Supplementary Methods

RNA preparation

Total RNA was extracted from lymphoblastoid cell lines of the 270 individuals of the HapMap ([1]; Coriell, Camden, New Jersey, United States). Two, one-quarter scale Message Amp II reactions (Ambion, Austin, Texas, United States) were performed for each RNA extraction using 200 ng of total RNA as previously described [2]. 1.5 µg of the cRNA was hybridized to an array [3].

Gene expression quantification

To assay transcript levels in the cell lines, we used Illumina's commercial whole genome expression array, Sentrix Human-6 Expression BeadChip version 1 (Illumina, San Diego, California, United States) [4]. These arrays utilize a bead pool with ~48,000 unique bead types (one for each of 47,294 transcripts, plus controls), each with several hundred thousand gene-specific 50mer probes attached.

On a single BeadChip, six arrays were run in parallel as described [3]. Each bead type (probe) is present on a single array on average 30 times. Each of the two IVT reactions from the 270 samples was hybridized to two arrays each, so that each cell line had four replicate hybridizations. cRNA was hybridized to arrays, and subsequently labelled with Cy3-streptavidin (Amersham Biosciences, Little Chalfont, United Kingdom) and scanned with a Bead Station (Illumina) as previously described in Stranger *et al.* [2].

Post-experimental raw data normalization

With the Illumina bead technology, a single hybridization of RNA from one cell line to an array produces on average approximately 30 intensity values for each of 47,294 bead types. These background-corrected values for a single bead type are subsequently summarized by Illumina software and output to the user as a set of 47,294 intensity values for each individual hybridization [5]. In our experiment, each cell line was hybridized to 4 arrays, thus resulting in 4 reported intensity values (as averages of the values from the 30 beads per probe) for each of the 47,294 bead types. To combine data from our multiple replicate hybridizations, raw data were read using the beadarray R package [6] and then normalized on a log scale using a quantile normalization method [7] across replicates of a single individual, followed by a median normalization method across all 270 individuals. These normalized values are used in subsequent analyses.

Heritability estimation

To estimate heritability of expression phenotypes, we used the data for the trios of the CEU and YRI populations. We performed midparent-offspring regression for each of the probes on the array. The slope of the line is the estimate of heritability for that probe.

Population differentiation analysis

To avoid potential age and batch effects at the establishment of the cell lines, that would impact the estimation of population differentiation estimates occurring we used the trios to establish critical values of differentiation. We expect that differences between CEU or YRI parents and CEU or YRI children represent the null distribution of differences in median expression values since they are related. Therefore, we can use them to establish critical thresholds of differentiate in gene expression levels that correspond to the null distribution. For both the CEU and YRI populations we estimated the degree of median deviation in expression levels between parents and children such that 5% of expression phenotypes exceed the threshold. This results in a difference in median $\log_2$ values of 0.22, which corresponds to a 16% difference in median expression levels.

Association analyses

The association analysis employed: 1) Normalized $\log_2$ quantitative gene expression measurements for the 210 unrelated individuals of each HapMap population using the Illumina Sentrix Human-6 Expression BeadChip, 2) SNP genotypes for the unrelated individuals of each HapMap population with minor allele frequency above 5% from the HapMap phase II map for each population (version 21, NCBI Build 35).

Of the 47,294 probes for which we collected expression data, we initially selected a set of 14,925 probes to analyze, corresponding to autosomal genes that were highly variable within or between populations (as described in Stranger *et al.* [3]). We subsequently discarded from our analyses any probe that mapped to more than one Ensembl gene (Ensembl version 42) or that had an associated SNP underlying the probe sequence. This resulted in a set of 14,456 probes that were analyzed in the association analyses, corresponding to 13,643 unique autosomal genes.

Association and multiple-test correction (individual populations)

For each of the selected probes interrogating expression and for each SNP, we fit a linear regression model as previously described [2,3]. We also performed Spearman Rank Correlation. Both of these analyses were applied to each population separately, including the unrelated individuals only.

In order to assess significance of associations of expression variation to SNP genotype using the linear regression model, we performed 10,000 permutations of each expression phenotype relative to the genotypes. We performed both *cis* and *trans* analyses as follows:

For the *cis*- association, we limited the analysis to those probes and SNPs (MAF > 5%) where the distance from probe genomic midpoint to SNP genomic location was less than or equal to 1Mb. An association to a gene expression phenotype was considered significant if the p-value from the analysis of the observed data (nominal p-value) was lower than the threshold of the 0.001 tail of the distribution of the minimal p-values (among all comparisons for a given gene) from 10,000 permutations of the expression phenotypes [8,9].

For the *trans*- association, we selected a subset of phase II HapMap SNPs that have a higher probability of being functional than randomly selected SNPs of the genome. We selected SNPs of four categories: i) All SNPs with significant *cis*- associations, ii) All nsSNPs (rs numbers from Ensembl v41, genotypes extracted from HapMap v21), iii) All splice SNPs (rs numbers from Ensembl v41, genotypes extracted from HapMap v21), and iv) microRNA SNPs (as annotated in miRBase; genotypes from HapMap v21). Together these categories comprised a set of approximately 29,000 SNPs with MAF > 5% in each of the four populations. We limited the analysis to those probes and SNPs where the probe and SNP were on different chromosomes, or where the probe and SNP were on the same chromosome but the distance from probe midpoint to SNP position was greater than 1Mb. An association to a gene expression phenotype was considered significant if the p-value from the analysis of the observed data (nominal p-value) was lower than the threshold of the 0.001 tail of the distribution of the minimal p-values (among all comparisons for a given gene) from 10,000 permutations of the expression phenotypes [8,9]. To assess whether any of the 4 categories of SNPs was over- or underrepresented among the significant *trans*-associations (relative to the SNPs tested), a Fisher Exact Test was employed.

To assess significance of associations between SNP genotype and expression variation as quantified by the Spearman rank correlation, we used 10,000 permutations.

Association and multiple-test correction (multiple population panels)

With the aim of increasing the power of our *cis*- association analysis, data were combined (normalized expression values and SNP genotypes) for unrelated individuals of multiple populations to comprise three different multiple population analysis panels: 1) CEU-CHB-JPT-YRI, 2) CEU-CHB-JPT, and 3) CHB-JPT. The *cis*- association was performed separately for each of these panels using linear regression as described above, only considering those SNPs located less than 1Mb away from the probe midpoint.

Conditional permutations were performed to assess significance of the nominal p-values [10,11]. For each of the 14,456 probes in each multiple population panel, expression values were permuted among individuals of a single population, followed by regression analysis of the grouped multi-population expression data against the grouped multi-population permuted SNP genotypes. Associations were considered significant if the nominal p-value was lower than the threshold of the 0.001 tail of the distribution of the minimal p-values from 10,000 permutations of the expression phenotypes.

Genome annotation in associated regions

From the significant *cis*- associations detected in the single population analyses, a non-redundant list of the most significant SNP per gene was mapped to the genomic coordinates of 3 categories of genome annotation: i) exons (Ensembl v41), ii) promoters (100kb regions upstream of 5'UTRs; Ensembl v41), iii) Conserved non-genic sequences (Conserved non-coding regions that do not overlap Ensembl genes (Ensembl v41). In addition, the full set of non-redundant SNPs that was tested in the *cis*- analysis was mapped to the annotation. An enrichment or deficit of associated SNPs (relative to tested SNPs) in a given annotation category was assessed with the Fisher's Exact Test.

Gene ontology annotation in associated regions

We tested for significant enrichment or deficit of genes of specific gene ontology annotation among the set of genes exhibiting *cis*- associations relative to the set of 13,643 genes tested. Gene ontology categories were obtained from the European Bioinformatics Institute's Gene Ontology Association (GOA) database (www.ebi.ac.uk/GOA). All tested genes were mapped to GO-slim categories, which are meta-categories of gene ontology categories. Enrichment or deficit of each GO-slim category among the significant *cis*- associated genes was evaluated through the use of Fisher's Exact Test.

Evolutionary analyses of associated regions

Multiple species alignments

We investigated the depth of conservation for all SNPs exhibiting a significant *cis*-association at the 0.01 permutation threshold (from the single population analysis). We queried Galaxy for alignments of all the SNPs tested in this study, to sequence from 7 additional species (chimpanzee, mouse, rat, dog, chicken, fugu, frog). We ascertained the depth of the conservation of the human polymorphic sites by counting the instances where an alignment was obtained for each of the following sets: 1) All species (human + all 8), 2) All mammals and chicken, 3) All mammals and fish, 4) At least three mammals, 5) At least three mammals and chicken, and 6) At least three mammals and fish.

SNPs that satisfied the above criteria for alignment were explored with regards to their allelic state. We counted the instances where all bases at a position of homology were identical to the chimpanzee sequence.

A Fisher's Exact Test was used to investigate whether *cis*- associated SNPs at the 0.01 threshold show higher levels of conservation with respect to all the SNPs tested in this study.

Accession Numbers

The expression data reported in this paper have been previously deposited in the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) database (Series Accession Number GSE6536 [3]).

References

1. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
2. Stranger, B.E. et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).
3. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
4. Kuhn, K. et al. A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res* **14**, 2347-56 (2004).

5. Dunning, M.J., Thorne, N.P., Camilier, I., Smith, M.L. & Tavaré, S. Quality control and low-level statistical analysis of Illumina BeadArrays. *Revstat* **4**, 1-30 (2006).
6. Dunning, M.J., Smith, D.R., Thorne, N.P. & Tavaré, S. beadarray: An R Package to analyse llumina BeadArrays. *R News* **6**, 17 (2006).
7. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
8. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-71 (1994).
9. Doerge, R.W. & Churchill, G.A. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285-94 (1996).
10. Koren, M. et al. ATM haplotypes and breast cancer risk in Jewish high-risk women. *Br J Cancer* **94**, 1537-43 (2006).
11. Lee, S.I., Pe'er, D., Dudley, A.M., Church, G.M. & Koller, D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A* **103**, 14062-7 (2006).