# Gene expression levels are a target of recent natural selection in the human genome.

Supplementary Information

Sridhar Kudaravalli[1], Jean-Baptiste Veyrieras[1],
Barbara E. Stranger[2], Emmanouil T. Dermitzakis[2], Jonathan K. Pritchard[1,3]

Research Article

[1]Dept of Human Genetics, The University of Chicago,
920 E 58th St CLSC 507, Chicago IL 60637, USA.
[2]Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus, Hinxton,
Cambridge, CB10 1SA, UK.
[3]Howard Hughes Medical Institute.

Correspondence to skudarav@uchicago.edu, md4@sanger.ac.uk, pritch@uchicago.edu.

Keywords: Population Genetics, Recent Positive Selection,
eQTL mapping, Humans, iHS.

December 9, 2008

# Contents

# List of Figures



# List of Tables

# 1 Comparing HapMap Phase I and Phase II iHS signals

We applied a method for detecting very recent positive selection [Voight et al. 2006] to HapMap phase II data. We find that a majority of the signals signals reported in Phase I data were replicated in Phase II data. Voight et al. [2006] divided the genome into 100Kb segments and identified the top 1% of such 100Kb segments with the strongest evidence for selection in each population. We find that 64% (YRI), 76% (CEU) and 75% (ASN) of the autosomal 100Kb segments identified in phase I data are in the top 5% in Phase II data. In order to account for changes in the genome build and differing SNP densities, the top hits from phase I 100Kb segments were extended by 5Kb on either side. Then replication between Phase I and Phase II data improved ( 76%(YRI), 81%(CEU) and 82%(ASN)). Figure 1 shows a comparison of Phase I and Phase II iHS signals from chromosome 2 in CEU and is meant to serve as a representative example of the replication of iHS signals.

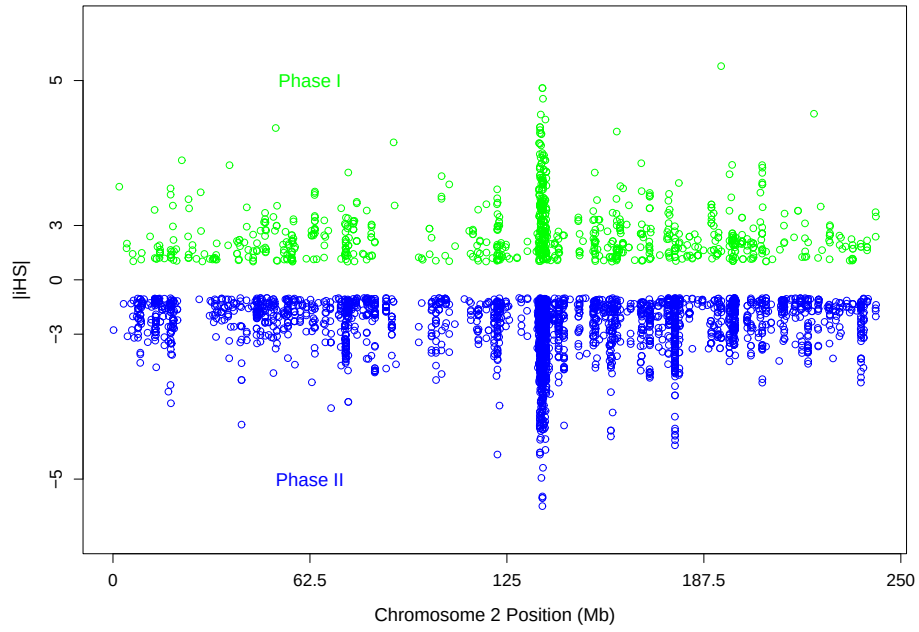


Figure 1: **Comparing the signals for selection in Phase I and Phase II HapMap data.** Chromosome 2 SNPs with an $|iHS| > 2.5$ in CEU are plotted to compare signals for selection between Phase I and Phase II HapMap data. The upper half of the plot shows Phase I data and the lower half shows Phase II data. The biggest peak in both plots corresponds to the Lactase region.

# 2 Calibration of the test for association between gene expression and SNP

In order to check if our test of association is well calibrated we used a permutation based method. Our test for association is based on a linear regression model where we regress genotypes from a SNP against the expression level of a gene within 100Kb as described in the Methods section of the paper. The permutation method involves randomizing the genotypes and repeating our test for association. When the proportion of eQTLs were plotted as a function of the frequency of the SNP we find no unusual deviation from the expected levels. We defined a SNP as an eQTL if the p-value of the association is less than 0.0001. Figure 2 shows that the proportion of the eQTLs across various frequency bins is mostly uniform at the 0.0001 level across the frequency bins which is in line with our expectation.

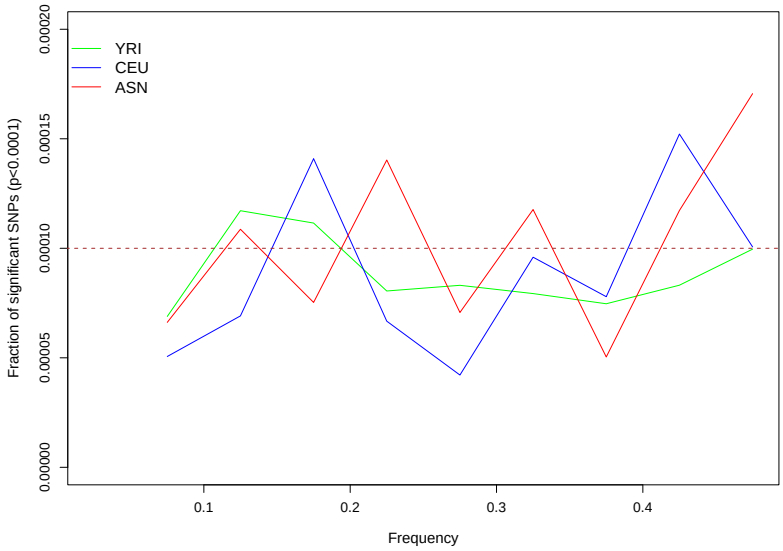


Figure 2: **Examining the calibration of gene expression association testing.** The proportion of significant eQTLs was determined by calculating the fraction of SNPs at a given frequency that are associated with the expression profile of a gene. The association is defined as being significant if the p-value $< 0.0001$.

To estimate the False Discovery Rate (FDR) of our method to detect eQTLs we used a permutation based method. FDR was assessed at two different levels: at the SNP level and at the gene level. At the SNP level the total numbers of SNPs that are strongly associated with gene expression ($p<0.0001$) were calculated in the permuted and observed data. If all SNPs were independent then the ratio of the numbers from the permuted (Expected False Positives) and the observed data would provide an estimate for the FDR. However, the SNP-level FDR is not well-defined because multiple SNPs can be associated sue to a single underlying eQTL.

At the gene level the total number of genes with at least one SNP that is is strongly associated (p<0.0001) with the gene is counted in the permuted and the observed data. The ratio of the number of such genes for permuted (expected false positives) and the real data gives an estimate of the FDR. The FDR levels at the SNP and the gene levels for each population are given in Table 1. We conducted this analysis for SNPs at different distances from genes: for genes with an eQTLs that lie strictly between 100Kb and 500Kb from the gene and genes with eQTLs that lie within 100Kb of the gene. We find that the FDR levels beyond 100Kb is extremely high and hence our primary analysis in the paper focuses on genes with eQTLs within 100Kb.

| | Number of SNPs | | | Number of Genes | | |
|---|---|---|---|---|---|---|
| | Observed | Expected | FDR | Observed | Expected | FDR |
| **< 100Kb** | | | | | | |
| CEU | 9259 | 283 | 0.031 | 638 | 115 | 0.180 |
| YRI | 8977 | 336 | 0.037 | 1060 | 186 | 0.176 |
| ASN | 24187 | 298 | 0.012 | 1289 | 94 | 0.073 |
| **> 100Kb** | | | | | | |
| CEU | 1241 | 1021 | 0.82 | 333 | 356 | 1.00 |
| YRI | 1242 | 1115 | 0.90 | 570 | 582 | 1.00 |
| ASN | 1738 | 1240 | 0.71 | 432 | 353 | 0.82 |

Table 1: **False Discovery Rate (FDR) of method for detecting eQTLs.** Gene expression data was permuted and our method for detecting eQTLs was repeated on this permuted data. FDR levels are assessed at two different levels. At the SNP level, we count the total number of SNPs that are strongly associated (p< 0.0001) with gene expression in the permuted dataset ("Expected"). This number is compared to the observed number of associated SNPs in the real data. At the gene level, the total number of genes that have at least one SNP strongly associated with their gene expression is calculated for the real ("Observed") and permuted data ("Expected"). The ratio of the "expected" to the "observed" numbers provides an estimate for the FDR. At the SNP level, due to LD, the interpretation of FDR levels becomes more complicated but it can serve as a guide post about the true FDR levels. This analysis was performed separately for each population and for genes that have SNPs that are strongly associated with the expression levels and that lie within 100Kb of a gene and those that lie between 100Kb and 500Kb from the gene. Clearly, the FDR levels are very high for genes that have eQTLs beyond 100Kb.

# 3 Alternate logistic regression models

We explored robustness of our results from the logistic regression model using various thresholds for the independent variables in the model. These include the distance to the transcription start site (dTSS), distance to transcription end site (dTES), $|iHS|$ value, cluster iHS threshold, p-value of the SNP-gene association to define an eQTL and measure of linkage disequilibrium (LD). We find that while each of these variables has an effect on the strength

of association between iHS and eQTLs, changing the threshold for one or more of these variables does not alter the direction of the association between iHS and eQTL. These analyses were done to show the direction of the association and hence bootstrapping was not conducted. The direction of the association between iHS and eQTL is robust to various thresholds except in the most extreme circumstances when the number of genes with an iHS and eQTL signal is very low.

Bootstrap analysis was performed only for various LD measures because of these variables LD is perhaps the most important variable that can generate a false positive association between iHS and eQTL as explained in the paper and hence its effect needs to be controlled very carefully. We incorporated several different measures of LD into the logistic regression model to assess the impact of iHS on the presence of an eQTL. Measures of LD include: (A) The number of SNPs that are in LD with the core SNP; (B) the point estimate of the recombination rate at each SNP; (C) the number of SNPs that are associated with the expression profile of a particular. For measure (A) LD was assessed using the $r^2$ statistic. For every SNP the number of other SNPs within 500Kb that have an $r^2$ greater than some threshold are counted. Such a measure informs us about how good a tag SNP a particular SNP is. Several thresholds of $r^2$ were used to define high LD. Our primary analysis uses this measure to control for LD and we used an $r^2 > 0.8$ as our threshold in counting the number of SNPs being tagged by a core SNP. For example, for SNP rs1234 we count how many SNPs within 500Kb from rs1234 have an $r^2 > 0.8$ with rs1234. This number is used as the surrogate for LD in our logistic regression model. This measure will reduce the effect size of iHS on eQTL because iHS signals are generally associated with higher LD levels. Measure (B) is the point estimate of the local recombination rate (referred to as $\rho$) [International HapMap Consortium 2007]. The recombination rate was estimated by the HapMap consortium using local LD patterns and is averaged across the three HapMap populations. $\rho$ estimates were downloaded from the HapMap website. Measure (C) is the number of SNPs within 500Kb of a gene that are significantly associated ($p< 0.0001$) with the expression level of the gene (referred to as neQTLs). The last measure of LD (neQTLs) generates a gene specific control for LD as opposed to the SNP specific measures for the first two methods. The odds ratios (OR) long with the 95% confidence intervals corresponding to the effect of iHS on eQTLs for the three measures of LD is tabulated in Table2.

In order to show that the choice of the distance from a gene or the strength of iHS signal does not alter the direction of our signal appreciably we conducted the logistic regression analysis at various distances from the gene and at different strengths of iHS signals (3, 4, and 5). The iHS signal strength is based on the clustering of SNPs with high $|iHS|$. For example, a SNP is placed in the top 5% if it has an $|iHS| > 2$ **and** the proportion of SNPs in a 75 SNP window around it is in the top 5% of the genome-wide distribution. We found in our simulations that while a single SNP with a high iHS may not be a good signal for selection a cluster of SNPs in close proximity with high iHS is a much better predictor of

| Population | LD measure | Odds Ratio | 95% CI | Number of Genes |
|---|---|---|---|---|
| YRI | $r^2$ | 2.49 | 1.23 - 4.27 | 35 |
| | $\rho$ | 3.29 | 1.50 - 8.62 | 35 |
| | neQTL | 2.54 | 1.55 - 2.57 | 35 |
| CEU | $r^2$ | 2.23 | 0.83 - 4.35 | 16 |
| | $\rho$ | 2.9 | 1.06 - 5.86 | 16 |
| | neQTL | 2.64 | 0.97 - 5.20 | 16 |
| ASN | $r^2$ | 1.45 | 0.82 - 2.38 | 47 |
| | $\rho$ | 1.90 | 1.13 - 3.00 | 47 |
| | neQTL | 1.76 | 1.07 - 2.63 | 47 |

Table 2: **Signal for enrichment of eQTLs among SNPs with signals of selection is robust to method used to control for LD.** For each population separately, we used 3 different measures of LD. For each measure the odds ratio (OR) of the effect of iHS on eQTL is calculated. The 95% confidence interval for this OR is estimated using a bootstrap approach. The three measures of LD are 1) For each SNP the number of other SNPs, within 500Kb, in LD ($r^2 > 0.8$) is calculated. This is denoted by $r^2$. 2) The point estimate of the recombination rate at each SNP. This is denoted by $\rho$. 3) The number of SNPs around a gene that are significantly associated with the expression profile of that gene. This is denoted by neQTL. Each of these terms were included in the logistic regression model as a surrogate for LD and the analysis repeated.

selection.

| | 500Kb | | | 200Kb | | | 100Kb | | | 50Kb | | | 10Kb | | | 50Kb | | | 1Kb | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes |
| **LD: $r^2$** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 1.6 | 8.4 | 45 | 1.9 | 11.6 | 39 | 2.5 | 14.1 | 35 | 2.6 | 13.0 | 32 | 2.8 | 8.3 | 22 | 3.5 | 7.1 | 20 | 3.1 | 3.0 | 14 |
| Top 2.5% | 1.8 | 9.6 | 31 | 2.1 | 11.1 | 29 | 3.0 | 14.6 | 26 | 2.9 | 12.6 | 24 | 3.1 | 7.7 | 17 | 3.6 | 6.0 | 15 | 4.5 | 3.0 | 10 |
| Top 1% | 2.5 | 11.4 | 16 | 2.6 | 10.9 | 15 | 3.8 | 13.3 | 13 | 3.3 | 10.6 | 12 | 2.6 | 5.1 | 9 | 2.7 | 3.4 | 8 | 5.0 | 1.8 | 4 |
| **LD: $\rho$** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 2.6 | 18.2 | 45 | 3.0 | 20.0 | 39 | 3.3 | 19.5 | 35 | 3.1 | 16.2 | 32 | 3.1 | 9.7 | 22 | 3.9 | 8.1 | 20 | 3.3 | 3.4 | 14 |
| Top 2.5% | 2.9 | 17.5 | 31 | 3.4 | 19.1 | 29 | 3.9 | 18.6 | 26 | 3.4 | 14.6 | 24 | 3.0 | 7.7 | 17 | 3.4 | 5.9 | 15 | 3.8 | 2.7 | 10 |
| Top 1% | 3.8 | 16.7 | 16 | 4.0 | 16.4 | 15 | 4.2 | 14.7 | 13 | 3.2 | 10.1 | 12 | 2.4 | 4.8 | 9 | 2.5 | 3.2 | 8 | 3.8 | 1.5 | 4 |
| **LD: neQTLs** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 2.3 | 15.0 | 45 | 2.4 | 14.6 | 39 | 2.5 | 13.4 | 35 | 2.2 | 9.6 | 32 | 2.4 | 6.6 | 22 | 3.3 | 6.7 | 20 | 3.1 | 3.1 | 14 |
| Top 2.5% | 2.9 | 16.7 | 31 | 3.2 | 16.7 | 29 | 3.4 | 15.3 | 26 | 2.7 | 10.9 | 24 | 2.7 | 6.5 | 17 | 3.3 | 5.6 | 15 | 4.0 | 2.8 | 10 |
| Top 1% | 3.7 | 15.5 | 16 | 3.8 | 14.6 | 15 | 3.5 | 11.8 | 13 | 2.4 | 7.2 | 12 | 2.1 | 3.9 | 9 | 2.5 | 3.2 | 8 | 4.6 | 1.7 | 4 |
| Genes with eQTLs | 1630 | | | 1215 | | | 1060 | | | 921 | | | 659 | | | 558 | | | 312 | | |

Table 3: **Odds Ratio of a SNP with high iHS being an eQTL at various distances from the gene in the YRI.** The odds ratios were calculated using logistic regression model at different strengths of signals for iHS. The numbers in the parentheses indicate the z-score from the logistic regression model, number of genes with at least one SNP with an iHS signal ($|iHS| > 2$ and lies in a cluster of other SNPs with high $|iHS|$) and an eQTL signal (p-value<0.0001). LD is included into the model is two different ways. The first surrogate for LD is the number of SNPs that were significantly associated (p< 0.0001) with the gene expression. The second measure is the number of SNPs in LD ($r^2 > 0.8$) with a SNP.

| | 500Kb | | | 200Kb | | | 100Kb | | | 50Kb | | | 10Kb | | | 50Kb | | | 1Kb | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes |
| **LD: $r^2$** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 1.7 | 9.0 | 22 | 1.6 | 7.4 | 18 | 2.2 | 10.9 | 16 | 2.5 | 10.8 | 15 | 2.0 | 4.4 | 10 | 2.0 | 3.4 | 9 | 2.3 | 1.6 | 6 |
| Top 2.5% | 2.1 | 9.8 | 12 | 1.9 | 7.9 | 10 | 3.1 | 12.7 | 9 | 3.6 | 11.9 | 9 | 2.3 | 3.8 | 5 | 3.0 | 3.9 | 5 | 3.0 | 1.8 | 4 |
| Top 1% | 3.2 | 12.1 | 5 | 2.2 | 7.8 | 5 | 3.5 | 10.4 | 4 | 3.5 | 8.6 | 4 | 6.7 | 3.8 | 2 | 6.0 | 2.7 | 2 | 0.7 | -0.4 | 2 |
| **LD: $\rho$** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 2.5 | 16.3 | 22 | 2.4 | 13.8 | 18 | 2.9 | 14.9 | 16 | 3.2 | 13.8 | 15 | 2.9 | 7.0 | 10 | 2.8 | 5.1 | 9 | 3.4 | 2.5 | 6 |
| Top 2.5% | 2.8 | 14.6 | 12 | 2.9 | 14.1 | 10 | 4.0 | 15.8 | 9 | 4.4 | 14.1 | 9 | 3.4 | 5.8 | 5 | 4.2 | 5.2 | 5 | 4.5 | 2.6 | 4 |
| Top 1% | 3.7 | 13.9 | 5 | 3.6 | 13.2 | 5 | 4.4 | 12.7 | 4 | 4.8 | 11.0 | 4 | 14.1 | 5.7 | 2 | 11.9 | 3.8 | 2 | 1.4 | 0.4 | 2 |
| **LD: neQTLs** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 2.3 | 13.6 | 22 | 2.4 | 13.1 | 18 | 2.6 | 12.5 | 16 | 2.6 | 10.3 | 15 | 1.8 | 3.7 | 10 | 1.7 | 2.5 | 9 | 2.2 | 1.6 | 6 |
| Top 2.5% | 3.0 | 14.3 | 12 | 3.5 | 15.3 | 10 | 4.2 | 15.1 | 9 | 4.1 | 12.0 | 9 | 2.1 | 3.4 | 5 | 2.5 | 3.1 | 5 | 2.8 | 1.7 | 4 |
| Top 1% | 4.5 | 14.9 | 5 | 4.3 | 13.6 | 5 | 4.5 | 11.6 | 4 | 4.1 | 8.7 | 4 | 5.4 | 3.5 | 2 | 5.8 | 2.7 | 2 | 0.8 | -0.2 | 2 |
| Genes with eQTLs | 970 | | | 741 | | | 638 | | | 571 | | | 427 | | | 367 | | | 229 | | |

Table 4: **Odds Ratio of a SNP with high iHS being an eQTL at various distances from the gene in the CEU.** The odds ratios were calculated using logistic regression model at different strengths of signals for iHS. The numbers in the parentheses indicate the z-score from the logistic regression model, number of genes with at least one SNP with an iHS signal ($|iHS| > 2$ and lies in a cluster of other SNPs with high $|iHS|$) and an eQTL signal (p<0.0001). LD is included into the model is two different ways. The first surrogate for LD is the number of SNPs that were significantly associated (p< 0.0001) with the gene expression. The second measure is the number of SNPs in LD ($r^2 > 0.8$) with a SNP.

| | 500Kb | | | 200Kb | | | 100Kb | | | 50Kb | | | 10Kb | | | 50Kb | | | 1Kb | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes | Odds ratio | z score | Genes |
| **LD: $r^2$** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 1.7 | 13.0 | 64 | 1.6 | 10.3 | 56 | 1.4 | 7.0 | 47 | 1.2 | 3.1 | 39 | 1.1 | 0.4 | 23 | 0.9 | -0.5 | 21 | 0.8 | -0.6 | 12 |
| Top 2.5% | 1.8 | 11.6 | 36 | 1.7 | 9.6 | 27 | 1.4 | 5.5 | 22 | 1.0 | 0.1 | 21 | 1.0 | -0.2 | 13 | 0.7 | -1.4 | 10 | 0.5 | -1.5 | 4 |
| Top 1% | 1.8 | 9.0 | 15 | 1.9 | 8.3 | 14 | 1.1 | 1.3 | 11 | 0.8 | -1.6 | 8 | 1.1 | 0.6 | 6 | 1.0 | 0.1 | 5 | 0.6 | -0.9 | 3 |
| **LD: $\rho$** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 2.1 | 20.2 | 64 | 2.1 | 17.9 | 56 | 1.9 | 12.8 | 47 | 1.6 | 7.7 | 39 | 1.4 | 3.0 | 23 | 1.2 | 1.2 | 21 | 1.0 | -0.0 | 12 |
| Top 2.5% | 2.5 | 19.1 | 36 | 2.4 | 16.9 | 27 | 2.0 | 10.5 | 22 | 1.4 | 3.8 | 21 | 1.5 | 2.3 | 13 | 1.1 | 0.3 | 10 | 0.7 | -0.8 | 4 |
| Top 1% | 3.0 | 16.5 | 15 | 2.9 | 14.4 | 14 | 1.6 | 4.8 | 11 | 1.2 | 1.1 | 8 | 1.6 | 2.2 | 6 | 1.5 | 1.4 | 5 | 0.9 | -0.1 | 3 |
| **LD: neQTLs** | | | | | | | | | | | | | | | | | | | | | |
| Top 5% | 2.0 | 16.9 | 64 | 2.0 | 14.6 | 56 | 1.8 | 10.3 | 47 | 1.5 | 6.2 | 39 | 1.1 | 0.5 | 23 | 0.8 | -1.1 | 21 | 0.9 | -0.4 | 12 |
| Top 2.5% | 2.8 | 20.5 | 36 | 2.7 | 18.0 | 27 | 2.1 | 10.9 | 22 | 1.5 | 4.5 | 21 | 1.5 | 2.5 | 13 | 1.2 | 0.6 | 10 | 0.7 | -0.7 | 4 |
| Top 1% | 3.3 | 17.2 | 15 | 3.0 | 14.1 | 14 | 1.8 | 5.5 | 11 | 1.2 | 1.6 | 8 | 1.7 | 2.3 | 6 | 1.7 | 1.7 | 5 | 1.0 | -0.1 | 3 |
| Genes with eQTLs | 1721 | | | 1438 | | | 1289 | | | 1181 | | | 960 | | | 847 | | | 534 | | |

Table 5: **Odds Ratio of a SNP with high iHS being an eQTL at various distances from the gene in the ASN.** The odds ratios were calculated using logistic regression model at different strengths of signals for iHS. The numbers in the parentheses indicate the z-score from the logistic regression model, number of genes with at least one SNP with an iHS signal ($|iHS| > 2$ and lies in a cluster of other SNPs with high $|iHS|$) and an eQTL signal (p<0.0001). LD is included into the model is two different ways. The first surrogate for LD is the number of SNPs that were significantly associated (p< 0.0001) with the gene expression. The second measure is the number of SNPs in LD ($r^2 > 0.8$) with a SNP.

In addition to the measures of LD, distance from the gene and strength of iHS cluster signal we also tested the robustness of our results to various thresholds of $|iHS|$ score and the eQTL p-value (Table 6. Here the $|iHS|$ refers to the actual iHS score and the clustering iHS signal ranks in the top 5%. We find that, in general, the signals in the YRI and CEU are more robust to most changes in the thresholds compared to the ASN. This could possibly because if the higher false positive rates in the signals for selection in the ASN.

| Population | $\|iHS\| > 2$ | | | $\|iHS\| > 2.5$ | | | $\|iHS\| > 3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| eQTL p-value< 0.0001 | Odds ratio | z-score | Genes | Odds ratio | z-score | Genes | Odds ratio | z-score | Genes |
| YRI | 2.5 | 14.1 | 35 | 2.7 | 11.1 | 26 | 4.0 | 11.3 | 15 |
| CEU | 2.2 | 10.9 | 16 | 3.2 | 11.1 | 12 | 2.5 | 5.0 | 7 |
| ASN | 1.4 | 7.0 | 47 | 1.3 | 3.2 | 26 | 1.3 | 1.8 | 17 |
| eQTL p-value< 0.00001 | Odds ratio | z-score | Genes | Odds ratio | z-score | Genes | Odds ratio | z-score | Genes |
| YRI | 2.5 | 10.8 | 20 | 2.2 | 6.1 | 14 | 3.0 | 6.7 | 8 |
| CEU | 2.8 | 10.9 | 10 | 3.4 | 9.1 | 8 | 2.3 | 3.2 | 6 |
| ASN | 0.9 | -1.1 | 30 | 0.7 | -2.7 | 16 | 0.3 | -3.9 | 6 |

Table 6: **Odds Ratio of a SNP with high $|iHS|$ being an eQTL using various thresholds of iHS.** We see a significant effect of iHS of the presence of eQTLs at various thresholds for iHS. The effect of iHS on eQTLs remains almost the same even if the threshold for eQTL p-value is changed. These results are generated using SNPs within 100Kb of the gene. In addition to each SNP having an iHS greater than the threshold they should also rank in the top 5% based on clustering of other high $|iHS|$ SNPs.

# 4 Comparing gene expression patterns across populations

The number of genes with at least one SNP associated with the gene expression level is the lowest among the CEU population. It is about 40% lower compared to YRI and ASN (638 Vs 1060 and 1289 respectively). Similarly the number of genes with at least one SNP with strong evidence for association that also has significant evidence for selection is lowest in CEU. It is less than 50% of the number of such genes in YRI and ASN(16 Vs 35 and 47 in YRI and ASN, respectively). So in order to check if the expression patterns are significantly different between the CEU and the other two populations we conducted principal component analysis on the expression patterns from the three populations. Figure 3 clearly shows that the pattern of expression in the CEU is quite different from that for YRI and ASN. In addition to this we also find that among the probes that differ significantly in the mean expression levels between CEU and the other two populations, CEU had the higher mean expression level in 80% of the probes. Differences in the expression levels for each pairwise comparison of the populations were assessed using the Kruskal-Wallis test and probes were classified as having significantly different mean expression levels if the p-value is $< 10^{-10}$. We also find that among probes that have significantly different mean expression

levels between YRI and ASN, 65% of the probes have higher mean expression levels in the ASN. So the non-African populations have higher expression levels for probes that have significantly different expression levels between the populations.

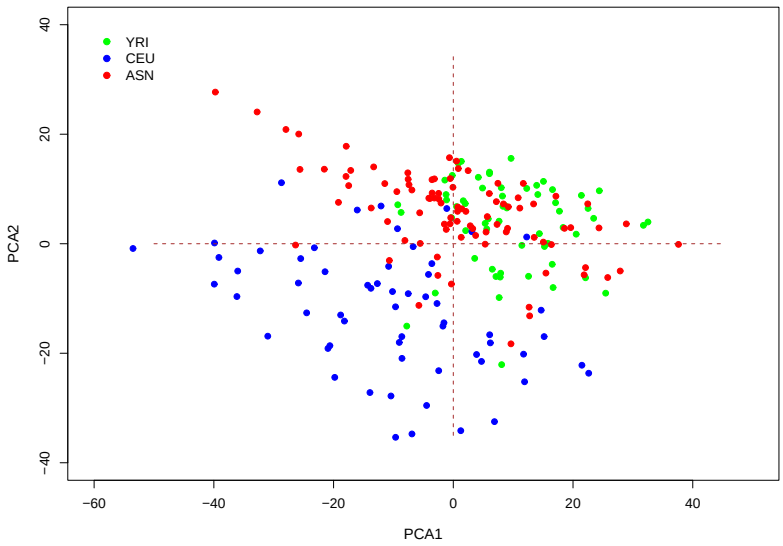


Figure 3: **Patterns of expression are significantly different in the CEU compared to ASN and YRI.** Principal components analysis of the expression patterns in the three HapMap population groups show that the CEU are separated from the other two populations. Each data point corresponds to an individual from the the HapMap populations.

# 5 $F_{st}$-based analysis

Extreme differences in the allele frequencies between populations may arise due to a selective sweep in one of the populations. So if regulation of gene expression is a common target of natural selection then SNPs with high frequency differences across populations are more likely to be associated with gene expression profiles of neighboring genes.

We implemented a robust approach to check if high $F_{st}$ SNPs are more likely to be eQTLs. For each pairwise population comparison the SNPs were binned according to the frequency difference between the two populations. Then for each such bin the expected number of eQTLs in one of the two populations was computed based on the frequency spectrum of the SNPs in that bin and the observed rate of eQTLs in that populations. This expected number of eQTLs is compared to the observed number. For example, for a CEU-YRI comparison, SNPs with an absolute frequency difference $f$ are collected into a bin. Then the expected number of eQTLs for this bin in the YRI (or CEU) are calculated based on the frequency spectrum of these SNPs in the YRI (or CEU) multiplied by the probability of observing an eQTL in the YRI (or CEU) across the frequency spectrum. For

each population we compared to expected number of eQTLs to those observed and found that at higher frequency differences we see a greater than expected number of eQTLs. This is especially true at the extreme frequency differences as seen in Figures 4, 5, and 6.

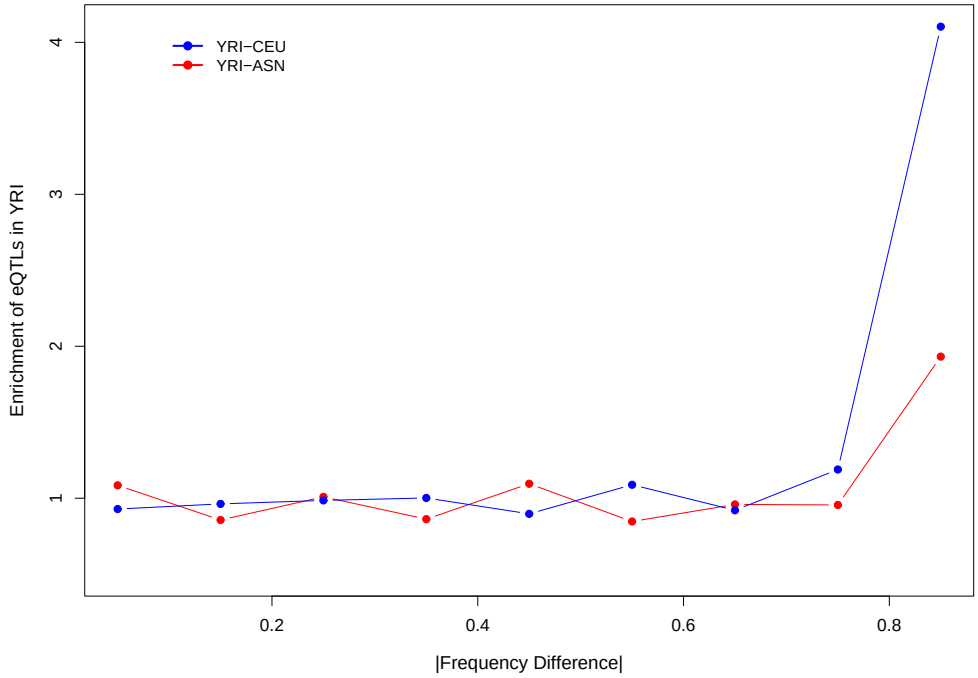


Figure 4: **As the difference in SNP frequencies between YRI and the other two populations increases there is an enrichment of SNPs associated with gene expression. The ratio of the observed to expected number of eQTLs is plotted for each frequency bin. If there is no enrichment of eQTLs the ratio of the two should be close to 1. A value $> 1$ indicates enrichment of eQTLs and a value $< 1$ indicates deficiency of eQTLs.**

# 6 Effect of local SNP density

As an episode of strong selection has an effect on the local SNP density, we wanted to check if local SNP density has a bearing on the relationship between iHS and eQTLs. For each gene the number of SNPs within 100Kb with a minimum allele frequency greater than 0.05 and whose ancestral state is known was computed. This number was used as an additional covariate in the logstic regression model. We found a significant but inconsistent effect (Table 7) of the local SNP density on the relationship between iHS and *cis* eQTL.

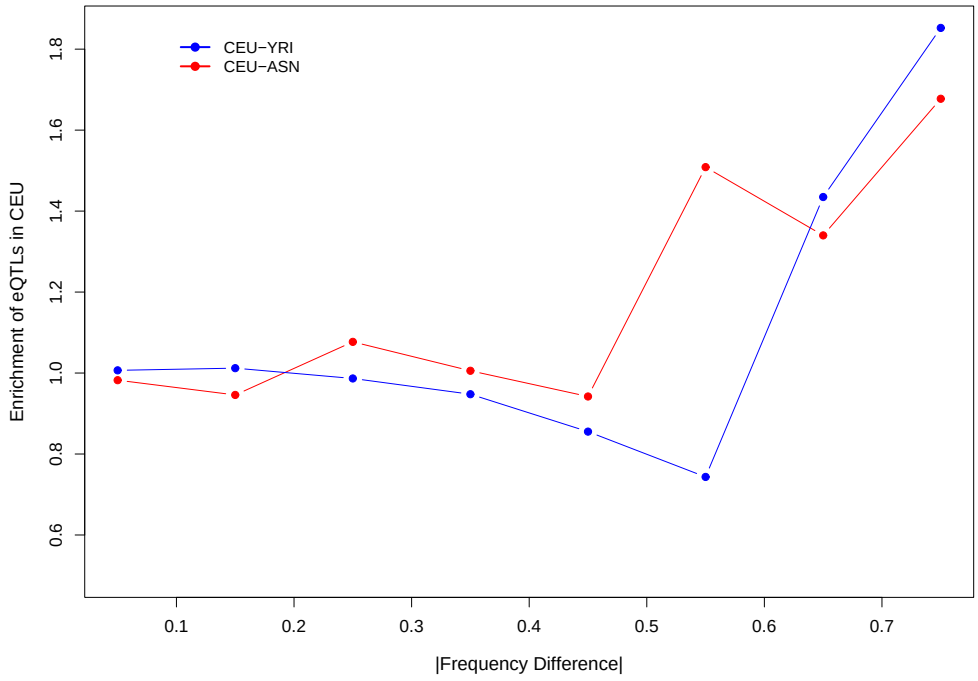


Figure 5: **As the difference in SNP frequencies between CEU and the other two populations increases there is there is an enrichment of SNPs associated with gene expression. The ratio of the observed to expected number of eQTLs is plotted for each frequency bin. If there is no enrichment of eQTLs the ratio of the two should be close to 1. A value $> 1$ indicates enrichment of eQTLs and a value $< 1$ indicates deficiency of eQTLs.**

# 7 Purifying Selection

Purifying selection is characterized by a shift in the frequency spectrum towards rarer alleles. This is because new deleterious mutations are eliminated constantly and are tolerated only at low frequencies, depending on the effective population size. It is possible that purifying selection also targets gene expression regulatory regions. In order check if such regions are important targets of purifying selection a straightforward comparison of eQTLs across frequency bins is not appropriate because the power to detect an eQTL is reduced as the minimum allele frequency decreases. So we implemented an alternate scheme where the expression level of each heterozygote was compared against homozygotes of the common allele of the same SNP. The distribution of expression levels of the homozygotes was normalized and the expression level of each heterozygote was converted to a z-score based on the normalized homozygote scores. So if the low frequency SNPs have disproportionately high effect on the expression levels of a gene then we should see a skew in the tails of the distribution of the z-scores for heterozygotes from the low frequency SNPs compared to high

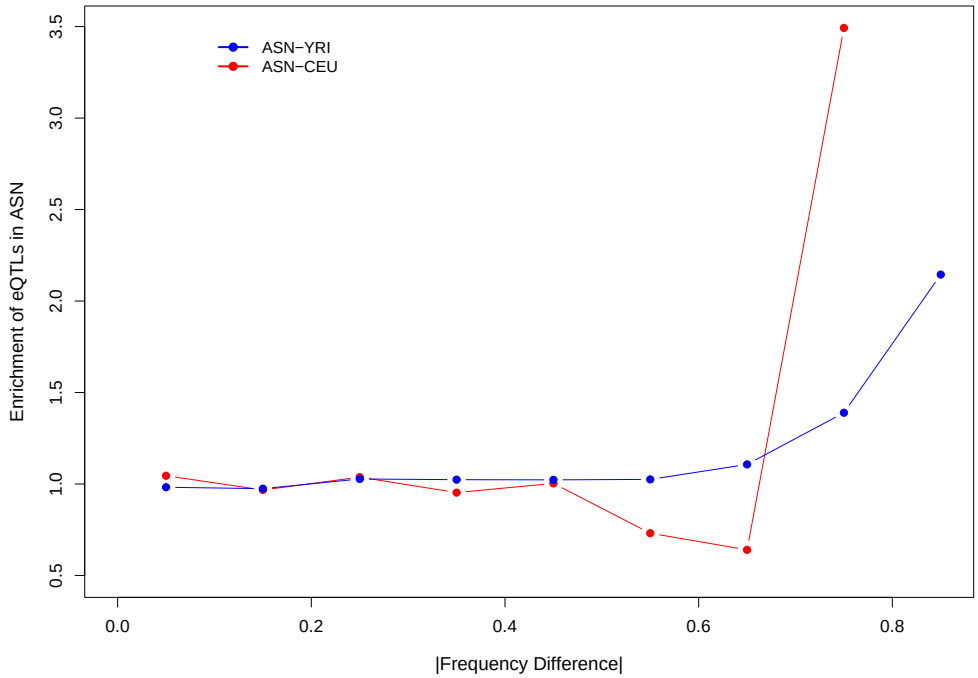


Figure 6: **As the difference in SNP frequencies between ASN and the other two populations increases there is an enrichment of SNPs associated with gene expression. The ratio of the observed to expected number of eQTLs is plotted for each frequency bin. If there is no enrichment of eQTLs the ratio of the two should be close to 1. A value $> 1$ indicates enrichment of eQTLs and a value $< 1$ indicates deficiency of eQTLs.**

frequency SNPs. Only SNPs with at least 30 homozygotes of the common allele were analyzed and that lie within 20Kb of the transcription start site of a gene were analyzed. Figure 7 was generated using this approach for all the three populations. We see slight evidence for a skew in the tails of the expression levels for low frequency heterozygotes compared to the remaining SNPs. The lack of very convincing evidence is perhaps not surprising considering the fact that we do not have complete SNP information across the genome and on average a low frequency SNP would be expected to be a poorer tag SNP compared to a high frequency SNP.

# 8 Enrichment of eQTLs in Most Conserved regions

In order to check if conserved regions are more likely to harbor eQTLs we checked to see if SNPs that are strongly associated ($p<0.0001$) with gene expression are more likely to be with conserved regions compared to outside such regions. The physical coordinates for conserved regions were downloaded from the UCSC database ([Karolchik et al. 2003]).

| Population | Effect Size Without SNP density | Effect Size With SNP density |
|---|---|---|
| YRI | 2.41 | 2.03 |
| CEU | 2.29 | 2.85 |
| ASI | 1.41 | 1.4 |

Table 7: The effect of the local SNP density on the relation ship between iHS and *cis* eQTL was examined by including and excluding the local SNP density term in the model.

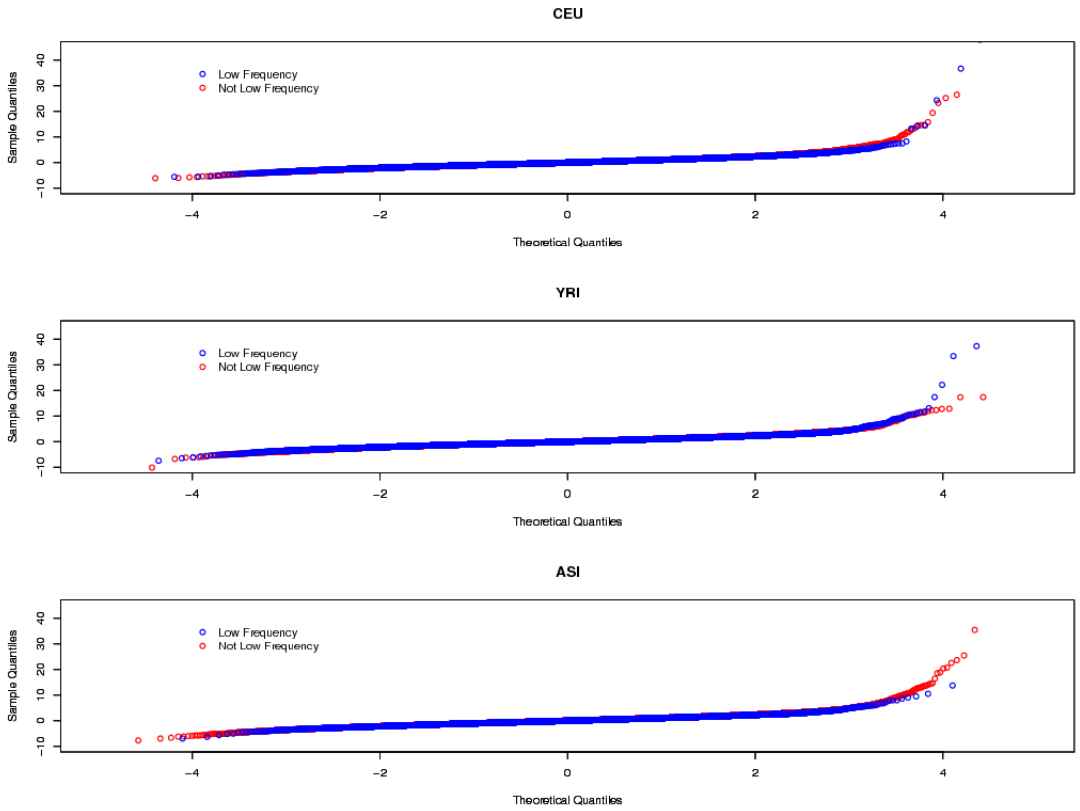


Figure 7: **Comparing the expression profiles of low frequency SNPs with the remaining SNPs** The expression levels of heterozygotes from low frequency SNPs ($n <= 5$,n-number of heterozygotes) and SNPs that are at higher frequencies. Normalized scores were generated as described in the text.

These represent top 5% of the most conserved regions in the human genome. We find that SNPs with strong association with gene expression are significantly more likely to be present in the most conserved region compared to outside such regions. This is true for eQTLs in all three populations (Table 8). When SNPs that lie within a gene, whose expression levels are being tested for association, are excluded we still see an enrichment of eQTLs among most conserved regions. For example, if we are testing for association between SNPs and the expression profile of Gene XYZ; all SNPs that lie within Gene XYZ are excluded from the

analysis. But SNPs may lie in other neighboring genes. So relative to each gene, SNPs that lie in the noncoding most conserved regions are more likely to be associated with expression levels compared to the noncoding SNPs that lie outside most conserved regions.

| | YRI | | CEU | | ASN | |
|---|---|---|---|---|---|---|
| | Not MC | MC | Not MC | MC | Not MC | MC |
| Not eQTL | 169943 | 7644 | 92694 | 4400 | 157199 | 7072 |
| eQTL | 8436 | 541 | 8717 | 542 | 22926 | 1261 |
| | p-value: $7.6 \times 10^{-15}$ | | p-value: $7.7 \times 10^{-9}$ | | p-value: $1.4 \times 10^{-10}$ | |
| Not eQTL | 135360 | 5863 | 74127 | 3316 | 129164 | 5593 |
| eQTL | 5298 | 317 | 5930 | 354 | 16189 | 807 |
| | p-value: $4.5 \times 10^{-8}$ | | p-value: $4.8 \times 10^{-7}$ | | p-value: $2.5 \times 10^{-4}$ | |

Table 8: **Comparing if eQTLs are enriched among most conserved regions in noncoding regions.** SNPs that are significantly associated (p< 0.0001) with gene expression are significantly more likely to be present in most conserved regions. This holds true even when the data is restricted to SNPs that lie outside the gene whose expression is being tested for association (see text). p-values were generated using a chi-square test of the contingency table for each population.

# 9 Gene Ontology Analysis

We also conducted analysis to identify gene ontology (GO) categories that are enriched for our signal [Thomas et al. 2003 2006]. In order to test if any our set of genes with eQTLs centered around signals for selection are enriched for biological processes we conducted gene ontology (GO) analysis. Using the stringent threshold that was used for our main analysis results only in about 30 genes that have eQTLs within signals for selection in each of the populations. So in order to increase the number of genes, the gene expression p-value threshold was increased to 0.001 from 0.0001. The set of genes for GO analysis were chosen such that they have at least one SNP with a p-value<0.001, an $|iHS| > 2$ and lies in a cluster of high $|iHS|$ SNPs. There are 92, 138 and 123 such genes in the CEU, YRI and ASN respectively. These sets of genes were contrasted against the genes that have at least 1 SNP with a gene expression p-value<0.001 but not necessarily have a high $|iHS|$. There are 4145, 6328 and 5044 such genes in the CEU, YRI and ASN respectively. The results of the GO analysis for each population are tabulated below (Table 9).

No category is significant after Bonferroni correction for multiple testing, however several of the top categories overlap with those identified in previous studies [Clark et al. 2003; Haygood et al. 2007; Voight et al. 2006] including categories related to carbohydrate and steroid metabolism in the YRI, mRNA processing, spermatogenesis and gametogenesis in

the CEU and sulfur metabolism in the ASN. GO categories related to immune function are significantly enriched in the CEU and ASN.

| GO category | YRI | CEU | ASN |
|---|---|---|---|
| Amino acid activation | | 0.01 | |
| Amino acid catabolism | 0.048 | | |
| Blood clotting | | 0.044 | |
| Carbohydrate metabolism | 0.00135 | | |
| Cell cycle | 0.018 | | |
| Cell proliferation and differentiation | 0.034 | | |
| Cell structure | | 0.049 | |
| Chromosome segregation | | 0.015 | |
| Coenzyme and prosthetic group metabolism | 0.017 | | |
| Constitutive exocytosis | | 0.04 | |
| DNA metabolism | 0.0074 | | |
| DNA replication | 0.00084 | | |
| Ferredoxin metabolism | 0.021 | | |
| Gametogenesis | | 0.011 | |
| Immunity and defense | | | 0.0046 |
| MHCII-mediated immunity | | 0.021 | 0.0026 |
| MHCI-mediated immunity | | | 0.00015 |
| mRNA end-processing and stability | | 0.0056 | |
| mRNA polyadenylation | | 0.0056 | |
| Other carbohydrate metabolism | 0.018 | | |
| Other transport | | 0.039 | |
| Pentose-phosphate shunt | | 0.04 | |
| Pre-mRNA processing | | 0.045 | |
| Protein glycosylation | 0.021 | 0.022 | |
| Regulation of lipid, fatty acid and steroid metabolism | 0.017 | | |
| Small molecule transport | | 0.025 | |
| Spermatogenesis and motility | | 0.0089 | |
| Steroid metabolism | 0.035 | | |
| Stress response | | | 0.046 |
| Sulfur metabolism | | | 0.026 |
| T-cell mediated immunity | | 0.046 | 0.0007 |
| Vitamin metabolism | 0.000274 | | |

Table 9: **Biological Process Gene Ontology categories that are enriched for signals of selection overlapping eQTLs.** Genes that have an eQTL with a p-value$< 0.001$ and evidence for selection that ranks in the top 5% within each population were analyzed for enrichment of GO categories. All categories that show an enrichment with a p-value$< 0.05$ in any population are included in the above table. Empty fields for a population indicate lack of significance of that category.

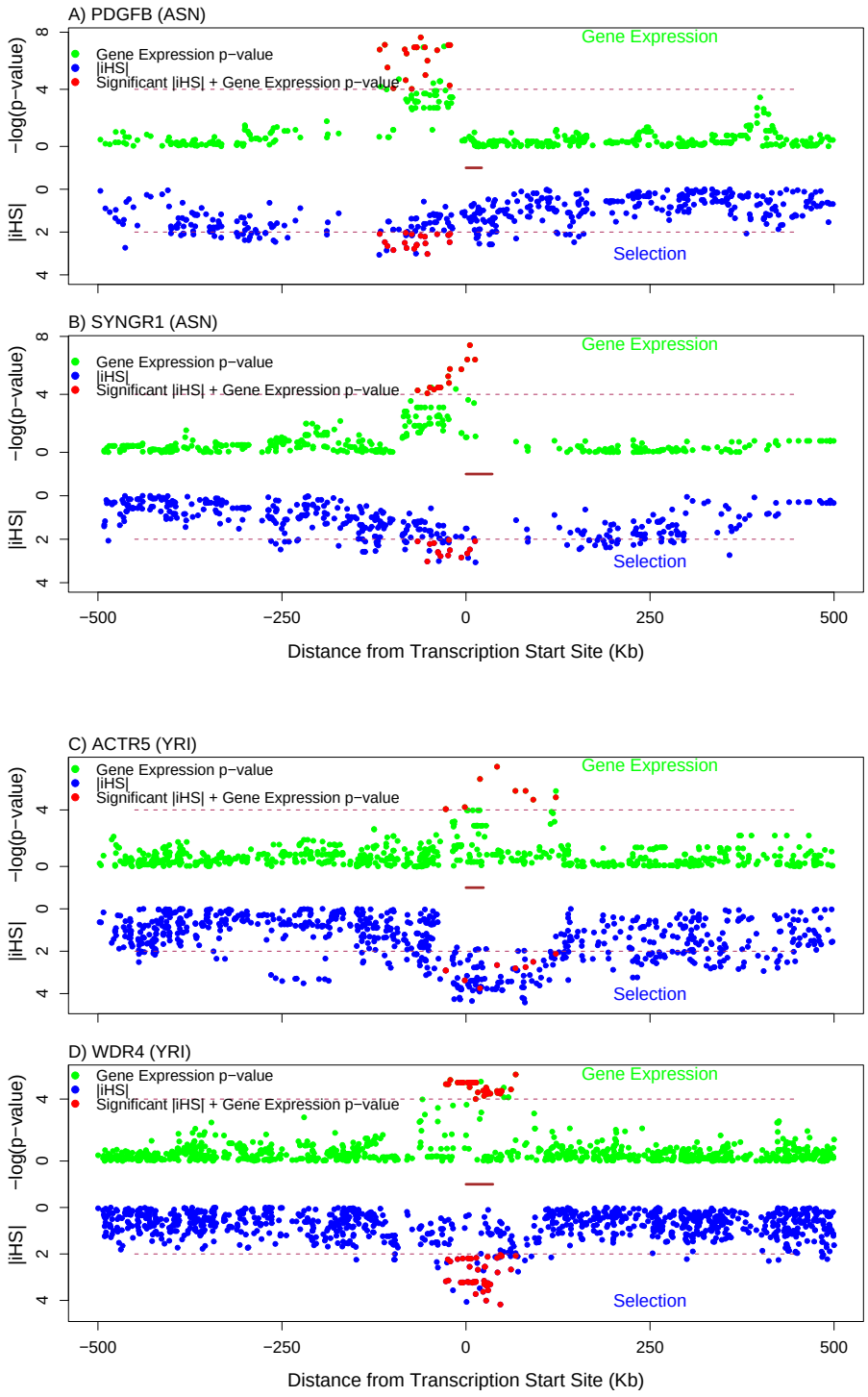


Figure 8: **More examples in which an eQTL is centered on a strong signal of selection.** The upper half of each plot shows the strength of association between SNPs and gene expression levels (plotted as -$\log_{10}$(ps) of the indicated gene. The lower half of each plot indicates $-|iHS|$ scores for the same set of SNPs. The red points indicate SNPs that are both strongly associated with expression (p$< 10^{-4}$) and have high $|iHS|$ ($> 2$). The positions of the genes of interest are indicated by the red bars at the center of each plot. **(A)** Data from PDGFB (ASN). **(B)** Data from SYNGR1 (ASN). **(C)** Data from ACTR5 (YRI). **(D)** Data from WDR4 (YRI).

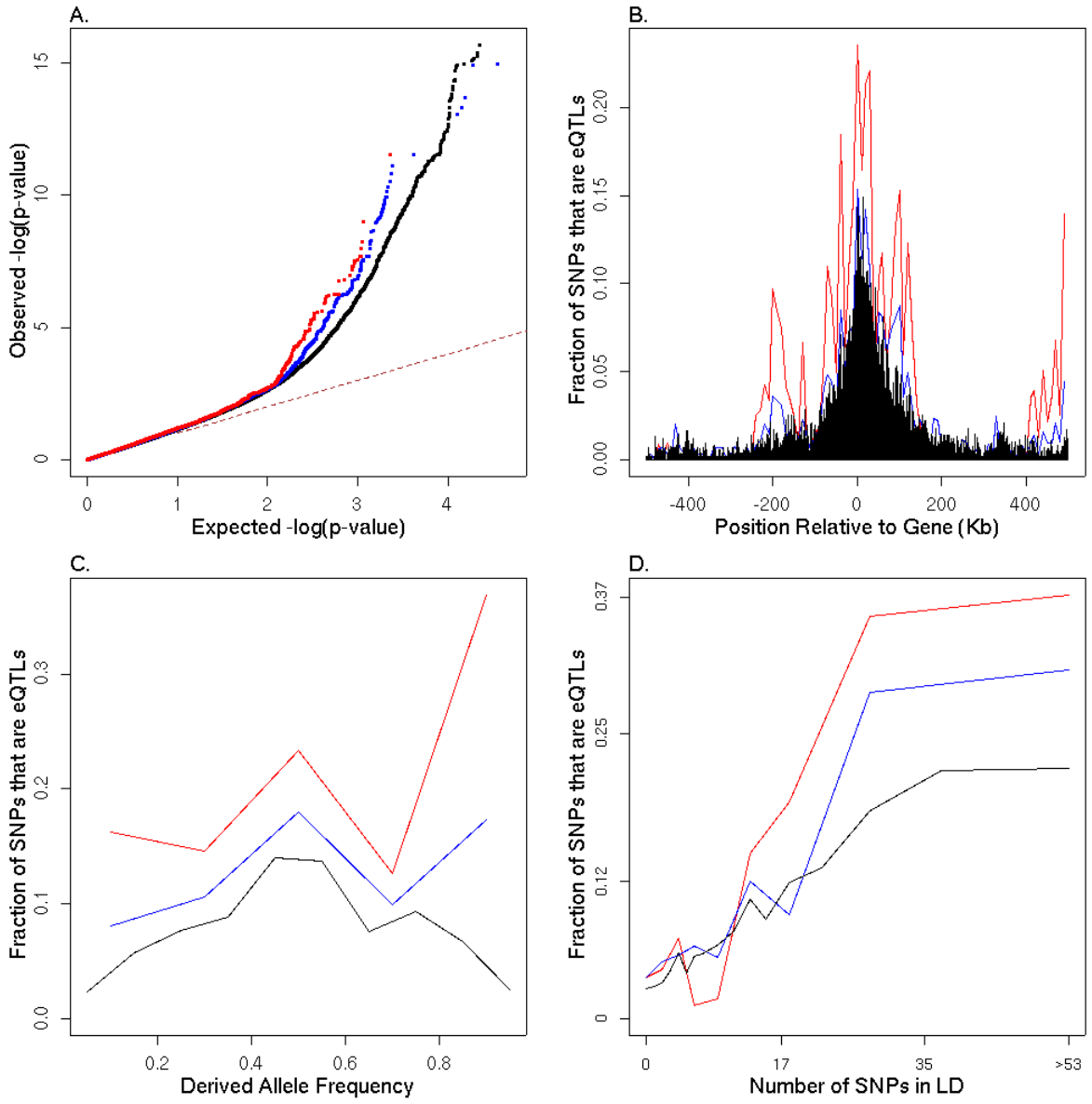


Figure 9: **The abundance of eQTL signals in SNPs with and without evidence for selection in CEU.** In each plot black indicates SNPs without any evidence for selection, blue is for SNPs with an $|iHS| > 2$ and red is for SNPs with significant evidence for selection as ascertained by clustering of high $|iHS|$ SNPs. **(A)** Plots the expected distribution of the p-values from the linear regression model for *cis* eQTLs against the observed p-values. The blue points show the p-values for SNPs with high $|iHS|$. The red points plot the distribution for SNPs with significant evidence for selection based on clustering of SNPs with high $|iHS|$. **(B)** SNPs with high iHS show an enrichment for eQTLs after at various distances from the transcription start site. **(C)** SNPs with high iHS tend to be enriched for eQTLs after controlling for allele frequency. **(D)** SNPs with high iHS show greater levels of eQTLs after controlling for LD levels, as measured by the number of SNPs in high LD ($r^2 > 0.8$).

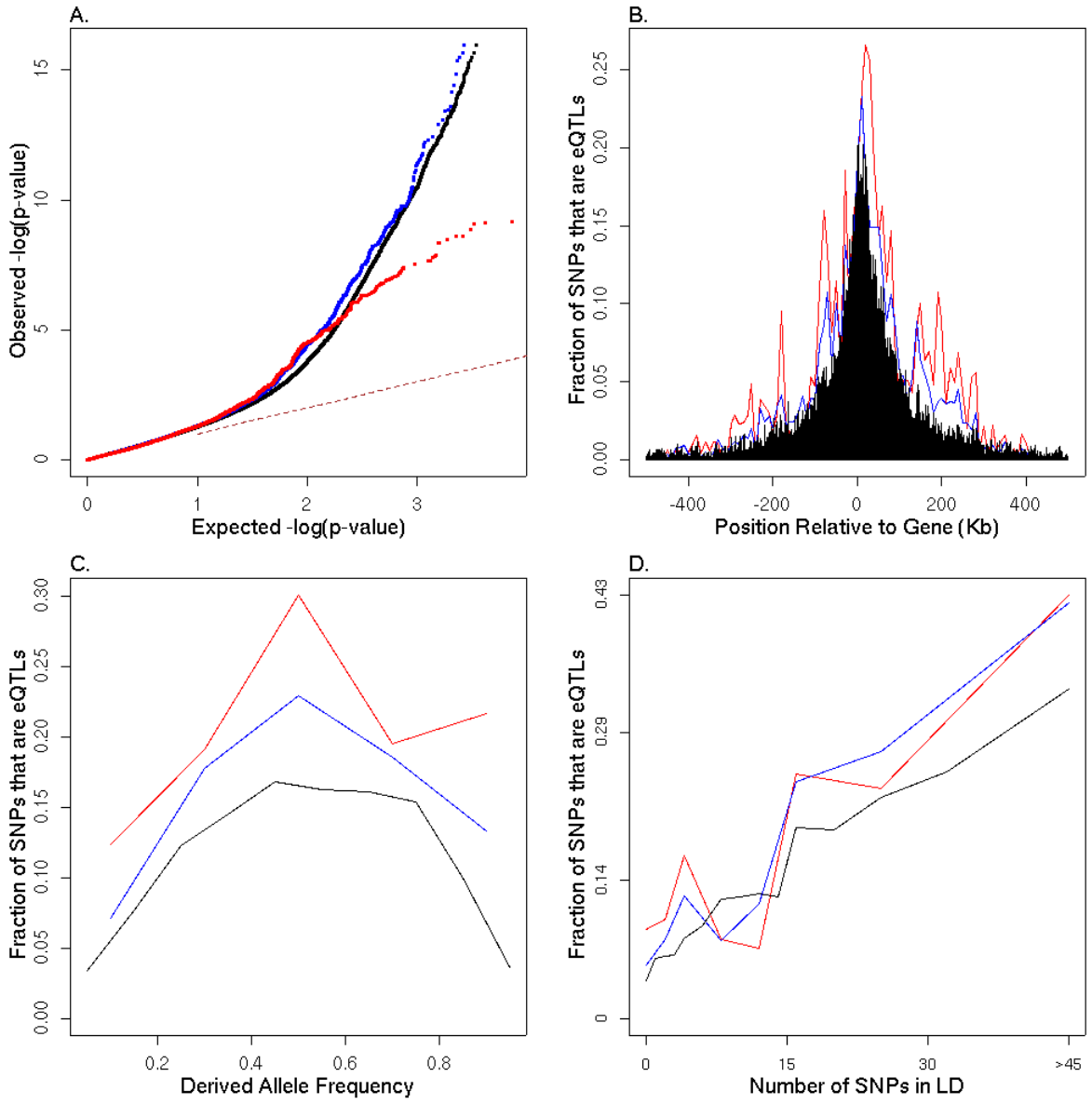


Figure 10: **The abundance of eQTL signals in SNPs with and without evidence for selection in ASN.** In each plot black indicates SNPs without any evidence for selection, blue is for SNPs with $|iHS| > 2$ and red is for SNPs with significant evidence for selection as ascertained by clustering of high $|iHS|$ SNPs. **(A)** Plots the expected distribution of the p-values from the linear regression model for *cis* eQTLs against the observed p-values. The blue points show the p-values for SNPs with high $|iHS|$. The red points plot the distribution for SNPs with significant evidence for selection based on clustering of SNPs with high $|iHS|$. **(B)** SNPs with high iHS show an enrichment for eQTLs after at various distances from the transcription start site. **(C)** SNPs with high iHS tend to be enriched for eQTLs after controlling for allele frequency. **(D)** SNPs with high iHS show greater levels of eQTLs after controlling for LD levels, as measured by the number of SNPs in high LD ($r^2 > 0.8$).

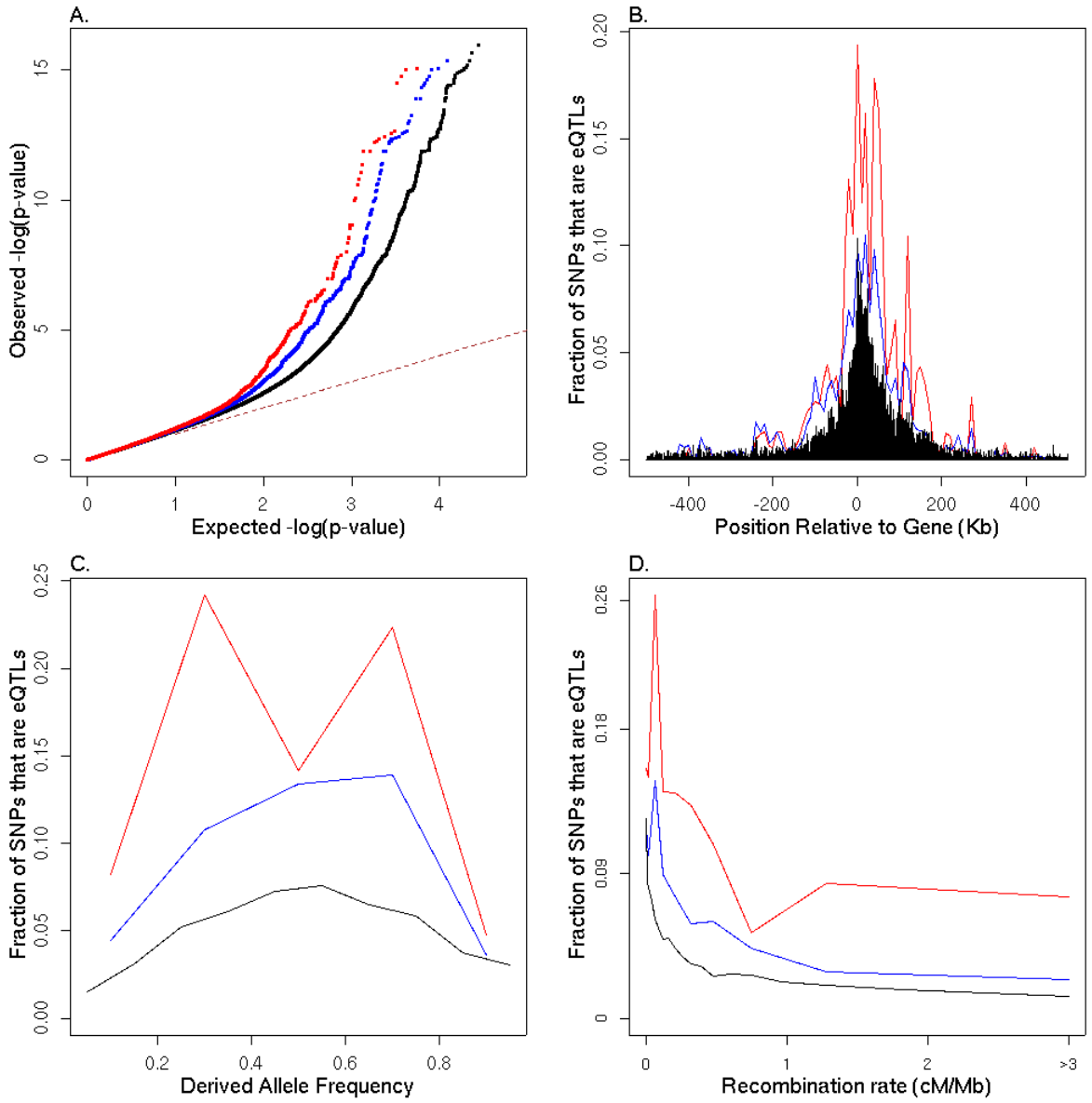


Figure 11: **The abundance of eQTL signals in SNPs with and without evidence for selection in YRI. LD is controlled using the local recombination rate.** In each plot black indicates SNPs without any evidence for selection, blue is for SNPs with high $|iHS|$ and red is for SNPs with significant evidence for selection as ascertained by clustering of high $|iHS|$ SNPs. **(A)** Plots the expected distribution of the p-values from the linear regression model for *cis* eQTLs against the observed p-values. The blue points show the p-values for SNPs with high $|iHS|$. The red points plot the distribution for SNPs with significant evidence for selection based on clustering of SNPs with high $|iHS|$. **(B)** SNPs with high iHS show an enrichment for eQTLs after at various distances from the transcription start site. **(C)** SNPs with high iHS tend to be enriched for eQTLs after controlling for allele frequency. **(D)** SNPs with high iHS show greater levels of eQTLs after controlling for LD levels, as measured by the local recombination rate.

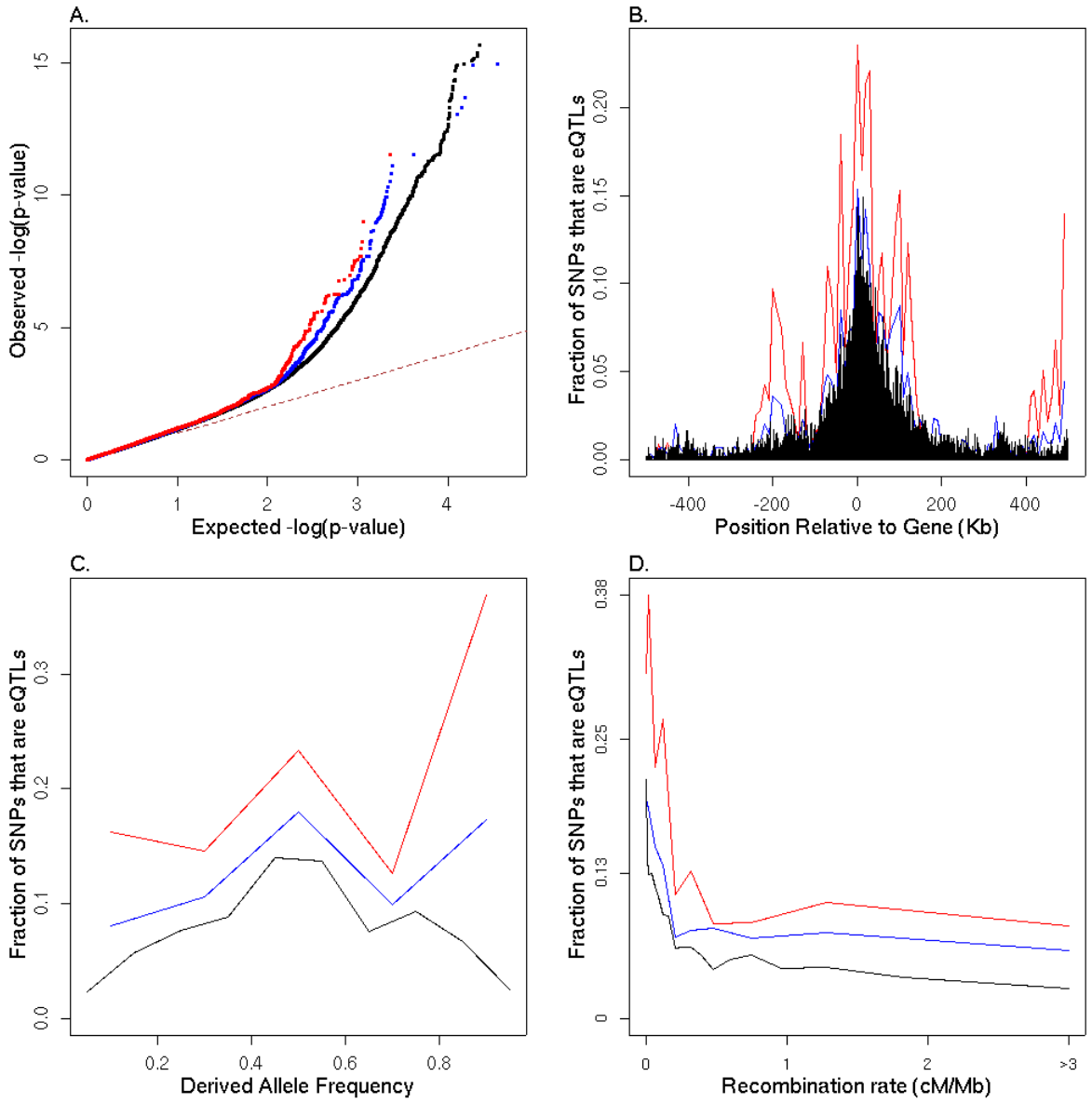


Figure 12: **The abundance of eQTL signals in SNPs with and without evidence for selection in CEU. LD is controlled using the local recombination rate.** In each plot black indicates SNPs without any evidence for selection, blue is for SNPs with high $|iHS|$ and red is for SNPs with significant evidence for selection as ascertained by clustering of high $|iHS|$ SNPs. **(A)** Plots the expected distribution of the p-values from the linear regression model for *cis* eQTLs against the observed p-values. The blue points show the p-values for SNPs with high $|iHS|$. The red points plot the distribution for SNPs with significant evidence for selection based on clustering of SNPs with high $|iHS|$. **(B)** SNPs with high iHS show an enrichment for eQTLs after at various distances from the transcription start site. **(C)** SNPs with high iHS tend to be enriched for eQTLs after controlling for allele frequency. **(D)** SNPs with high iHS show greater levels of eQTLs after controlling for LD levels, as measured by the local recombination rate.

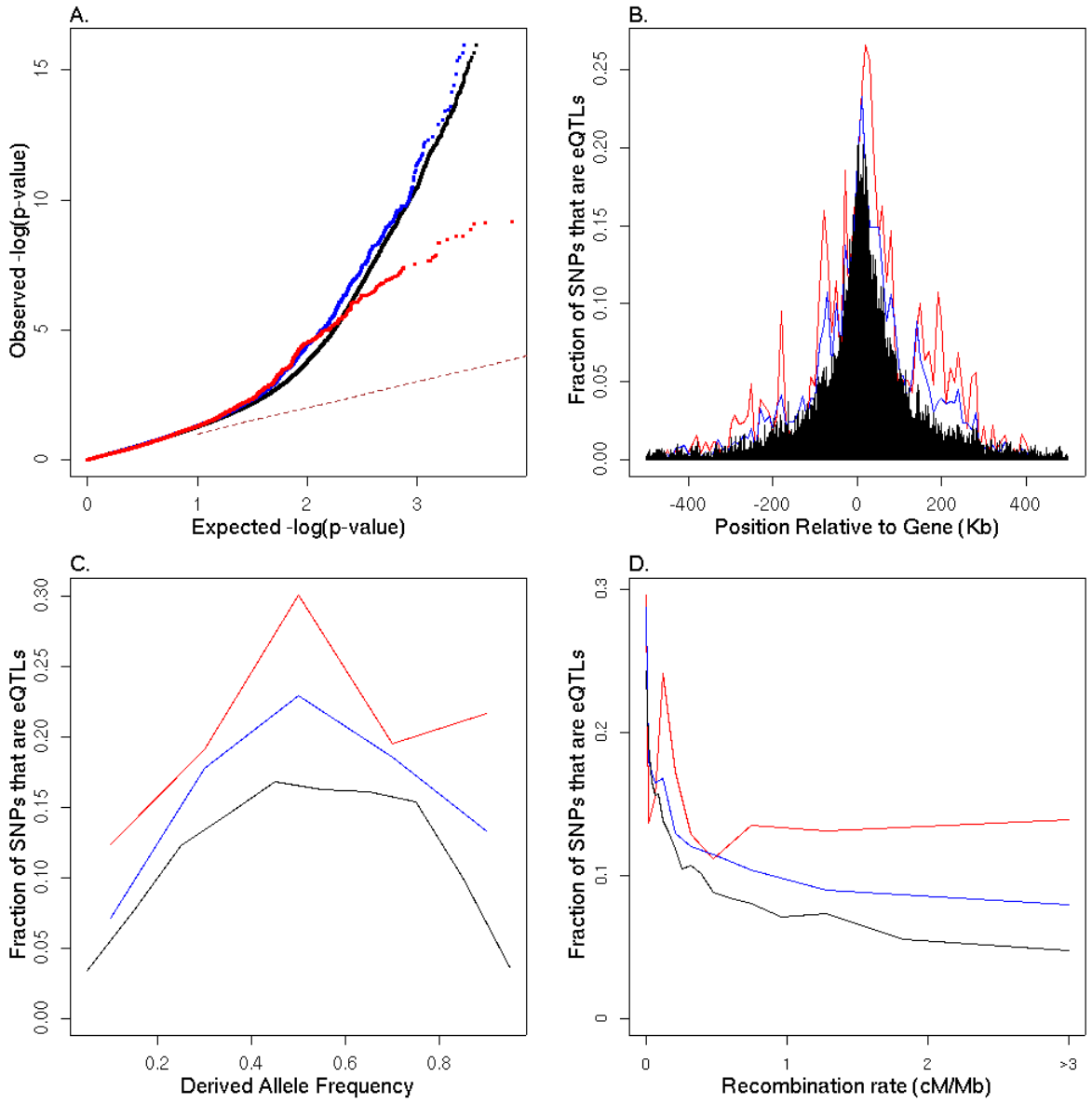


Figure 13: **The abundance of eQTL signals in SNPs with and without evidence for selection in ASN. LD is controlled using the local recombination rate.** In each plot black indicates SNPs without any evidence for selection, blue is for SNPs with high $|iHS|$ and red is for SNPs with significant evidence for selection as ascertained by clustering of high $|iHS|$ SNPs. **(A)** Plots the expected distribution of the p-values from the linear regression model for *cis* eQTLs against the observed p-values. The blue points show the p-values for SNPs with high $|iHS|$. The red points plot the distribution for SNPs with significant evidence for selection based on clustering of SNPs with high $|iHS|$. **(B)** SNPs with high iHS show an enrichment for eQTLs after at various distances from the transcription start site. **(C)** SNPs with high iHS tend to be enriched for eQTLs after controlling for allele frequency. **(D)** SNPs with high iHS show greater levels of eQTLs after controlling for LD levels, as measured by the local recombination rate.

| CEU | YRI | ASN |
|---|---|---|
| 3107(HLA-C) | 10653(SPINT2) | 60558(GUF1) |
| 55112(WDR60) | 79443(FYCO1) | 65055(REEP1) |
| 113277(TMEM106A) | 5468(PPARG) | 79080(CCDC86) |
| 91612(CHURC1) | 5066(PAM) | 79623(GALNT14) |
| 8904(CPNE1) | 22934(RPIA) | 80008(TMEM156) |
| 134728(IRAK1BP1) | 51318(MRPL35) | 720(C4A) |
| 126248(PQWD) | 132299(OCIAD2) | 5155(PDGFB) |
| 9135(RABEP1) | 8904(CPNE1) | 8706(B3GALNT1) |
| 167227(DCP2) | 7391(USF1) | 64105(CENPK) |
| 23350(SR140) | 25888(ZNF473) | 3117(HLA-DQA1) |
| 91646(ECAT8) | 94103(ORMDL3) | 3107(HLA-C) |
| 10905(MAN1A2) | 132321(LOC132321) | 55081(IFT57) |
| 389362(LOC389362) | 84545(MRPL43) | 129684(CNTNAP5) |
| 10352(WARS2) | 79913(ACTR5) | 135295(SRrp35) |
| 347733(TUBB2B) | 54107(POLE3) | 4259(MGST3) |
| 7280(TUBB2A) | 8034(SLC25A16) | 121053(C12orf45) |
| | 4605(MYBL2) | 8904(CPNE1) |
| | 51231(VRK3) | 10475(TRIM38) |
| | 113675(SDSL) | 64976(MRPL40) |
| | 221264(C6orf199) | 3127(HLA-DRB5) |
| | 25961(NUDT13) | 177(AGER) |
| | 1763(DNA2L) | 440279(UNC13C) |
| | 10781(ZNF266) | 132946(ARL9) |
| | 84282(RNF135) | 64847(SPATA20) |
| | 10785(WDR4) | 11148(HHLA2) |
| | 160140(C11orf65) | 79174(CRELD2) |
| | 284443(ZNF493) | 91419(XRCC6BP1) |
| | 79631(EFTUD1) | 90416(CCDC32) |
| | 114801(KIAA1913) | 9648(GCC2) |
| | 11112(HIBADH) | 6166(RPL36AL) |
| | 5074(PAWR) | 6875(TAF4B) |
| | 4763(NF1) | 7587(ZNF37A) |
| | 4123(MAN2C1) | 135293(ACY1L2) |
| | 23729(CARKL) | 9145(SYNGR1) |
| | 51329(ARL6IP4) | 128272(ARHGEF19) |
| | | 326625(MMAB) |
| | | 55840(EAF2) |
| | | 285961(LOC285961) |
| | | 219972(MPEG1) |
| | | 8732(RNGTT) |
| | | 4276(MICA) |
| | | 10957(PNRC1) |
| | | 3712(IVD) |
| | | 50854(C6orf48) |
| | | 3116(HLA-DPB2) |
| | | 283635(C14orf24) |
| | | 114789(SLC25A25) |

Table 10: **List of genes with eQTLs centered in signals for selection.** Entrez Gene IDs are provided for the genes for each population where there is at least 1 SNP with 100Kb of the gene that is significantly associated ($p<0.0001$) with gene expression, has an $|iHS| > 2$ and is present in a cluster of other SNPs with high $|iHS|$. Gene symbols are provided next to each gene ID in parantheses.

# References


Clark, A., S. Glanowski, R. Nielsen, P. Thomas, A. Kejariwal, M. Todd, D. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. White, J. Sninsky, M. Adams, and M. Cargill (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science 302*, 1960–1963.

Haygood, R., O. Fedrigo, B. Hanson, K. Yokoyama, and G. Wray (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet. 39*, 1140–1144.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature 449*, 851–861.

Karolchik, D., R. Baertsch, M. Diekhans, T. Furey, A. Hinrichs, Y. Lu, K. Roskin, M. Schwartz, C. Sugnet, D. Thomas, R. Weber, D. Haussler, and W. Kent (2003). The UCSC Genome Browser Database. *Nucleic Acids Res. 31*, 51–54.

Thomas, P., M. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res. 13*, 2129–2141.

Thomas, P., A. Kejariwal, N. Guo, H. Mi, M. Campbell, A. Muruganujan, and B. Lazareva-Ulitsky (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res. 34*, W645–650.

Voight, B., S. Kudaravalli, X. Wen, and J. Pritchard (2006). A map of recent positive selection in the human genome. *PLoS Biol. 4*, e72.