variant count on the y axis against expression bin on the x axis should yield a horizontal regression line. In the presence of rare variants that decrease expression, there should be larger counts in the low expression bins, toward the left in the plots in Figure 2, and similarly rare variants that increase expression should yield larger counts in the higher expression bins to the right. A general bias toward either effect would result in a significant linear slope term in a regression model. However, if both effects are present, a characteristic "smile" plot would ensue, the significance of which would be reflected in the quadratic term of a regression. We further assessed departure from the null by evaluating the significance of the complete quadratic model relative to 10,000 permutations of the full genotype and gene expression matrices, subsequently adjusting for various covariates to gain further insight into the nature of the burden of rare variants for extreme expression.

Although deleterious coding variants are generally loss of function, deleterious regulatory variants have similar a priori probabilities of increasing or decreasing transcript abundance. In fact, evolutionary studies[14,15] imply that gene expression is generally subject to moderate stabilizing selection, which acts to maintain transcript levels close to an optimal level. Thus, it is unlikely that many large effect mutations remain in the gene pool for extended periods of

[1]School of Biology and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA; [2]Wallace H. Coulter Department of Biomedical Engineering, Laboratory of Biomolecular Engineering and Nanomedicine, Georgia Institute of Technology, Atlanta, GA 30332, USA; [3]Department of Bioengineering, Rice University, Texas Medical Center, Houston, TX 77030, USA
*Correspondence: greg.gibson@biology.gatech.edu
http://dx.doi.org/10.1016/j.ajhg.2015.12.023. ©2016 by The American Society of Human Genetics. All rights reserved.

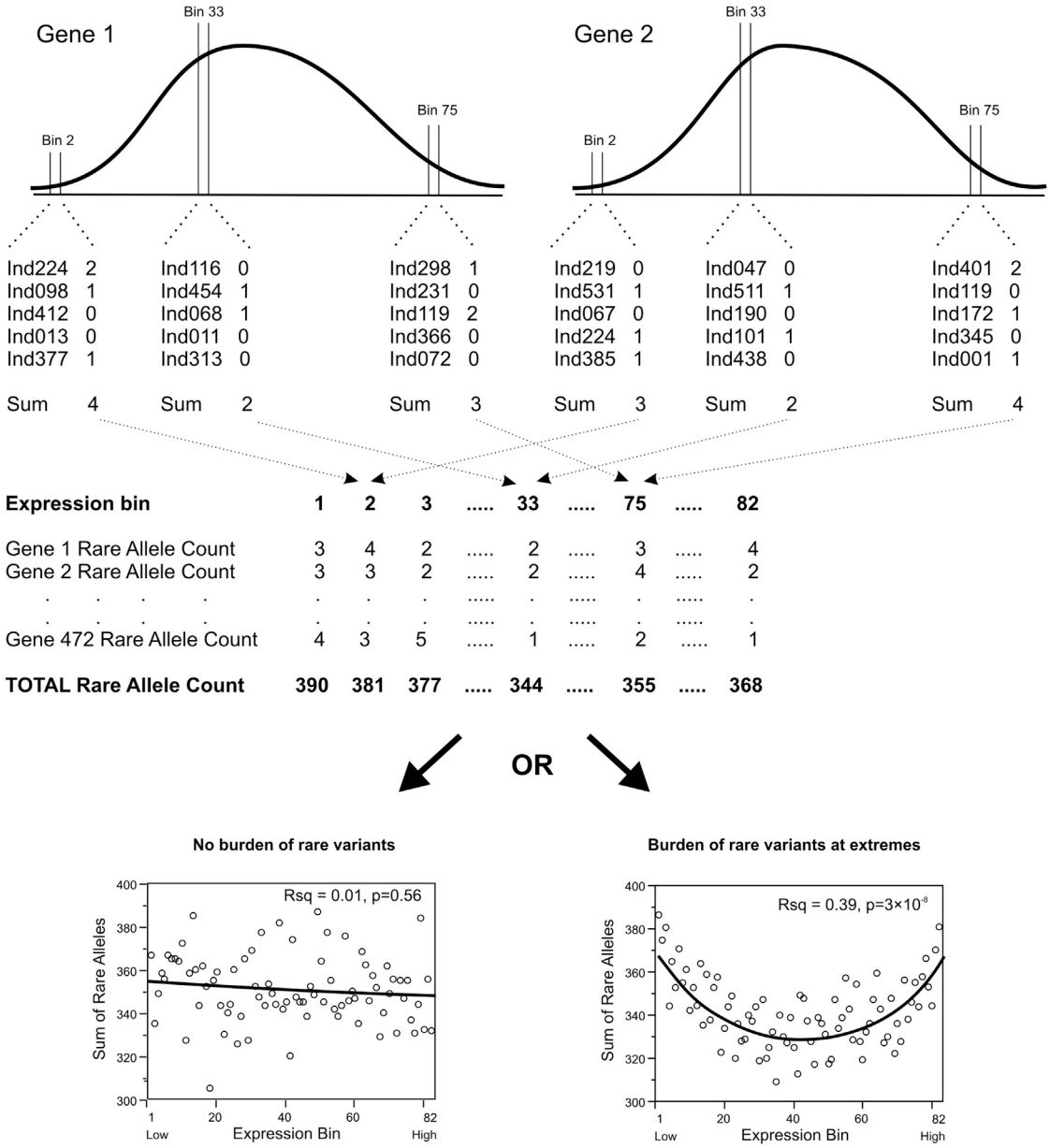

Figure 1. Schema Showing the Pooling Strategy to Evaluate Rare Variant Enrichment
For each gene, the normalized gene expression measures across all 410 individuals are sorted into 82 bins, resulting in somewhat normal frequency distributions shown in the top panels. Subsequently, the number of rare variants in the 2-kb promoter of each allele in that bin is tallied: for example, there are 2, 1, 0, 0, and 1 rare variants in the promoters of the 5 individuals (both alleles) in the second bin for gene 1, summing to 4, whereas the second bin for gene 2 has 3 rare variants. These expression bin rare allele counts are then summed over all 472 genes and plotted from lowest to highest bin to yield plots at the bottom of the figure that represent two alternative results. In the absence of a burden of rare variants at the extremes, there is neither a significant slope nor quadratic fit (left plot), whereas an excess of variants at both extremes produces a concave "smile" regression (right plot). If there were an excess at only the low or high expression, the linear slope would be significant.

time, because purifying selection ensures that only moderate effect alleles persist. One class of haplotype that is particularly interesting in this regard is the few percent of alleles that can be traced to introgression from Neandertal or Denisovan populations.[16–18] Although divergent at the genotype level and occasionally associated with extreme traits such as high-altitude adaptation,[19] the maintenance of such haplotypes over several tens of thousands of generations leads to the expectation that they are not likely to collectively influence gene expression, a proposition that we also test here.

Material and Methods

Targeted Promoter Sequencing

DNA was obtained from 410 participants in the Atlanta CHDWB study,[20,21] under approval of the Emory University and Georgia

Tech IRBs for genetic studies, and including written informed consent. The 410 samples comprised 274 females and 136 males, and included 297 individuals of European ancestry, 95 of African ancestry, and 18 of Asian ancestry. The age at entry into the program and initial sampling of blood spanned from 19 to 83 with a mean of 50.

A set of 500 genes was initially chosen for targeted promoter sequencing, but 28 of these located in the HLA complex were removed due to irregular read counts in the complex, resulting in a set of 472 genes that were included in the study. The full list is included as Table S1, and metrics of sequence quality along with a list of all promoter variants is available as additional material at the author's website (see Web Resources). Among the 472 genes, 207 have common *cis*-eQTL, 244 are represented on the Metabochip, and 220 are represented on the Immunochip, 145 of which are on both chips.

Whole-genomic DNA was isolated from buffy coats using Flexigene DNA kits (QIAGEN). The location of the major transcription start site (TSS) of each gene was extracted from the UCSC Genome Browser (accessed May 2012), and oligonucleotide probes were designed with the Illumina Design Studio so as to pull down 1 kb upstream and 1 kb downstream of the major TSS for each of the 472 genes. 50-mer oligonucleotide probes were designed to ensure that the percentage of the total length of all regions targeted for enrichment was not less than 90%. Sequence capture libraries were generated and pooled with Illumina TruSeq DNA Sample Preparation Kits and TruSeq Custom Enrichment Kits. Paired end 100 bp sequencing was then performed on an Illumina HiSeq 2500 at Georgia Tech. Approximately 75% of the aligned reads with BWA mapped to the 2-kb promoter regions of the 472 genes, indicating good enrichment to the targeted regions. The average read depth across the dataset was more than 600×, with more than 90% of the reads in the 2-kb promoter regions having more than 20× read depth.

Variant Calling

After short read alignment, a variety of different strategies for variant calling were evaluated and contrasted, leading to the decision to use the output of the GATK Unified Genotyper algorithm[13] applied to all samples in a pooled analysis. First, the BWA aligner was used to align fastq files to HuRef19. The total number of aligned reads per sample ranged from approximately 2 million to 40 million, with a mean of 14 million and standard deviation of 5 million. The mapped reads proportion was between 94.1% and 99.9% with a mean of 99.3% and a standard deviation of 1.05%. The percentage of concordantly paired reads ranged from 92.9% to 98.9% with a mean of 97.9% and a standard deviation of 1.2%. The GATK VQSR tool was used for variant filtering using the Illumina Omni chip array based on the 1000 Genomes Project as the training data with the highest-confidence SNPs from the 1000 Genomes Project's call set used to validate the SNPs, utilizing a machine learning approach to optimize cutoffs for QD (quality by depth ratio), FS (Fisher's exact test of strand bias), MQ (mapping quality metric), HaplotypeScore, MQRankSum (Mann-Whitney rank sum test for mapping qualities), and ReadPosRankSum (Mann-Whitney rank sum test for the distance of reads with the alternate allele to the end of the read).

After variant calling, we detected 17,584 raw SNPs in total, but these were reduced to 10,451 SNPs passing the filters, that lie within the 2-kb promoter regions of 472 genes. 8,833 of the SNPs are rare (defined as MAF < 0.05 in our dataset, which is concordant with 1000 Genomes frequencies) and 1,618 are common (with MAF ≥ 0.05), which averages 1.5 rare variants per promoter per individual. Approximately 60% of these rare variants are private, meaning that they were observed in only a single individual. The number of rare, common, and private SNPs per gene, along with an estimate of the polymorphism rate (π) per gene in the 2-kb sequenced region, as well as a matrix of rare allele counts per gene in each individual are available at the author's website.

Verification by Sanger Sequencing and Genotyping

To verify the accuracy of the high-throughput sequencing, we Sanger sequenced 500 bp segments of two genes, *TRAF3IP3* (OMIM: 608255) and *HSPA8* (OMIM: 600816). The sequenced region of TRAF3IP3 was chr1: 209,929,132–209,929,708, in which two rare SNPs and four common SNPs were observed in the 410 samples. All of these were included in 96 samples that were Sanger sequenced, and all were validated. The sequenced region of *HSPA8* was chr11: 122,932,665–122,933,158, which contained 18 rare and 5 common SNPs in the 96 sequenced samples, which were again verified by Sanger sequencing. A handful of other individuals were nominally positive at some of the rare sites, but manual inspection of the traces revealed poor-quality sequence toward the ends of the reads in those individuals suggesting a false positive rate that in any case would be less than 0.5%. No other variants that were not present in the GATK analysis were called with high confidence by Sanger sequencing, whereas all common variants were also validated by the Sanger sequencing.

Whole-genome genotypes either from Illumina OmniExpress or Core+Exome arrays, imputed onto 1000 Genomes with Impute2,[22] were also available for the majority of individuals. Extremely high concordance was observed. These genotypes were thus used for common variant eQTL analysis, which will be described in detail elsewhere. We also interrogated whether rare SNPs lie within the Illumina probes, but found just three examples in two genes, so these do not explain enrichment of rare variants downstream of the promoter with gene expression.

Gene Expression Profiling

Transcript abundance measures were generated in two batches using Illumina-HT12 human gene expression arrays. RNA was prepared from whole-blood samples collected and stored in Tempus tubes (Life Technologies), according to manufacturer-recommended protocols, and quality was confirmed with an Agilent Bioanalyzer such that all samples had RIN numbers greater than 8. The first batch of samples was processed for hybridization and bead intensity extraction by Expression Analysis and the second by HudsonAlpha. The raw data are available at the Gene Expression Omnibus (GEO) but additional data processing steps were employed for this study to account for batch effects that might have skewed the rare variant association statistics.

Raw expression data in the form of average bead intensities from the Illumina Genome Studio were first transformed to log2 values and then processed with two standard approaches for removing surrogate variables. For PEER analysis[23] of the Europeans only (Figure S1A), batch effects were removed with ComBat,[24] and then age and gender were fit in a general linear model, before fitting the PEER algorithm with 20 factors selected, 6 of which were notably stronger than the remainder. As an additional mode of analysis, surrogate variable analysis (SVA),[25] we considered the full dataset, after removing batch effects with COMBAT, then used SVA with age as the biological variable, fitting

a single surrogate variable identified by the open source R code. Because ethnicity still explained 9% of the variance and age and gender each approximately 0.5%, we removed these as linear terms, then extracted just the 279 European-ancestry samples for rare variant enrichment tests as described below, generating very similar results as with PEER. However, for most analyses reported below, we used normalization based on supervised normalization of microarrays (SNM) algorithm in R,[26] without fitting the surrogate variables. We fit age as the biological variable and removed effects of batch and ethnicity by including these as adjustment variables with the rm = True option. Individual effects were accounted for as the intensity-dependent variable.

We next extracted the 472 genes for which we have promoter genotypes and averaged the estimates for 172 genes that are represented by two (132 genes) or more (40 genes) probes in the Illumina-HT12 arrays. Each gene expression distribution was then transformed to the same scale, namely to z-scores, which are standard normal distributions with a mean of 0 and standard deviation of 1. To ensure that there was no overall batch effect on the variances (namely, that individuals from one batch are not, for technical reasons, more likely to have extreme values), we fit the z-scores by batch and combined them into a single gene expression dataset that was used to generate all of the results reported here. Quantile normalization[27] was also performed in parallel (Figure S1B) because it is commonly used in the literature. Quantile normalization ensures that the distribution of abundance estimates of the entire gene expression profile of each individual is the same but does not ensure adjustment of covariates influencing the variance (or average expression) of each gene. It was applied to the raw log2 distributions of each individual, and probes for the same gene were again averaged prior to assigning genes to expression bins for the regression on rare variant counts.

Rare Variant Burden Test

In order to evaluate whether there was a relationship between transcript abundance and number of rare variants in the promoter, for each gene, each individual was placed in one of 82 equal-sized bins of five individuals based on the rank of the batch-adjusted z-scores. We then tallied the number of rare variants in the promoter regions of those five individuals and subsequently summed the rare variant counts across all 472 genes to achieve statistical power to detect the overall burden. With a MAF cutoff of 0.05, only 1 homozygote is expected per gene, but because most rare variants are rarer than this, the actual number of homozygotes is too small to impact the sums, but they were counted twice in the tally at the gene level. The bin size was chosen as a compromise between smooth fitting of the quadratic regression and compensating for noise in individual gene expression measures assessed by microarray. However, additional analyses were performed with 41 bins of size 10 individuals (Figure S1E) or with MAF < 0.01 (Figure S1F), neither of which had a meaningful effect on the conclusions because both remain highly significant for enrichment at both extremes. We then evaluated the deviation of the distribution from the null hypothesis of no relationship by fitting a quadratic model where the linear term captures bias toward enrichment for either higher or lower expression and the quadratic term captures the effect of bias at both extremes simultaneously.

The significance values of the two terms were observed to be very similar to the empirical p values obtained by permuting the sum counts against the bin number. We also performed a more robust permutation to shuffle the genotype and gene expression vectors, keeping the full vector of promoter counts within each individual (and the full vector of expression ranks) constant so as to preserve any biological covariance. With the appropriately normalized gene expression data, such permutations generally resulted in flat regressions of allele count on expression bin, with non-significant linear and quadratic terms. We then evaluated the significance of the actual data by documenting how many permutations out of 10,000 have a more significant overall model fit, which is just a few cases, strengthening support for the inference (1) that the normalization has removed systematic biases and (2) that there is a true burden of rare variants at either extreme of the transcript distribution, averaging across 472 transcripts.

A further adjustment was made to account for unequal total read counts among individuals or in specific genes as follows. For the analyses involving mixed races, we performed a haplotype burden analysis by collapsing all multi-SNP promoters down to a count of 1, instead of the actual number of rare variants. This should be conservative because it will tend to underestimate the contributions of two or more variants in a single promoter. We also fit a "joint" analysis (Figure S1C) where instead of binning gene expression solely within individuals, we generated 82 bins of 2,100 gene expression measures (5 × 420), based in the ranked z-scores of all genes in all individuals, and performed the regression on the summed allele counts associated with the 2,100 measures. In this procedure, each gene contributes slightly disproportionately to each bin, yet the overall result was again retained.

Adjustment of Rare Variant Burden Test for Covariates

A variety of biological factors could mask the effect of rare variants by increasing the variance of gene expression. Two obvious effects are the contribution of common eQTLs, which will tend to cause individuals with the less active polymorphism to be in lower expression bins, and *trans*-acting sources of gene expression covariance. Because peripheral blood preserved in Tempus tubes is a complex mixture of leukocytes (residual red blood cell and platelet gene expression is thought not to contribute strongly to observed transcript abundance), an obvious source of covariance is cell counts. Cell counts (lymphocytes, monocytes, neutrophils, erythrocytes, and platelets) explain on average just 6.0% of the variance of each transcript abundance measure in our dataset, which is actually one quarter of the amount explained by seven empirically determined common axes of covariance (average 23.8%: compare Figures S2A and S2B for distributions of variance explained). These axes probably reflect a mixture of the contributions of cell counts and coordinate gene regulation for example by interferon or other systemic factors. The seven axes are defined by the first principal component of the expression of ten "blood informative transcripts (BITs)" per axis, where the BITs have been defined by comparison of multiple blood gene expression datasets.[28] Note that fitting PC1 to the 5 or 100 most correlated transcripts in each axis results in almost identical axis scores.

Simple linear regression was used to fit both eQTL and co-expression axes. For the eQTL adjustment, we first performed whole-genome *cis*-eQTL analysis on the full CHDWB dataset and identified significant eQTLs located within 5 kb of the TSS or in the gene body for 207 of the 472 genes at $p < 10^{-4}$, observing more than 70% overlap with the blood eQTL browser variants derived from meta-analysis of more than 5,000 samples.[29] For approximately one fifth of the genes, multiple additional *cis*-eQTLs were observed conditioned on the primary eQTL. We used stepwise linear regression to fit these empirical eQTLs in

our dataset, also including age and sex as covariates in the model (although neither age nor sex account for more than a few percent of the variance of any of the 472 genes). The residuals from the eQTL fit were then used for axis adjustment, for which we computed the seven PC1 scores from the full SNM normalized gene expression matrix, and then identified which axis was most strongly correlated with the expression of each gene (8,654 were influenced by an axis at $p < 10^{-5}$, including 398 of the 472 genes included in the enrichment analyses). Univariate linear regression was then used to fit the relevant axis for each gene, yielding gene expression residuals that were taken forward to the adjusted burden test.

Partitioning the Sources of Rare Variant Burden on Gene Expression

Several potential modifiers of the rare variant contribution were evaluated by dividing the total European ancestry dataset into subsets and comparing the model fit. For example, to evaluate whether suspected regulatory sites are more likely to harbor rare variants, we downloaded the RegulomeDB[30] assignments for each SNP and contrasted sites with scores in the ranges 1–4 (likely regulatory) or 5–7 (weak or no evidence for regulatory potential). Similar analyses reported in Table S1 contrast SNPs upstream and downstream of the TSS; SNPs in genes in the upper or lower halves of the overall average transcript abundance spectrum; SNPs in genes in the upper or lower halves of the average promoter polymorphism distribution; SNPs in genes with or without common eQTLs; and SNPs in genes represented on the Metabochip,[31] Immunochip,[32] or neither. For each comparison, we report the significance of the quadratic term and the linear term as well as the overall model fit alongside the average number of rare variants per gene in the two samples being compared.

Assessment of Archaic Origins of Haplotypes

Because 289 of the promoter sequences in the European ancestry samples (0.2%) have three or more substitutions relative to the reference human genome, we asked whether they might be derived from archaic genomes. For each of the 472 genes, vcftools v.3.0[33] was used to query the online sequences of the Neandertal individual from an Altai Mountain cave[34] and the Denisovan individual from the same cave.[35] 62 of the genes were not covered in the sequence. For the remaining 410 genes, we identified 195 positions where one of the 7,779 polymorphisms in our European individual promoter sequence set matched either of the archaic genomes rather than HuRef19. 75 genes had haplotypes with 3 or more sites different from HuRef19; of these, 18 had haplotypes matching a Neandertal haplotype. 31 genes had haplotypes with 4 or more sites different from HuRef19; of these, 8 had haplotypes matching a Neandertal haplotype. In several cases, a Denisovan haplotype was similar to the Neandertal one, but the Neandertal matched the human sequences, so as expected all archaic alleles identified in our dataset are most likely of Neandertal origin. Table S2 shows the genotypes of the 22 individuals with archaic haplotypes in the 8 genes with 4 or more divergent sites.

The expected proportion of Neandertal alleles genome-wide is between 1% and 2%, which is an order of magnitude greater than the proportion observed in our data. A very conservative lower bound on the proportion of divergent Neandertal promoters is 67 of 289 haplotypes, or 0.06% of all promoters. Including promoters with just 2 non-reference sites, approximately 10% of which match a Neandertal allele, this proportion rises to 0.1%. Single divergent sites at least double this proportion again. But owing to the possibility of recurrent mutation and recombination, we chose not to include such sites in analysis of association of Neandertal alleles with gene expression, because they need to be assessed by reference to long-range archaic haplotype blocks. It should also be noted that the Altai individual represents only a fraction of all archaic polymorphisms. Our observed proportion of Neandertal-derived alleles in the divergent promoters, at least 20%, can be considered to be within the expected range given that some alleles are also related to rare haplotypes observed in the African-ancestry individuals and might have been retained from the out-of-Africa dispersal.

Replication Dataset

In order to replicate the rare variant enrichment on a completely independent dataset generated with different gene expression and genotyping technologies, we identified a small cohort of 75 individuals with whole-genome sequence and whole-blood RNA-seq at the Duke Center for Human Genome Variation (D.B. Goldstein, P.I.). Most of these individuals are from a schizophrenia study. Permission to perform genetic analysis was obtained under IRB approval of Duke University, affirmed by the Georgia Tech IRB, and written informed consent was obtained from all study subjects, their parent/guardian, or legally authorized representative. Analysis of the principal components of the genotypes indicated that the sample includes 49 individuals of African ancestry, 24 of European ancestry, and 2 of Asian ancestry.

Whole-genome sequences were obtained on Illumina HiSeq2000 automated DNA sequencers and genotypes were called individually with the GATK algorithm. RNA-seq of whole blood preserved in Tempus tubes was performed also by paired end 100 bp sequencing on the Illumina platform. Raw read counts were log2 transformed and mean centered, and linear regression fitting each of the seven axes of variation (represented by PC1 of the blood informative transcripts) as well as the overall PC1 of gene expression variation (which is correlated with genetically determined ancestry). Subsequently, we assigned the rank of each gene in each individual and performed quadratic regression of the total number of individuals with at least one rare allele count for the gene (MAF < 0.05) in each of 75 ranks. That is, rather than pooling five individuals per bin, the analysis was essentially on bin sizes of 1, necessitated by the small sample of individuals.

Experimental Validation of SNP Effects by Genome Editing

Four SNPs were chosen for experimental validation by CRISPR/Cas9-mediated genome editing.[36] Two (chr6: 24,667,167 in *TDP2* [OMIM: 605764] and chr20: 33,999,719 in *UQCC1* [OMIM: 611797]) were associated with loss of gene expression and two (chr5: 54,603,837 in *DHX29* [OMIM: 612720] and rs182080358 in *COMMD4* [HGNC: 26027]) with gain of gene expression in the CHDWB targeted sequencing analysis (Table S3). These four sites were all present as private alleles in one individual and were in the outlier set for an effect size greater than 2 SDs, visible in Figure 2A. For each SNP, we generated 11 or 12 independent approximately 10-cell K562 clones targeted by guide RNAs. Although K562 cells are erythroleukemic, rather than lymphoid or myeloid, because the variants are promoter proximal, we reasoned that they might have effects generally on transcript abundance and this cell line is well established for CRISPR experiments. Each promoter region in the cell line was

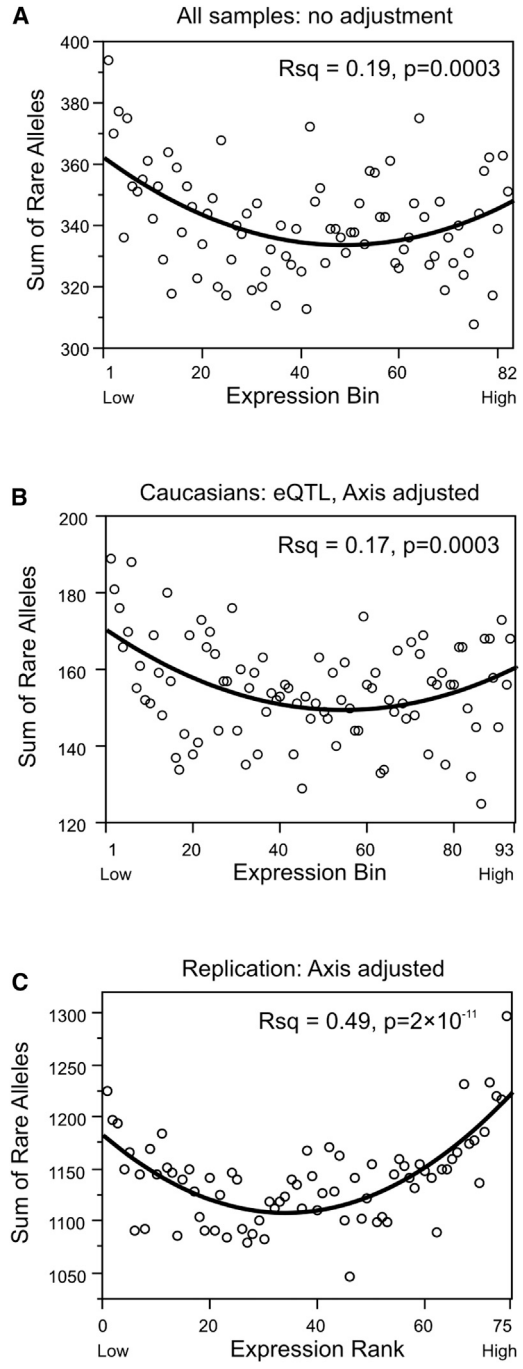

Figure 2. Relationship between Rare Variant Counts and Transcript Abundance

Each plot shows the cumulative number of rare variants in equal sized bins across the indicated number of genes and individuals, with lowest expression bin to the left and highest to the right. Lines indicate the best fit quadratic model.

(A) 472 genes in 410 individuals of mixed ethnicity (5 individuals per bin), gene expression data normalized by SNM with variance adjustment (whole model $R^2 = 0.19$, $p = 0.0003$).

(B) 472 genes in 279 Europeans (93 bins of 3 individuals) after removing effects of common eQTLs and conserved axes of covariation ($R^2 = 0.17$, $p = 0.0003$).

(C) 4,633 genes in 75 replicates after removing effects of conserved axes of covariation and PC1 ($R^2 = 0.49$, $p = 2 \times 10^{-11}$).

re-sequenced to confirm that it is homozygous for the reference allele, also as reported for the publically available K562 sequence.

DNA oligonucleotides containing a G followed by 19-nt guide sequence were kinased, annealed, and ligated into pX330 (gift from Feng Zhang, Addgene plasmid #42230). The four guide RNAs were *COMMD4*, 5′-GCGCCAAGAAGCCAGGGCCC-3′; *DHX29*, 5′-GCTCTCACTGCTCCCAAAAA-3′; *TDP2*, 5′-GTGCGCA GGCGCCTGTGTCA-3′; and *UQCC*, 5′-GGTGAAGGAGTAATTTT CTA-3′. Once constructed, the plasmids were sequenced to confirm the guide strand region using the primer CRISPR_Seq (5′-CGATACAAGGCTGTTAGAGAGATAATTGG-3′). K562 cells (1×10^6) were transfected by nucleofection with 1 μg CRISPR plasmid construct for *COMMD4*, *DHX29*, *TDP2*, or *UQCC* and 300 ng of pmaxGFP, according to manufacturer's recommended protocol (Lonza). GFP expression was analyzed 72 hr after transfection at which time DNA and RNA were prepared from pooled cells for preliminary analysis.

After a period of 4 to 10 days of growth, transfected cells were sorted by fluorescence activated cell sorting (FACS) and GFP-positive single cells or 10-cell colonies were sorted into 96-well plates. These were allowed to incubate for 2–4 weeks or until confluent and then subjected to T7E1 mutation detection assay[37] in which the region of interest was amplified and 200 ng of purified PCR product was re-annealed and digested with T7 endonuclease 1. Cleavage was confirmed by the appearance of reduced molecular weight bands on 2% agarose gels that were quantified by ImageJ in order to estimate the fractional heterozygosity.

After confirming disruption of each relevant SNP by the T7E1 assay, five clones were chosen for each of the four genes with average heterozygosity between 16% and 23%. Droplet PCR,[38] which is capable of detecting a 20% modulation of gene expression, was used with *HPRT* (OMIM: 308000) as a uniform control gene in all analyses, with *UQCC* as a second control for *TDP2* and *DHX29*, and with *TDP2* as the second control for *UQCC* and *COMMD4* disrupted clones. In each experiment we contrasted the relative expression of the knocked out gene in the five clones to its average expression in five clones carrying a knockout of another gene (that is, *TDP2* for *DHX29* and vice versa; *UQCC* for *COMMD4* and vice versa), performing a t test of the comparison. Differential expression was computed by formulating the ratio of each ddPCR count for the gene of interest to each reference gene, normalizing these ratios such that the average in the control cells was 1, then averaging the ratios to the two references. The average and significance of the change in expression is reported in Table S3 and Figure S3, which shows the proportional reduction or gain in signal for the five clones of each type, measured in two technical replicates of each clone.

Results

Figure 2A illustrates the core result that there is enrichment of rare variants for both increased and decreased gene expression in the full sample (model $R^2 = 0.19$, $p = 0.0003$, permutation $p = 0.0002$). This is true for a variety of modes of normalization of the gene expression data detailed in the Material and Methods section, including a conservative strategy involving removal of batch effects with Combat[24] followed by fitting age and gender followed by fitting PEER factors[23] to just the European-ancestry individuals (Figure S1A), simple quantile normalization[27]

(Figure S1B), surrogate variable analysis,[25] and an approach that first pooled the z-scores for all of the gene expression measures before assigning bins (Figure S1C). In subsequent analyses we employed supervised normalization of microarrays (SNM)[26] because it optimally controls for confounding between technical and biological sources of variance. The full dataset consisted of 297 individuals of European ancestry, 18 of East Asian ancestry, and 95 of predominantly African ancestry, but because African ancestry is associated with mild differential expression of one third of the genes and an almost 4-fold elevation in rare variant counts in the promoter regions (consistent with HapMap estimates[39,40]), we also considered each of the two larger population groups separately. The same trend was observed in both the European and African American samples but with only marginal significance in the latter due to the reduced sample size. Similarly, the result holds when we replaced total rare variant counts with rare haplotype counts, noting that some individuals carry alleles that differ from the reference allele at two or more sites. Because some genes show flatter distributions of expression in African ancestry samples, there is an excess of African ancestry at the extremes for the combined analysis, which we conservatively corrected by additionally standardizing the gene expression within each population group, which reduced the strength of the association with rare variant counts, particularly for increased expression (Figure S1D). Larger sample sizes will be required to explore whether there are population differences in rare variant contributions.

Further evidence that the regression models truly capture enrichment for rare variants at extremes of gene expression comes from the observation that the fit and significance improve after adjustment for covariates that are known to influence gene expression. Figure 2B shows the model fit for the Europeans (so as to avoid false positives due to population stratification) when gene expression bins were reassigned to the residuals after also fitting known common variant eQTL effects[29] for each gene, as well as principal component scores for seven common axes of peripheral blood gene expression covariance.[28] These seven axes collectively explain an average of 26% of the expression variance of the 472 genes (range 2% to 89%), which is approximately four times as much as the variance explained by the major cell types in whole blood (average 6%, range 0% to 44%; Figure S5). Fitting common eQTLs and the axis scores improves the model R^2 from 0.05 ($p = 0.07$) to 0.17 ($p = 0.0003$, permutation $p < 0.0001$).

Next we asked whether the enrichment might be attributed to particular classes of gene or gene region, summarizing the results in Table S1 and Figure S4. First, categorizing all variants with respect to predicted regulatory potential according to RegulomeDB[30] classifications confirms that variants that lie within features such as DNase hypersensitive sites or transcription factor binding sites (classes 1–4) are more enriched at the extremes than all other variants in classes 5–7 ($p = 0.002$ versus 0.08). Second, surprisingly, the enrichment was much stronger for variants located downstream of the transcription start site ($p = 0.0007$) versus upstream ($p = 0.09$). Third, we investigated whether there was a difference between overall low-abundance and high-abundance transcripts by dividing the dataset into two halves according to that parameter, but no difference was noted. Analyzing the genes separately in sets with low, intermediate, and high levels of promoter polymorphism suggests that the enrichment is observed across all levels of polymorphism. Notably as well, the burden remains when we reduced the minor allele frequency threshold to 0.01 (Figure S1F), reflecting the fact that the majority of rare variants are found in five or fewer individuals.

The most striking differential enrichment was observed with respect to gene function. The 472 genes were chosen for analysis in order to obtain approximately equal representation with respect to two criteria: whether or not they contain known common variant eQTLs, and whether or not they are thought to be associated with common chronic diseases. Genes not represented on the MetaboChip[31] are significantly more likely to be enriched for rare regulatory variants ($p = 9 \times 10^{-6}$ versus 0.73 for metabolic disease-related genes). To confirm that this is significant, we compared the deviation in R-squared values with those of 1,000 random partitions of the European dataset keeping the number of individuals in the two sets constant, and observed that the Metabochip deviation is toward the tail ($p = 0.007$). A similar trend was observed for genes on the Immunochip,[32] $p = 0.005$ versus 0.11, but this difference was not significant relative to the random partitions. These trends toward reduced rare regulatory SNP presence in promoters of disease-associated genes might be explained by relaxation of purifying selection on genes not associated with disease.

The possibility that some genes are more tolerant of regulatory variants is also implied by the observation that genes with common eSNPs are much more likely to harbor rare promoter-proximal variants affecting gene expression ($p = 0.0006$) than those without ($p = 0.11$). Furthermore, genes with low promoter polymorphism ($\pi_{5'P}$) relative to coding region polymorphism (π_{cod}) highlighted in Figure 3A are highly significantly depleted for rare variants in the top decile of effect sizes inferred from our data (Figure 3B; t test contrasting number in bottom 10% of low $\pi_{5'P}$ / π_{cod} genes against the remainder, $p = 3 \times 10^{-5}$). The deficit is not due solely to low polymorphism because the contrast remains significant when $\pi_{5'P}$ is included as a covariate, and the bottom 10% of $\pi_{5'P}$ genes overall are as likely to harbor large effect rare variants as all other genes. This result raises the possibility that more extensive analyses of rare regulatory variant association with transcript abundance in different tissues might give rise to a measure of promoter region tolerance to functional mutation similar to the RVIS[41] and constraint scores[42] that classify genes with respect to intolerance to disruptive and pathogenic protein mutations.

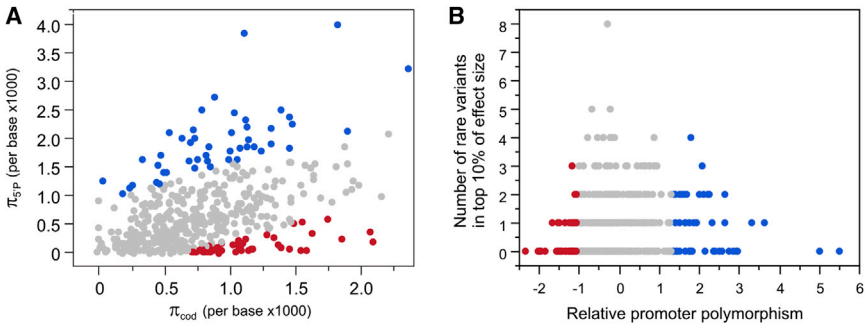

Figure 3. Genes with a Low Ratio of Promoter to Coding Polymorphism Are Intolerant to Large Effect Regulatory Variants
(A) Regression of $\pi_{5'P}$ on π_{cod} highlighting the bottom 10% of genes with the largest negative residuals (lowest ratio, red) and top 10% (blue).
(B) Plot of relationship between the number of variants per gene in the top decile of estimated effect size from all 8,833 rare variants (MAF < 0.05) and residual from the regression in (A) as a measure of relative promoter polymorphism.

Another potential source of functional regulatory variation is divergent promoters derived from admixture with archaic humans. We identified 31 genes in the dataset that harbor haplotypes with 4 or more rare variants in Europeans, and for 8 of these genes (26%), the divergent haplotype matches informative Neandertal or Denisovan sequences. Similarly, 15 of 75 (20%) haplotypes with 3 or more rare variants appear to be archaic, whereas less divergent haplotypes are more likely to be generated by successive mutations in the human lineage. The remaining divergent haplotypes might represent Neandertal alleles not found in the Altai cave individual, as-yet-unrecognized archaic lineages, or fast-evolving alleles also present in the African gene pool. Two of the non-archaic divergent haplotypes, both also found in multiple African ancestry individuals in our sample, were likely to associate with elevated expression, in *CDK10* (2 individuals, two-tailed t test $p = 1.2^{-10}$) and *NDUFB10* (12 individuals, two-tailed t test $p = 0.0028$). By contrast, there was no tendency for any of the archaic haplotypes to be associated with extreme expression in peripheral blood. Further analyses imputing ancestry based on long-range haplotypes from multiple Neandertal individuals, and assessments in multiple tissues, might nevertheless define roles for archaic promoter polymorphism in gene regulation.

In order to estimate the frequency and magnitude of effects that would be consistent with the observed enrichments, we performed a simulation study where effects were assigned to the empirically determined genotypes and added to randomly generated and normally distributed gene expression values. Figure 4A shows the actual observed effects from the data, and this distribution is compared with those drawn from a gamma(1.5, 0.12) distribution (Figure 4B) that generates results comparable to the observed enrichment (Figure 4C). The estimated effect sizes under this model are somewhat smaller than those estimated from the data, probably a consequence of sampling overestimation in the experiment and the absence of technical noise in the simulated data. However, the result suggests that observed effect sizes are very rarely greater than 2 SDs and that most of the enrichment is driven by rare variants with influences comparable to those of common eQTLs, namely allelic substitutions in the range of 0.5 to 1 SD.

Next we replicated the result with an independent dataset consisting of 75 whole-genome sequences of mixed ancestry, linked to whole-blood RNA-seq profiles. Figure S5 shows that the same trend was observed with the 472 genes as in the Atlanta cohort, but owing to the small sample size of individuals the result is only significant at $p = 0.008$. However, increasing the analysis to include the same total number of comparisons as in the CHDWB, namely 4,633 genes, and adjusting for the axes of variation as well as a strong surrogate variable corresponding to ancestry provides clear replication of the rare variant enrichment ($R^2 = 0.49$, $p = 2 \times 10^{-11}$, permutation $p < 0.0001$; Figure 2C).

To experimentally validate rare variant regulatory effects predicted from the statistical analysis, we used CRISPR/Cas9 to mutagenize four sites that had estimated effect sizes greater than 5-fold higher or lower than the population mean in a single individual, in K562 erythroleukemia cells.[43] 11 individual clones were grown for each disruption and once cleavage was confirmed via the T7E1 assay,[37] indicating 16% to 23% average heterozygosity, 5 clones were chosen for RNA abundance measurement. Droplet digital RT-PCR,[38] which quantifies relative abundance by counting the number of nanodroplets from a dilution of the RNA sample that yield PCR product, assessed relative to control genes, was used to demonstrate that disruption of three of the four sites resulted in reduced or increased transcript abundance (Table S3). Disruption of rs182080358 associated with increased expression of *COMMD4* weakly increased transcript abundance in three of the four clones and greatly increased it in another. All five clones with disruptions in *UQCC* reduced expression almost by half, whereas disruption of a negative control site had no effect and four of five clones with disruptions in *TDP2* reduced expression to varying degrees (Figure S3). Because our CRISPR protocol causes small deletions rather than targeted replacement of polymorphisms, it is expected that the disruptions remove binding sites for transcriptional activators in each case. This could cause loss or gain of expression depending on the nature of the transcription factor, but it is notable that in each case the effect of disruption was in the same direction as the nucleotide substitution. The difference in estimated effect size might be attributed to over-estimation of the effect from the microarray data and underestimation in the CRISPR clones that have incomplete heterozygosity because K562 cells are largely triploid.[43]

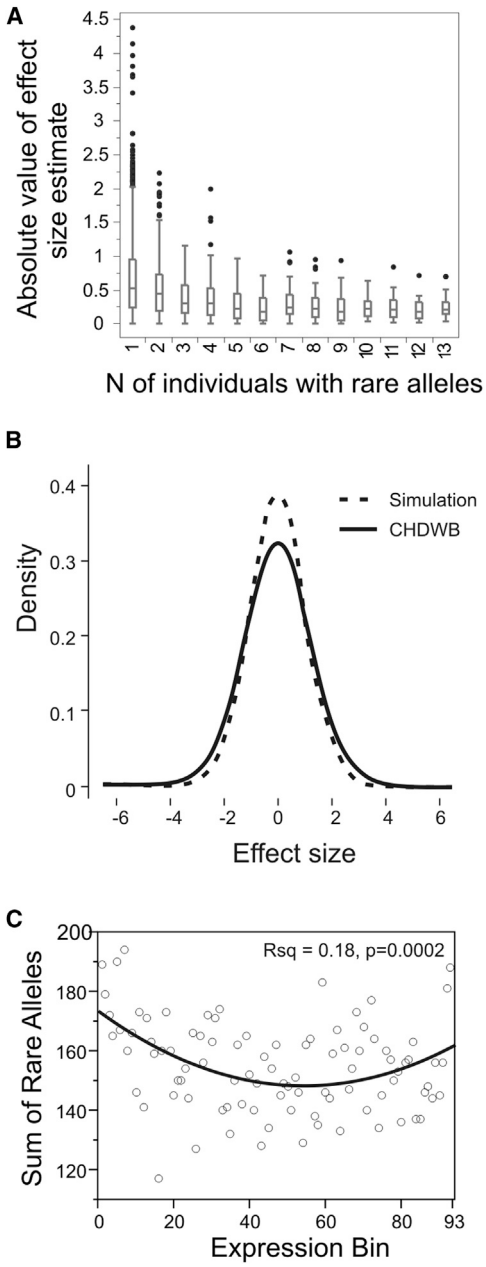

Figure 4. Estimation of Effect Sizes of Rare Variants
(A) The distribution of absolute values of the estimated effect size from the CHDWB data as a function of the number of alleles in the sample. Boxes show mean and interquartile range with whiskers at 1.5 times IQR with outliers as single points.
(B) Comparison of estimated effect size distribution from the data (bold curve) and in the simulation (thin black curve) showing slight excess of larger effect variants in the observed data.
(C) Simulated model fit assuming a gamma distribution with gamma(1.5,0.12), for 472 genes in 279 individuals, showing excess of extreme expression as in the actual data ($R^2 = 0.18$, $p = 0.0002$).

Discussion

These experiments demonstrate that the combination of gene sequencing with transcript profiling in peripheral blood has good potential as a screening approach to identify rare promoter-proximal variants that significantly disrupt the regulation of gene expression. Although the average effect sizes we estimate are not large, they are on a par with those estimated for disease-associated common eQTLs.[29] That is to say, large-effect rare variants typically alter expression of a gene between one half and one standard deviation unit. For a transcript otherwise expressed in the inter-quartile range, namely in the middle half of the distribution, this will often be enough to cause it to shift it to the upper or lower decile where it would presumably be more likely to contribute to pathology or abnormality.

Two other studies have also found indirect evidence for rare variant influences on gene expression in cell lines, albeit spread across extended *cis*-regulatory regions. Montgomery et al.[44] performed eQTL analysis on the 1000 Genomes Project lymphoblast cell lines and observed that transcripts with rare instances of allele-specific expression (ASE) had a median of four perfectly concordant rare putative regulatory variants within 100 kb of the TSS, compared with three for control transcripts not exhibiting ASE. Interestingly, the enrichment was greater for non-synonymous than synonymous coding ASE variants and biased to lower expression, suggesting an epistatic interaction between regulatory and structural polymorphism. They also found that very rare alleles in conserved putative regulatory sites were more likely to be 2 SDs from the mean. Reanalyzing the Geuvadis lymphoblast cell line data, Zeng et al.[45] focused on rare instances where the correlation between two transcripts was aberrant in a few individuals and described an enrichment of private variants specifically in the vicinity of enhancers in the 1 Mb region of the genes.

We do not observe any tendency for variants to be more likely to increase or decrease gene expression, which is consistent with the inference that globally, transcript abundance is under stabilizing selection.[14,15] Recent results from yeast[46] also indicate that there is selection against noise promoted by regulatory variants, so it will be interesting in larger studies to evaluate whether genes with reduced promoter relative to coding polymorphism also tend to have reduced expression variability due to rare variants.

If our experimental strategy can be applied to tissues or cell types directly relevant to certain pathologies, such as neurons or cardiac cells, it could enhance efforts to infer whether rare variants in regulatory regions are functionally deleterious. It is possible that iPSC culture might be beneficial in this regard, particularly if the effects on transcription are greater in uniform cell culture than in mixed whole-blood cell populations from individuals who experience a wide range of environments. By extrapolation from our data, we estimate that the average individual probably carries several dozen such variants that cause gene expression to be toward the extreme, and consequently these cannot be ignored as a potential source of disease-related pathology. Determination of whether or

not rare regulatory variants are sufficient to cause rare diseases in the same manner as structural variants are inferred to will require experimental designs targeted to individuals with particular congenital disorders.

Accession Numbers

Short read sequences have been deposited at the SRA (accession number phs001021.v1.p1) with restricted access requiring approval by dbGaP under the terms of the consent provided by CHDWB participants. The raw data are available at the Gene Expression Omnibus as accession GEO: GSE61672.

Supplemental Data

Supplemental Data include five figures, three tables, and an Appendix and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.12.023.

Acknowledgments

This work was supported by NIH Program Project 1-P01-GM0996568 (Project 3, G.G. Principal Investigator), directed by Bruce Weir (University of Washington). We are grateful to Cathy Laurie and Sarah Nelson for assistance with imputation, Shweta Biliya for sequencing, Sahar Gelfman and David Goldstein for provision of the replication dataset (which was generated with the support of NIH grant RC2MH089915), Mary-Jane Chandler for assistance with the ancestral genome analysis, and Cathy Laurie, Joe Lachance, Patrick McGrath, and Fred Vannberg for comments on the manuscript. We thank all of the staff and participants at the CHDWB for their willing assistance, especially Ken Brigham for his vision for the Center.

Received: August 17, 2015
Accepted: December 30, 2015
Published: February 4, 2016

Web Resources

The URLs for data presented herein are as follows:

GEO, http://www.ncbi.nlm.nih.gov/geo/
HUGO Gene Nomenclature Committee, http://www.genenames.org/
Neandertal and Denisovan genome sequences, http://cdna.eva.mpg.de/neandertal/altai/
OMIM, http://www.omim.org/
Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/sra
UCSC Genome Browser, http://genome.ucsc.edu
Zhao et al. additional materials, http://cig.gatech.edu/supplementary/zhao-et-al-2016

References

1. Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
2. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190.
3. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.
4. Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O., et al. (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature *498*, 232–235.
5. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
6. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
7. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.
8. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
9. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
10. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
11. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.
12. Chandrasekharappa, S.C., Lach, F.P., Kimble, D.C., Kamat, A., Teer, J.K., Donovan, F.X., Flynn, E., Sen, S.K., Thongthip, S., Sanborn, E., et al.; NISC Comparative Sequencing Program (2013). Massively parallel sequencing, aCGH, and RNA-Seq technologies provide a comprehensive molecular diagnosis of Fanconi anemia. Blood *121*, e138–e148.
13. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
14. Lemos, B., Meiklejohn, C.D., Cáceres, M., and Hartl, D.L. (2005). Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution *59*, 126–137.
15. Hodgins-Davis, A., Rice, D.P., and Townsend, J.P. (2015). Gene expression evolves under a House-of-Cards model of stabilizing selection. Mol. Biol. Evol. *32*, 2130–2140.

16. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354–357.
17. Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017–1021.
18. Qin, P., and Stoneking, M. (2015). Denisovan ancestry in East Eurasian and Native American populations. Mol. Biol. Evol. *32*, 2665–2674.
19. Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., et al. (2014). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature *512*, 194–197.
20. Brigham, K.L. (2010). Predictive health: the imminent revolution in health care. J. Am. Geriatr. Soc. *58* (*Suppl 2*), S298–S302.
21. Tabassum, R., Cunningham, L., Stephens, E.H., Sturdivant, K., Martin, G.S., Brigham, K.L., and Gibson, G. (2014). A longitudinal study of health improvement in the Atlanta CHDWB wellness cohort. J. Pers. Med. *4*, 489–507.
22. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.
23. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.
24. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.
25. Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.
26. Mecham, B.H., Nelson, P.S., and Storey, J.D. (2010). Supervised normalization of microarrays. Bioinformatics *26*, 1308–1315.
27. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.
28. Preininger, M., Arafat, D., Kim, J., Nath, A.P., Idaghdour, Y., Brigham, K.L., and Gibson, G. (2013). Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet. *9*, e1003362.
29. Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet. *45*, 1238–1243.
30. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.
31. Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J., et al. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. *8*, e1002793.
32. Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. Arthritis Res. Ther. *13*, 101.
33. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
34. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.
35. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.
36. Cho, S.W., Kim, S., Kim, J.M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat. Biotechnol. *31*, 230–232.
37. Mashal, R.D., Koontz, J., and Sklar, J. (1995). Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. Nat. Genet. *9*, 177–183.
38. Kiss, M.M., Ortoleva-Donnelly, L., Beer, N.R., Warner, J., Bailey, C.G., Colston, B.W., Rothberg, J.M., Link, D.R., and Leamon, J.H. (2008). High-throughput quantitative polymerase chain reaction in picoliter droplets. Anal. Chem. *80*, 8975–8981.
39. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
40. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
41. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. *9*, e1003709.
42. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of *de novo* mutation in human disease. Nat. Genet. *46*, 944–950.
43. Lozzio, B.B., and Lozzio, C.B. (1977). Properties of the K562 cell line derived from a patient with chronic myeloid leukemia. Int. J. Cancer *19*, 136.
44. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E.T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. *7*, e1002144.
45. Zeng, Y., Wang, G., Yang, E., Ji, G., Brinkmeyer-Langford, C.L., and Cai, J.J. (2015). Aberrant gene expression in humans. PLoS Genet. *11*, e1004942.
46. Metzger, B.P., Yuan, D.C., Gruber, J.D., Duveau, F., and Wittkopp, P.J. (2015). Selection on noise constrains variation in a eukaryotic promoter. Nature *521*, 344–347.

The American Journal of Human Genetics

Supplemental Data

A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood

Jing Zhao, Idowu Akinsanmi, Dalia Arafat, T.J. Cradick, Ciaran M. Lee, Samridhi Banskota, Urko M. Marigorta, Gang Bao, and Greg Gibson

Zhao *et al*, 2015. **Figure S1**

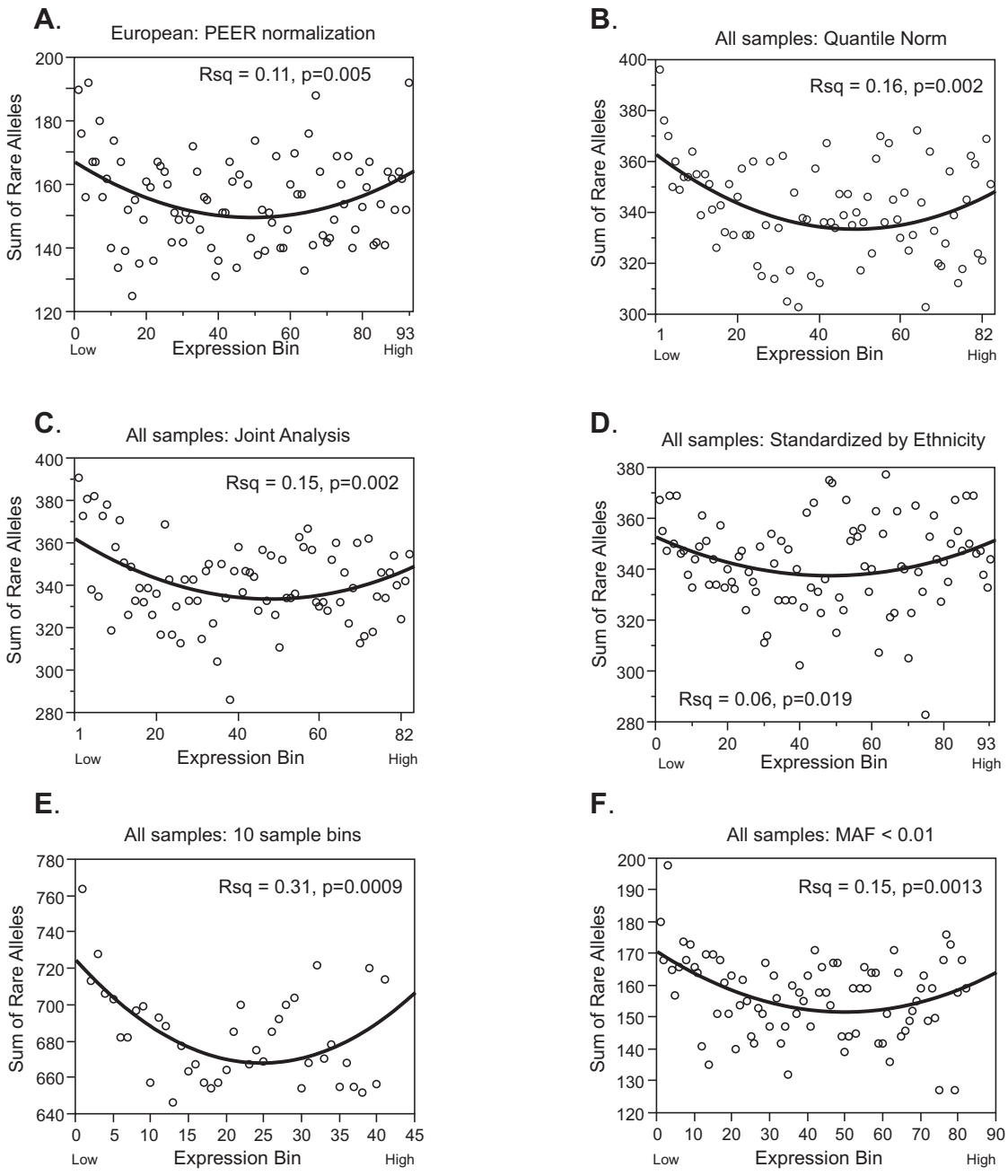

Figure S1. Quadratic regression plots with alternate normalization, bin size or MAF. As in Figure 1, each plot shows the quadratic regression fit of the sum of rare variants (A-E: MAF < 0.05; F, MAF < 0.01) in equal-sized expression bins from low to high (A-D,F: 5 individuals; E: 10 individuals) . Rsq and p-values refer to the full model R-squared and p-value from a quadratic regression. (A) European only analysis with PEER normalization; (B) Quantile of all samples, no adjustment; (C) Joint analysis of z-scores of all genes; (D) SNM normalization of all samples after standardization by ancestry; (E) as in Figure 1A but with bin size of 10; (F) All samples SNM but with MAF<0.01.

Zhao *et al*, 2015. **Figure S2**

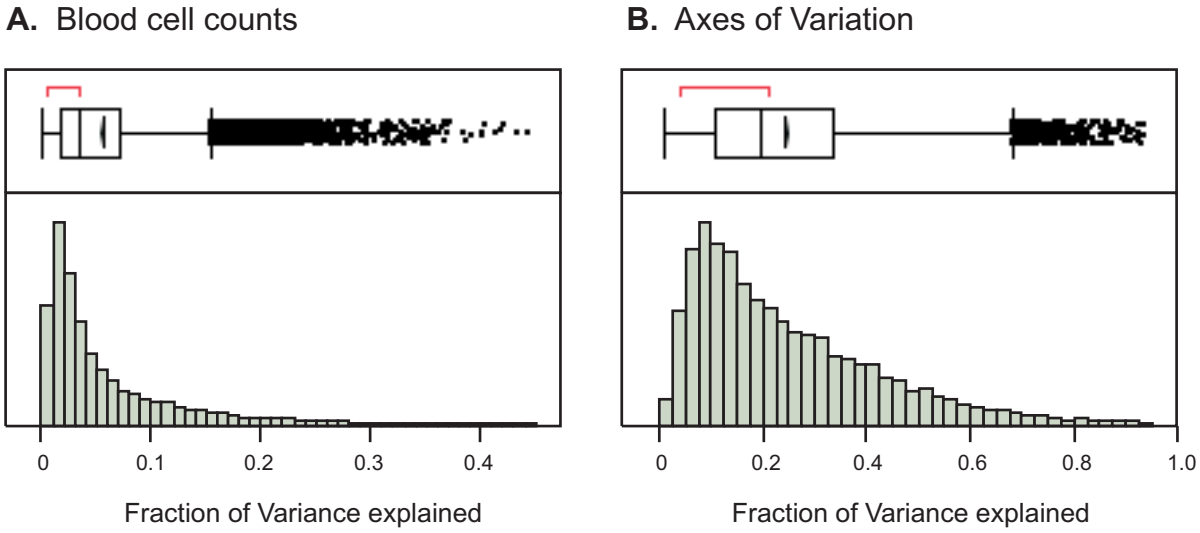

Figure S2. Variance explained by Blood Counts and Axes of Variation in CHDWB cohort. Histograms show the proportion of genes with the indicated fraction of variance explained by blood counts (lymphocytes, neutrophils, monocytes, red blood cells, platelets) and seven common Axes of variation respectively in (A) and (B), as the R-squared for the full model in a multiple regression. Boxes show the mean and inter-quartile range (25th to 75th percentile), and whiskers show 1.5X the respective IQR with outlier points shown as small squares.

Zhao *et al*, 2015. **Figure S3**

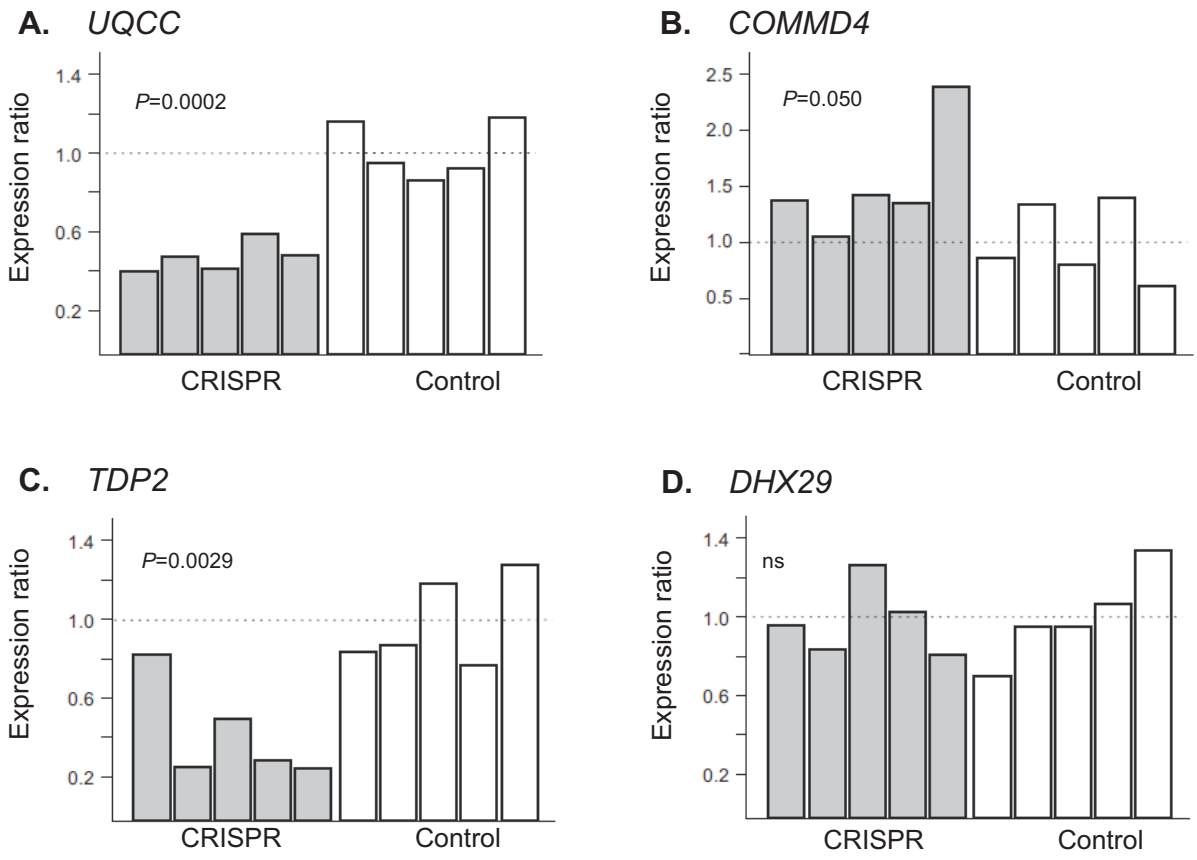

Figure S3. CRISPR / Cas9 mutagenesis validation of rare SNP regulatory effects. Average relative expression measures for the indicated transcript for five CRISPR mutagenized K562 10-cell clones in the same gene (gray) and in a different negative control gene (white). Dotted line represents mean expression of the control clones, to which the mutant clones are compared. See Supplementary methods for details of analysis.

Zhao *et al*, 2015. **Figure S4**

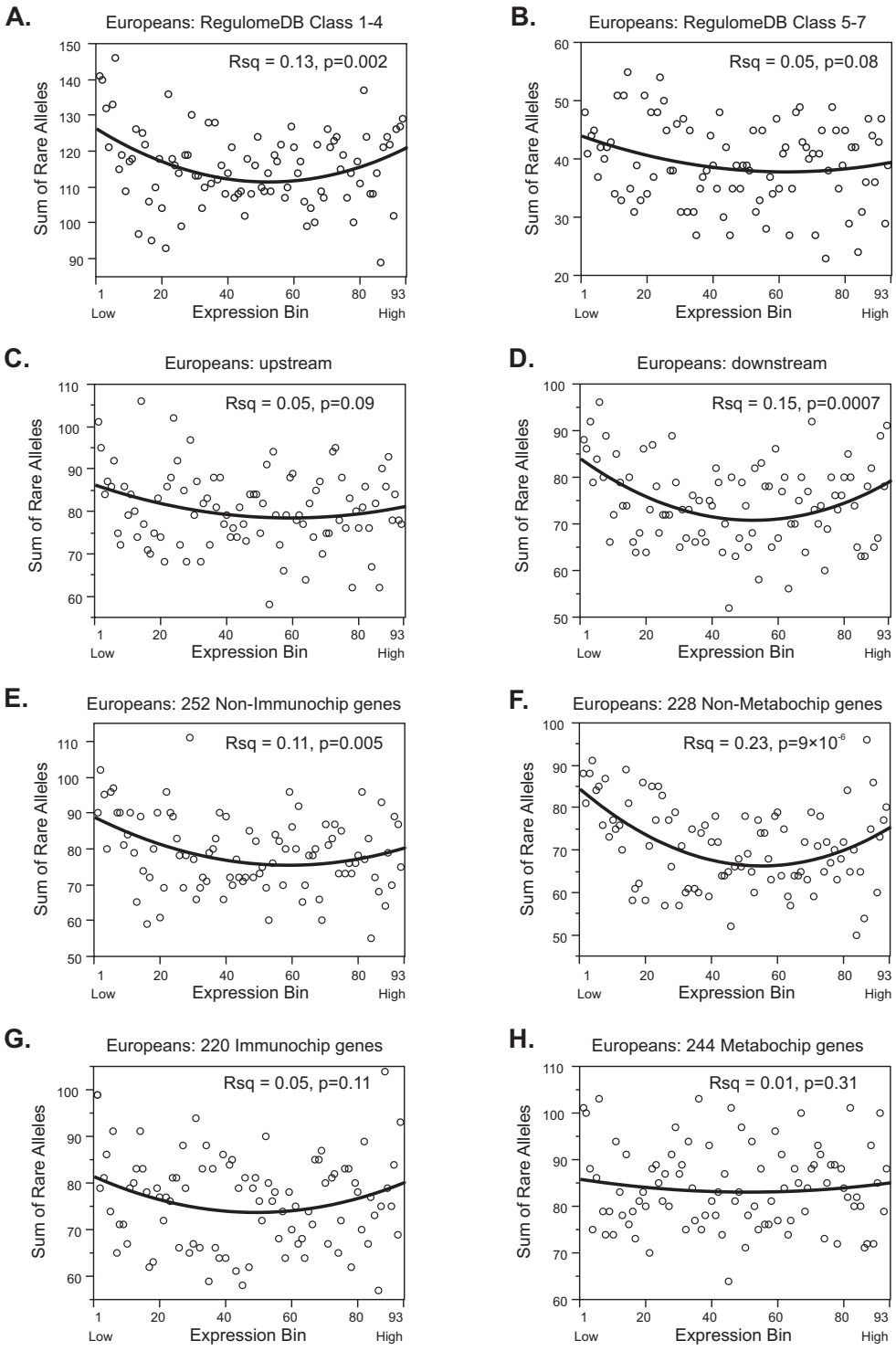

Figure S4. Quadratic regression plots for select sub-sets of SNPs or genes Rsq and p-values refer to the full model R-squared and p-value from a quadratic regression. See Suppl. Table 1 for significance of linear and quadratic terms separately.

Zhao *et al*, 2015. **Figure S5**

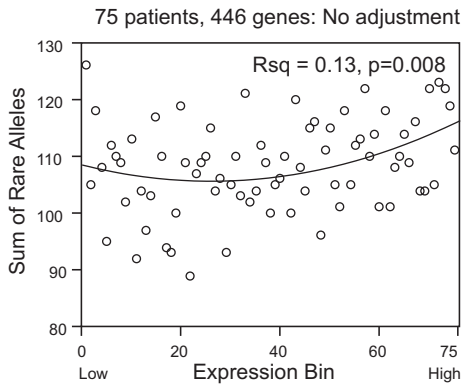

Figure S5. Quadratic regression plots for Replication dataset with 446 genes only, namely the same genes as in the CHDWB that are also present in the RNASeq dataset. Figure 1C shows an expanded replication dataset

Table S1 Comparisons of rare variant burden in subsets of SNPs and Genes

Comparison		First set				Second set			
		AvgCt[1]	Model[2]	Linear[3]	Quad[4]	AvgCt[1]	Model[2]	Linear[3]	Quad[4]
High polymorphism (242)	vs Low polymorphism (230)	104.9	14.4**	0.032	0.002	50.1	0.03[ns]	0.50	0.12
High expression (236)	vs Low expression (236)	74.9	10.7*	0.35	0.002	80.2	14.5**	0.0008	0.08
Metabochip (244)	vs non-Metabochip (228)	83.9	0.7[ns]	0.81	0.45	71.1	22.7***	0.006	<0.0001
Immunochip (220)	vs non-Immunochip (252)	76.1	4.8[ns]	0.72	0.04	78.9	11.2*	0.021	0.017
With eQTL (207)	vs No eQTL (265)	68.7	15.3*	0.027	0.001	86.4	4.9[ns]	0.33	0.06
Upstream (472)	vs Downstream (472)	80.5	5.1[ns]	0.12	0.12	74.5	15.0*	0.12	0.0004
RegulomeDB 1-4 (472)	vs RegulomeDB 5-7 (472)	115.7	12.7*	0.16	0.001	39.4	5.5[ns]	0.08	0.15

[1] Average count of rare variants in each bin in the dataset

[2] Model R-squared, with significance: * 0.01<p<0.001; ** 0.001<p<0.0001; *** p<0.0001

[3] p-value for linear term in the model

[4] p-value for quadratic term in the model

Table S2. Archaic Haplotypes detected in Caucasians

Gene 1: LILRB2. Two individuals with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site																					
rs118126757	19	54784069	G	C	0.00244									1		1	1												
rs147377983	19	54784404	T	C	0.00488																								
chr19:54784451	19	54784451	A	T	0.00122																								
rs191476279	19	54784500	T	C	0.00488																								
rs140128507	19	54784516	A	G	0.00122	1			1																				1
rs4806734	19	54784587	T	C	0.00366																								
rs188025641	19	54784621	T	C	0.00122					1					1						1								
chr19:54785045	19	54785045	G	A	0.00488															1									
rs146571067	19	54785085	G	C	0.00122	1	1		1																				1
chr19:54785088	19	54785088	G	C	0.00122																								
chr19:54785089	19	54785089	T	C	0.00366																								
rs112796768	19	54785092	G	C	0.00488						1																		
chr19:54785135	19	54785135	C	A	0.00488	1			1																				1
rs188272820	19	54785143	T	C	0.00488	1			1																				1
rs182465889	19	54785150	A	G	0.00122	1			1																				1
chr19:54785189	19	54785189	A	C	0.00488																			1					
rs149524204	19	54785217	A	G	0.00122	1	1		1																				1
chr19:54785227	19	54785227	A	G	0.00488																					1			
rs117972222	19	54785329	C	T	0.00976	1	1		1																				1
rs191923839	19	54785346	T	C	0.00488			1				1	1					1							2		1		
rs117542982	19	54785534	G	A	0.00122	1			1																				1
chr19:54785535	19	54785535	C	T	0.00488																	1							
rs117127854	19	54785610	T	C	0.00122	1			1																				1
chr19:54785651	19	54785651	T	C	0.00122														1										
chr19:54785721	19	54785721	A	C	0.00244																		1						
rs113145131	19	54785835	T	C	0.01098																								
rs113412095	19	54785863	A	C	0.00122																1								

Gene 2: TGF. Five individuals with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site																																								
rs116650206	3	100427151	A	T	0.00976																																											
rs116622528	3	100427159	T	C	0.00854	1						1		1																1			1				1											
rs114652332	3	100427191	T	C	0.00244																																											
rs62274023	3	100427207	C	T	0.0061											1	1												1																	1		
rs9853377	3	100427253	T	C	0.03293	1						1		1																1			1				1											
rs146889339	3	100427283	A	C	0.00122																																											
chr3:100427416	3	100427416	T	C	0.00122																																							1				
rs140713495	3	100427524	G	C	0.00366																																											
rs967308	3	100427676	C	T	0.03293	1						1		1																1			1				1											
rs145846729	3	100427729	T	C	0.00122													1																														
rs967309	3	100427738	T	C	0.03293	1						1		1																1			1				1											
chr3:100427772	3	100427772	T	G	0.00122																				1																							
chr3:100427953	3	100427953	C	T	0.00244					1	1																																					
rs75268146	3	100428061	C	T	0.03415	1						1		1																1			1				1											
rs28364602	3	100428066	G	A	0.00976																																											
rs12493765	3	100428072	A	C	0.00976										1						1				1	1								1														
rs141024756	3	100428195	C	G	0.00854	1						1		1																1			1				1											
rs190536612	3	100428205	G	A	0.00854	1						1		1																1			1				1											
rs112759386	3	100428243	T	C	0.0061																		1																									
chr3:100428245	3	100428245	A	T	0.00122																																								1			
rs72919417	3	100428286	T	G	0.02561	1						1		1																1			1				1											
rs111540048	3	100428298	A	G	0.0122																																											
rs114124529	3	100428356	G	A	0.02561			1	1				1						1	1		1	1		1			1			1	1			1	1		1	1	1	1	1	1				1	1

Gene 3: DLAT. Nine individuals with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site																								
rs12099187	11	111894560	T	C	0.00488																												
rs151035291	11	111894896	A	G	0.00732																												
rs78290067	11	111894991	C	T	0.03049									1																			
chr11:111895066	11	111895066	C	G	0.00122																												
chr11:111895067	11	111895067	C	G	0.00244																					1					1		
chr11:111895102	11	111895102	G	T	0.00122						1																						
rs73561950	11	111895193	T	G	0.03415	1	1		1	1	1						1			1			1		1		1					1
rs190887558	11	111895550	A	C	0.00488																												
rs114863504	11	111895560	A	G	0.00732																	1											
rs78298568	11	111895599	T	C	0.0439	1	1		1	1	1						1			1		1	1		1							1
chr11:111895636	11	111895636	A	G	0.00122											1																
chr11:111895704	11	111895704	A	C	0.00122																												
chr11:111895840	11	111895840	T	G	0.00244							1																			1		
chr11:111895960	11	111895960	G	C	0.00122																								1				
chr11:111895985	11	111895985	G	C	0.00366								1														1				1		
chr11:111895994	11	111895994	T	C	0.00122										1																		
chr11:111896040	11	111896040	A	G	0.00732																	1											
rs115067052	11	111896041	G	C	0.00732																	1											
chr11:111896072	11	111896072	C	A	0.00244			1																									
chr11:111896107	11	111896107	C	G	0.00122																							1					
rs150145390	11	111896242	A	G	0.00122																												
rs61757217	11	111896251	C	G	0.01098	1			1	1	1						1			1			1		1							1
chr11:111896339	11	111896339	A	T	0.00122																1												
chr11:111896404	11	111896404	T	C	0.00122													1															
rs201934276	11	111896514	C	T	0.00122																	1											
rs137879141	11	111896524	G	A	0.0122									1																			
rs192971364	11	111896531	A	G	0.00122																	1											

Gene 4: NOP10. Eight individuals with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site																								
rs146293018	15	34634434	C	T	0.0122	1	1	1					1					1		1				1	1		1		1			1
chr15:34634473	15	34634473	C	T	0.00122									1																			
rs113680082	15	34634512	C	G	0.00244																												
rs144573482	15	34634727	T	C	0.00488																												
rs148453358	15	34634853	G	T	0.0122	1	1	1					1					1		1				1	1		1		1			1
chr15:34634987	15	34634987	G	A	0.00122																												
rs146261631	15	34635241	G	C	0.00854			1	1			1			1						1												
chr15:34635348	15	34635348	A	G	0.00122																												
rs149606664	15	34635380	C	T	0.01341	1	1	1				1	1					1		1				1	1		1		1			1
rs139502071	15	34635383	T	C	0.0122	1		1					1					1		1				1	1		1		1			1
rs76537972	15	34635391	C	T	0.00366						1	1																		1			
rs76029512	15	34635393	A	G	0.00732										1	1						1	1					1					
rs186487749	15	34635399	G	C	0.00854			1	1			1			1					1													
rs75061777	15	34635597	A	G	0.00366																												
rs114643475	15	34635646	G	A	0.00122																												
rs60281344	15	34635684	T	A	0.00366					1																							

Gene 5: MRPL53. One individual with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site
rs112896280	2	74699066	C	G	0.00488			
chr2:74699141	2	74699141	A	G	0.00854			
chr2:74699184	2	74699184	T	C	0.00122			
rs148007344	2	74699584	C	T	0.00488	1	1	
rs78834087	2	74699595	T	G	0.01463			
rs201235080	2	74699675	T	C	0.00122			
rs200975054	2	74699680	A	C	0.00122			
rs141704877	2	74699715	C	T	0.0122			
rs77266305	2	74699922	C	T	0.00488	1	1	
chr2:74700003	2	74700003	T	C	0.00122			
rs183078378	2	74700031	C	T	0.00366			
chr2:74700040	2	74700040	A	G	0.00122			
rs13403485	2	74700190	C	A	0.01341			
chr2:74700213	2	74700213	G	A	0.00122			
rs139465854	2	74700214	A	G	0.00244			
rs141818840	2	74700242	A	T	0.00244			
rs17009955	2	74700451	C	A	0.0122	1	1	
rs183406812	2	74700504	C	G	0.00244			
chr2:74700609	2	74700609	A	G	0.00122			
chr2:74700656	2	74700656	T	C	0.00244			
chr2:74700659	2	74700659	A	G	0.00122			
chr2:74700754	2	74700754	A	G	0.00122			

Gene 6: ALDH3B1. One individual with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site
chr11:67775057	11	67775057	A	G	0.00244			
chr11:67775089	11	67775089	T	C	0.00122			
chr11:67775337	11	67775337	A	C	0.00122			
chr11:67775397	11	67775397	A	C	0.00122			
rs78790062	11	67775439	G	C	0.00366			
chr11:67775464	11	67775464	C	A	0.00122			
rs308340	11	67775505	A	G	0.02317			
chr11:67775709	11	67775709	C	G	0.00122			
chr11:67775821	11	67775821	T	C	0.00122			
rs75423674	11	67775822	A	G	0.02317	1	1	
chr11:67775842	11	67775842	A	C	0.00122			
chr11:67776124	11	67776124	A	G	0.00122			
chr11:67776137	11	67776137	C	T	0.00122			
chr11:67776152	11	67776152	T	C	0.00122			
chr11:67776153	11	67776153	A	G	0.00122			
chr11:67776279	11	67776279	A	G	0.00122			
rs117602660	11	67776283	T	C	0.01707		1	
rs308337	11	67776292	A	C	0.03659			
chr11:67776311	11	67776311	A	G	0.00122			
rs308336	11	67776320	C	T	0.03659			
rs184759902	11	67776353	T	C	0.00244			
rs3763941	11	67776362	T	C	0.0439			
rs140148305	11	67776387	T	C	0.00854			
chr11:67776410	11	67776410	A	G	0.00122			
rs557098	11	67776465	C	T	0.0378			
rs116715743	11	67776488	G	A	0.00122			
rs149823943	11	67776579	T	G	0.00488			
chr11:67776620	11	67776620	T	C	0.00122			
rs7113328	11	67776664	A	G	0.00366			
rs308335	11	67776686	G	A	0.03659			
rs308334	11	67776852	C	G	0.03659			
rs189473290	11	67776917	T	G	0.00244	1		
rs147659398	11	67776924	T	G	0.0061			
rs308333	11	67777026	A	G	0.03659			
chr11:67777046	11	67777046	T	C	0.00122			

Gene 7: C7orf25. 22 individuals with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site																								
chr7:42950799	7	42950799	C	T	0.00122																											
chr7:42951067	7	42951067	C	G	0.00122																											
rs148046461	7	42951174	A	G	0.00244																											
rs28372722	7	42951646	T	C	0.03171	1		1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1		1	1	1	2	1	1	1
chr7:42951652	7	42951652	A	G	0.00244																											
chr7:42951824	7	42951824	G	T	0.00244																											
chr7:42951879	7	42951879	T	G	0.00122																											
rs182726646	7	42951898	T	C	0.00122																			1								
rs7786568	7	42952016	G	C	0.00976																											
rs28372721	7	42952186	C	G	0.03171	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1		1	1	1	2	1	1	1
chr7:42952282	7	42952282	G	A	0.00122															1												
rs184091995	7	42952346	T	C	0.00122																	1										

Gene 8: RBBP4. 23 individuals with haplotypes introgressed from Archaics

Marker	Chr	Pos (hg19)	Allele 1	Allele 0	MAF	Altai Neanderthal sequence	Denisovan sequence	Caucasian individuals with at least 1 non-HuRef site																																																										
rs145261649	1	33116103	A	C	0.00122																					1																																									
rs138735449	1	33116168	T	C	0.00122																																																												
rs41265851	1	33116206	T	A	0.02317					1						1				1	1		1	1												1					1		1	1		1		1		1																
rs113601952	1	33116359	G	A	0.00976				1				1														1					1																	1					1		1										
rs188230717	1	33116546	T	C	0.00122																																																											
chr1:33116575	1	33116575	A	G	0.00122																																1																														
chr1:33116635	1	33116635	T	C	0.00122																																																											
chr1:33116811	1	33116811	G	A	0.00122																																																					1						
chr1:33116852	1	33116852	C	T	0.01829										1				1						1				1	1				1		1			1					1						1						1								1	
chr1:33116853	1	33116853	T	C	0.00122																																							1																			
chr1:33116961	1	33116961	A	G	0.00244																																																											
chr1:33117052	1	33117052	A	G	0.00122														1																																														
rs583114	1	33117071	T	G	0.02683																																																											
rs12407673	1	33117092	T	C	0.03415	1		1			1			1			1	1				1			1			1	1		1		1					1			1							1			1	1		1			1	1		1	1		1		1
chr1:33117099	1	33117099	T	C	0.00244							1																										1																													
rs12407680	1	33117137	G	C	0.03415	1	1	1			1			1			1	1				1			1			1	1		1		1					1			1							1			1	1		1			1	1		1	1		1		1
rs664567	1	33117231	G	A	0.02683																																																											
chr1:33117343	1	33117343	T	C	0.00122																																																											
rs146406493	1	33117429	A	G	0.0061																																																											
chr1:33117683	1	33117683	A	G	0.00122																																																											
rs182390022	1	33117735	T	C	0.00122																																																								1			
rs186772311	1	33117743	G	A	0.00244																																																											

Table S3. Validation of SNP effects by CRISPR/Cas9

Gene	SNP	Estimated effect[1]	Change in ddPCR[2]	p-value	N clones Δ[3]
UQCC	chr20:33999719	0.17X	0.48X	0.0002	5 of 5
TDP2	chr6:24667167	0.11X	0.46X	0.0029	4 of 5
COMMD4	rs182080358	12.5X	1.53X	0.050	4 of 5
DHX29	chr5:54603837	12.9X	0.98X	0.43	1 of 5

[1] Inferred from difference in log2 fluorescence intensity between the individual with the mutation and all other 409 samples

[2] Average change in ratio of drop digital PCR counts relative to two reference genes normalized to unity

[3] Number of the 5 clones deviant in the expected direction (see Suppl. Fig. 6).

APPENDIX: R code

//Code for PEER Normalization

Input is 279 Caucasian samples with 11,056 genes, Quantile normalized with SAS/JMP

//std for genes within batches(279, 11056G), get lists of std by batch

```
res <- apply(qnm_edata, 2, tapply, qnm_edata$Batch, scale)
```

//combine each gene, then combine all genes

```
std=matrix(rep(0),279,11056)
for(i in 1:11056){
tmp=do.call("rbind",res[[i]])
std[,i]=tmp[order(row.names(tmp)),]}
colnames(std)=colnames(qnm_edata)[2,11057]
rownames(std)=rownames(qnm_edata)
```

//ComBat fit batch

```
library(SVA)
edata=t(std)
batch=pheno$Batch
modcombat = model.matrix(~1, data=pheno)
combat_edata = ComBat(dat=edata, batch=batch, mod=modcombat, par.prior=TRUE, prior.plots=FALSE)
```

//transpose data

```
edata_ComBat=as.matrix(combat_edata)
transpose_ComBat_edata=t(edata_ComBat)
ComBat_edata=as.data.frame(transpose_ComBat_edata)
```

//Fit Age and Sex

```
rr=matrix(rep(0),279,11056)
for(i in 1:11056){
 ff=lm(ComBat_edata[,i]~pheno$Gender+pheno$Age)
 rr[,i]=ff$res
 }
colnames(rr)=colnames(transpose_ComBat_edata)
rownames(rr)=rownames(transpose_ComBat_edata)
```

//Fit Axes (analysis not reported in paper, for comparison with SVA)

```
exp=rr
res=matrix(rep(0),279,11056)
```

```
Pmin=rep(1,11056)
Cmin=rep(1,11056)

for(i in 1:11056){
	Pmin[i]=1
	Cmin[i]=1
	for(j in 1:7){
		fit=lm(exp[,i]~Axes[,j])
		if(summary(fit)$coef[2,4]<Pmin[i]){
			Cmin[i]=j
			Pmin[i]=summary(fit)$coef[2,4]
			res[,i]=fit$res}
		}
	}
Bind=cbind(Cmin,Pmin)
colnames(res)=colnames(transpose_ComBat_edata)
rownames(res)=rownames(transpose_ComBat_edata)
write.csv(Bind,file="279CAU_qnm_std_ComBat_fitAgeSex_fitAxis_coef.csv")
write.csv(res,file="279CAU_qnm_std_ComBat_fitAgeSex_fitAxis.csv")


//PEER -- NK20

library(peer)
expr=res
model = PEER()
PEER_setPhenoMean(model,as.matrix(expr))
PEER_setNk(model,20)
PEER_update(model)
factors = PEER_getX(model)
weights = PEER_getW(model)
precision = PEER_getAlpha(model)
residuals = PEER_getResiduals(model)
pdf('PEER_Nk20_Model.pdf')
PEER_plotModel(model)
dev.off()
pdf('PEER_Nk20_Precision.pdf')
plot(precision)
dev.off()
```

//SNM normalization (on 546 samples, 14111 probes)

//Take mean of probes for each gene (on 546 samples, 14111 probes) using apply function in R

//Remove duplicate sample GG2_0043 and American Indian sample(GG1-000149) --> 544sample_SNM.csv

//Read tab-delimited files for expression, biological, adjustment and samples

```
chdwb.snm = read.csv("C:/Documents/544sample_SNM.csv", header=T, row.names=1)
chdwb.bio = read.csv("C:/Documents/chdwb_ageBethnR_bio.csv", header=T, row.names=1)
chdwb.adj = read.csv("C:/Documents/chdwb_ageBethnR_adj.csv", header=T, row.names=1)
chdwb.int = read.csv("C:/Documents/chdwb_ageBethnR_int.csv", header=T)
```

//Create model matrices and run SNM

```
int.var = chdwb.int
int.var$Array = as.factor(int.var$Array)
adj.var = model.matrix(~.,chdwb.adj)
bio.var = model.matrix(~.,chdwb.bio)
raw.data = as.matrix(chdwb.snm)
snmR.chdwb = snm(raw.data,bio.var,adj.var,int.var,rm.adj=TRUE,num.iter=10)
write.table(snmR.chdwb$norm.dat, file = "C:/Documents/544sample_SNM_stdBatch.csv", sep=",",
col.names=NA)
```

//For SVA analysis, this step was included to use a linear model to fit age and gender to 472 genes and 411 samples among 544 which have genotyping information

```
        //input "544sample_SNM" to R as "exp"
        //input phenotype file to R as "pheno"
rr=matrix(rep(0),411,472)
for(i in 1:472){
 ff=lm(exp[,i]~pheno$Gender+pheno$Age)
 rr[,i]=ff$res
 }
write.table(rr, file = "C:/Documents/544sample_SVA_stdBatch_agegender.csv", sep=",", col.names=NA)
```

//Fit Independent Common eQTL (within 1kb upstream and gene region) for 411 samples with genotyping information

//Prepare eQTL files with row as samples, column as each common eSNP (order by coordinates) --> input to R as "gen"

//Prepare expression file with 207 genes with common cis-eQTLs (order by coordinates)
        --> input to R as "exp"

//Prepare another expression file with the left 265 genes without cis-eQTLs

//Prepare a file with two columns, the first column is gene name, the second column is the number of common eQTLs. --> input to R as "count"

```
j=1
res=exp
Pmin=rep(1,207)
Cmin=rep(1,207)
for(i in 1:207){
	Cend=j+count[i,2]-1
	for (m in j:Cend){
		fit=lm(exp[,i]~gen[,m])
		if(summary(fit)$coef[2,4]<Pmin[i]&&summary(fit)$coef[2,4]<0.05){
		Cmin[i]=m
		Pmin[i]=summary(fit)$coef[2,4]
		res[,i]=fit$res}
		}
	exp[,i]=res[,i]
	j=Cend+1}
```

	//Check Pmin, if there is at least one Pmin<0.05, then re-do the whole steps, until all Pmins>=0.05 for all genes

	//Output res, combine with the 265 gene expression.

//385 samples who have expression levels.were included in 411 samples --> 385sample_SNM_stdBatch_fitEqtl

//**Fit the most significant Axis** to 385 samples

	//input "385sample_SNM_stdBatch_fitEqtl" to R as "exp"

	//Prepare axis file with 7 axes --> input to R as "Axes"

```
res=matrix(rep(0),385,472)
Pmin=rep(1,472)
Cmin=rep(1,472)
for(i in 1:472){
	Pmin[i]=1
	Cmin[i]=1
	for(j in 1:7){
		fit=lm(exp[,i]~Axes[,j])
		if(summary(fit)$coef[2,4]<Pmin[i]){
			Cmin[i]=j
			Pmin[i]=summary(fit)$coef[2,4]
			res[,i]=fit$res}
		}
	}
```

//res --> expression data for 385 samples after normalization and adjustment

//279 are Caucasians among 385 samples. --> 279CAU_SNM_stdBatch_fitEqtlAxis

//**Rare variant association test**

```
        //input "279CAU_SNM_stdBatch_fitEqtlAxis" as "exp"

        //input rare variant counts per gene per sample as "snp"

        //the below code split genes to 93 bins with 3 genes in each bin for each sample

count=c(rep(0,93))
for(i in 1:472){
Rk=rank(exp[i,],ties.method="first")
        for(j in 1:279){
                zz=as.integer((Rk[j]+2)/3)
                if(snp[i,j]!=0){
                count[zz]=count[zz]+snp[i,j]}
        j=j+1}
i=i+1}
perc=1:93
```

        //"count" --> rare variant counts in each bin

        //"perc" --> bin number (expression bin from lowest to highest)

//**Perform quadratic test on count and perc**

SAS/JMP -- Analyze -- Fit Y by X -- Fit polynomial – quadratic