nature genetics

Population genomics of human gene expression

Barbara E Stranger[1], Alexandra C Nica[1], Matthew S Forrest[1], Antigone Dimas[1], Christine P Bird[1], Claude Beazley[1], Catherine E Ingle[1], Mark Dunning[2], Paul Flicek[3], Daphne Koller[4], Stephen Montgomery[1], Simon Tavaré[2], Panos Deloukas[1] & Emmanouil T Dermitzakis[1]

Genetic variation influences gene expression, and this variation in gene expression can be efficiently mapped to specific genomic regions and variants. Here we have used gene expression profiling of Epstein-Barr virus–transformed lymphoblastoid cell lines of all 270 individuals genotyped in the HapMap Consortium to elucidate the detailed features of genetic variation underlying gene expression variation. We find that gene expression is heritable and that differentiation between populations is in agreement with earlier small-scale studies. A detailed association analysis of over 2.2 million common SNPs per population (5% frequency in HapMap) with gene expression identified at least 1,348 genes with association signals in *cis* and at least 180 in *trans*. Replication in at least one independent population was achieved for 37% of *cis* signals and 15% of *trans* signals, respectively. Our results strongly support an abundance of *cis*-regulatory variation in the human genome. Detection of *trans* effects is limited but suggests that regulatory variation may be the key primary effect contributing to phenotypic variation in humans. We also explore several methodologies that improve the current state of analysis of gene expression variation.

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

Understanding the molecular basis of human phenotypic variation is a key goal of human genetics, encompassing disease susceptibility, variable response to drugs and ultimately treatment and public health. Over the past decades, studies have described and analyzed the genetic basis of human phenotypic variation ranging from whole-organism phenotypes such as height[1], to molecular-level phenotypes such as lipid abundance[2,3]. Previous studies have also investigated the effects of nucleotide variation in specific genes or genomic regions on complex and monogenic diseases. There has also been an explosion of genome-wide studies examining the genetic basis of complex diseases by exploring the effects of genetic variation such as SNPs[4–7] and copy number variants[8–10], some of which are clearly in noncoding regions of the genome[4–7,11]. Technological advances have now made genome-wide association studies a reasonable and affordable approach to the study of complex phenotypes[12].

Although association methodologies can identify genomic regions containing the genetic variants underlying disease and other phenotypes, on their own they provide little insight into which is the functional variant and/or mechanism. There is a need for methodologies that allow both the interpretation of functional consequences of variants and the description of functionally important variants[13,14]. Gene expression (that is, transcription) level is a quantitative phenotype that is directly linked to DNA variation and can be affected by polymorphisms in *cis*-regulatory regions[15–19] or by exonic variants altering transcript stability or splicing[20]. In addition, gene expression (or mRNA levels) can be measured accurately and consistently in tissues and cell lines in humans. Many studies have described the genetic basis of transcriptional variation and have convincingly demonstrated that it is a heritable trait[16,17,21]. The highly dense, phase II HapMap[22,23] now facilitates, for the first time, a fine-scale analysis of the genomic location and properties of the variants associated with gene expression. In addition, we can advance current knowledge by investigating both the genetic basis of gene expression variation within and between populations and its implications and relationship to genome function.

Here we have used transcriptional profiling of the 270 individuals in the four HapMap populations and their genotypes of nearly 4 million SNPs described by the HapMap Consortium[22,23] to elucidate some of the key features of the genetics of gene expression. We provide biological insights into variable gene expression and the fine-scale genomic properties of *cis*- and *trans*-acting regulatory variants. We also present methodological implementations that uncover shared functional genetic effects between populations, and describe the degree and characteristics of population differentiation with respect to genetic effects. Our analysis describes extensive replication of signals in multiple populations and provides a comprehensive exploration of biological signals underlying regulatory associations.

RESULTS

Data generation and biological properties of the data

We quantified gene expression in the 270 individuals genotyped in the International HapMap Project with Illumina's human whole-genome

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [3]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [4]Computer Science Department, Gates Building 1A, Stanford University, Stanford, California 94305-9010, USA. Correspondence should be addressed to E.T.D. (md4@sanger.ac.uk) or P.D. (panos@sanger.ac.uk).

Received 30 May; accepted 29 August; published online 16 September 2007; doi:10.1038/ng2142

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

expression (WG-6 version 1) arrays, which contain 47,294 probes in four technical replicates[19]. The population samples include 30 Caucasian trios of northern and western European origin (CEU), 45 unrelated Chinese individuals from Beijing University (CHB), 45 unrelated Japanese individuals from Tokyo (JPT), and 30 Yoruba trios from Ibadan, Nigeria (YRI). Expression signal values were $\log_2$-transformed and normalized first by quantile normalization across the four replicates of an individual, followed by median normalization across all 270 individuals to allow comparison of expression values across populations[24,25].

We estimated the median and variance of each of the 47,294 probe types for each population, and analyzed the distribution of variance and median values of normalized values by Gene Ontology (GO) categories[26] after summarizing them in GO Slim categories[27]. Specific GO Slim categories such as 'chaperone regulatory activity' showed an excess of high variance of gene expression, whereas genes with extracellular function showed low variation. 'Chaperone regulatory activity' genes and 'translational regulatory activity' genes had highest median expression across all populations. The latter category was mainly driven by the very high expression of ribosomal proteins.

We estimated heritability of expression phenotypes independently in the CEU and YRI trios by performing midparent-offspring regressions. Of the 47,294 probes analyzed, 4,829 and 6,482 (10% and 13%, respectively) demonstrated heritability greater than 0.2 in CEU and YRI, respectively, with an overlap of 958 genes; and 154 CEU genes and 217 YRI genes had heritability higher than 0.5 with an overlap of only nine genes. This observation suggests that even with only 30 trios per population we can detect a substantial number of heritable expression phenotypes, although the low heritability estimates indicate that there are considerable additional nongenetic sources of variation in gene expression.

We tested for population differences in gene expression because some studies have shown that it is considerable[16,28]. To avoid potential age and batch effects occurring at establishment of the cell lines, we used the trios to establish critical values of differentiation. This approach (**Supplementary Methods** online) results in a difference in median $\log_2$ values of 0.2 (16% difference in median expression). Using this threshold, we tested all pairs of populations, after pooling CHB and JPT into one population (ASN), and calculated the number of genes exceeding this threshold. In total, 5,359 genes exceeded the threshold in one or more of the three population pairs (**Supplementary Fig. 1** online), and most of them were different in only one population pair, as expected. If the number of genes expressed in lymphoblastoid cell lines is about 50% of the total (about 12,000 genes), then we estimate that the fraction of genes with significant gene expression differences between any two populations is between 17% and 29%. Caution should be taken with such a result, however, because we observed that the CEU population was the most divergent, although we expected it to cluster closer to ASN. This discrepancy is probably due to the much older age of the CEU cell lines relative to the most recently established YRI and ASN cell lines.

For subsequent analyses, we defined a reduced set of 14,456 probes (13,643 distinct genes) selected on criteria of variance and population differentiation (**Supplementary Methods**). This set is smaller than our previous set of 14,925 probes (14,072 genes; ref. 19), because we removed probes with either multiple mapping positions (370) or SNPs within them that generated false associations (99). All association analyses were performed with the 14,456 probe set, corresponding to 13,643 genes.

Expression levels as measured by two different arrays

To maximize the power to detect genetic effects, it is important to establish that expression measurements are robust to experimental variables. We previously described expression data for 60 CEU cell lines (all of the CEU HapMap parents) generated with Illumina's low-density (~700 genes) custom arrays[18]. A total of 539 probes on this custom array have sequences identical to probes on the genome-wide array, which allows direct comparison of expression signals across experiments.

For probes where there was variable signal intensity across individuals, there was highly significant correlation between signals from the two experiments (**Supplementary Fig. 2** online). Note that the RNA used in each experiment for a given individual was extracted from a different cell line batch. The RNA labeling, hybridization and normalization were also done independently. The high degree of correlation illustrates that the transcript measurements are stable despite differential growth and treatment of cell lines and samples.

Cis associations of gene expression with SNPs

We selected HapMap phase II SNPs with a minor allele frequency (MAF) of >5% from each of the four populations (CEU, CHB, JPT, and YRI; ~2.2 million SNPs per population). We tested for association between SNP variation and expression variation for each of 13,643 genes independently in each population, considering only unrelated individuals. For each population, we used a linear regression

Table 1 ***Cis*** **associations detected with different statistical models**

	0.001 permutation threshold				
Population	LR: significant genes	SRC: significant genes	Overlap, LR and SRC	% LR	% SRC
CEU	299	293	242	0.81	0.83
CHB	318	274	238	0.75	0.87
JPT	341	326	264	0.77	0.81
YRI	394	363	302	0.77	0.83
Nonredundant	831	783	642		
4 populations	62	57	46		
≥2 populations	310	283	239		
	0.01 permutation threshold				
	LR: significant genes	SRC: significant genes	Overlap, LR and SRC	% LR	% SRC
CEU	606	591	450	0.74	0.76
CHB	634	598	460	0.73	0.77
JPT	679	667	521	0.77	0.78
YRI	742	730	564	0.76	0.77
Nonredundant	1,746	1,737	1,282		
4 populations	114	99	87		
≥2 populations	533	513	423		

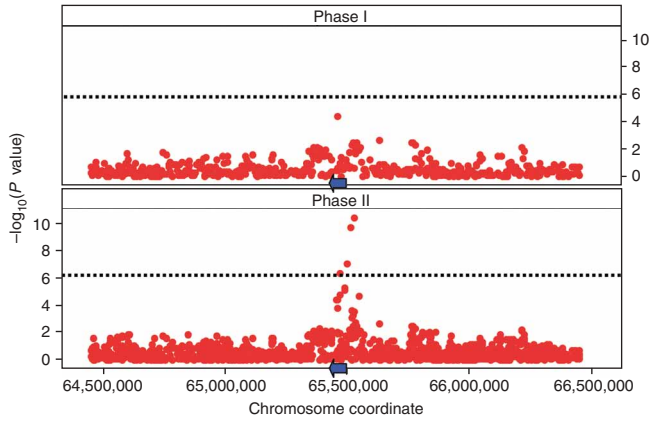


Figure 1 Associations of SNPs with expression of the gene *SPRED2* on chromosome 2. Top and bottom panels contrast the results obtained with phase I HapMap SNPs and phase II HapMap SNPs. Coordinates are in NCBI Build 35. Blue arrows represent the location (not to scale) and direction of transcription of the associated gene.

(LR) model. To analyze those SNPs that potentially act in *cis*, we tested only those SNPs located within a region 1-Mb upstream or downstream of the expression probe midpoint (the 'candidate region approach'[18]). For large genes (>500 kb), we also used the transcription start site (TSS) as the center of the 2-Mb window, discovering only one additional association relative to those discussed below.

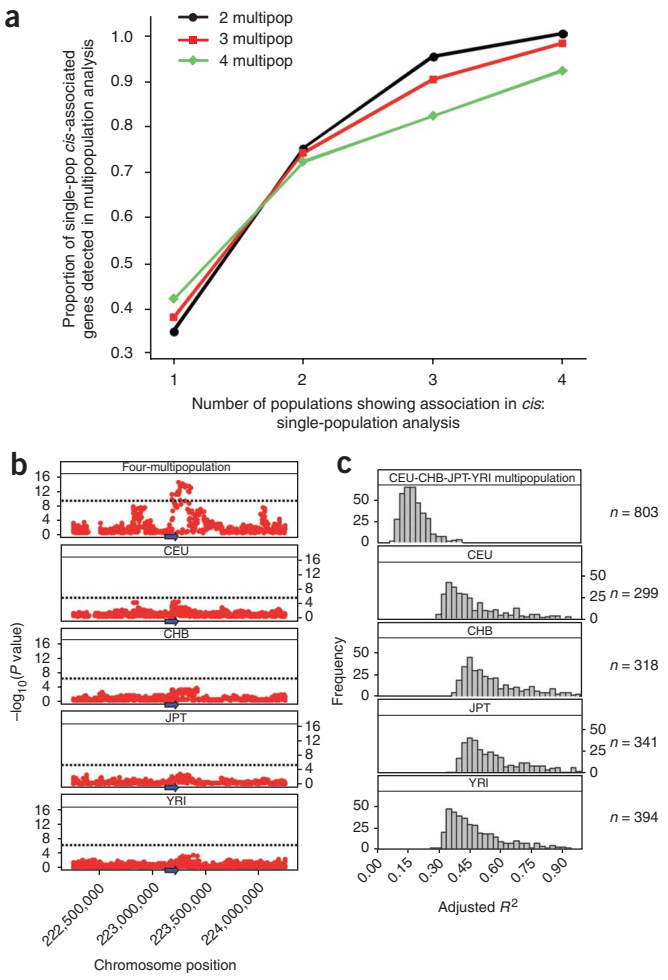
To determine the significance of the regression *P* values, we performed 10,000 permutations of the data independently for each gene (and population) and analyzed in depth those associations significant at the 0.001 permutation threshold (**Supplementary Methods**). At this level of significance, we expect roughly 14 genes to have at least one significant association by chance, and we detected 299, 318, 341 and 394 for CEU, CHB, JPT and YRI, respectively, with a false discovery rate (FDR) of 4–5% per population (**Table 1** and **Supplementary Table 1** online). In total, there is a nonredundant set of 831 genes showing a significant *cis* association in at least one population, 310 genes in at least two populations, and 62 in all four. As expected, owing to the small sample size, the detected genetic effects are large and the square of the correlation coefficient (R^2) ranges from 0.27 to almost 1. A total of 209 out of 299 CEU-significant genes and 247 out of 394 YRI-significant genes have heritability estimates above 0.2. Of the set of 831 genes with significant *cis* associations, 431 have heritability estimates above 0.2.

Figure 2 Comparison of *cis* associations detected between single- and multipopulation analysis. (**a**) Numbers of genes with significant *cis* associations determined by single- and multipopulation analysis, and proportion of overlap of associations across the two methodologies. (**b**) Associations of SNPs of the phase II HapMap with expression of the gene *SGPP2* on chromosome 2. Coordinates are in NCBI Build 35. Shown are four-population multipopulation analysis and individual population analysis for CEU, CHB, JPT and YRI. Blue arrows represent the location (not to scale) and direction of transcription of the associated gene. In this case, the SNP was not rare in any of the populations ($0.08 < \text{MAF} < 0.44$), but the effect was small ($R^2 = 0.25$, slope = 0.25); therefore, it was detected only when the populations were pooled and the sample size increased. (**c**) Comparison of the adjusted R^2 values (proportion of the variance in expression explained by the linear relationship between genotype and phenotype) of significant *cis* associations obtained from single- and multipopulation analysis (0.001 permutation threshold).

Heritability overall correlates reasonably well with *cis*-association significance (**Supplementary Fig. 3** online), but the heritability estimates have large variance and are to be taken with caution on a per-gene basis. When evaluated in the context of GO Slim terms, we detected a considerable deficit of genes involved in 'cellular processes' among the set of genes with significant *cis* associations. This observation is not surprising because segregating genetic variation that affects such basic cell functions would be expected to be detrimental.

When we compare the sets of genes showing *cis* associations identified with the phase I HapMap[19] to those identified with the phase II HapMap data, we observe that the phase I HapMap captured 79–87% (depending on the population) of the genes detected with phase II. Only in YRI is there is a substantial gain in numbers of *cis*-associated genes (**Fig. 1** and **Supplementary Fig. 4** online); in the other three populations, linkage disequilibrium decays much more slowly such that, instead of capturing most common haplotype diversity, in the YRI population the phase II HapMap captures additional functional genetic variation relative to phase I HapMap.

It would be desirable to be able to use the full sample of all 210 unrelated individuals to detect genetic effects on gene expression that are common but of smaller magnitude than those detected within each individual population. We cannot simply pool the samples without appropriate corrections, however, because population differentiation will generate spurious associations. Conditional permutations allow us to reveal the relevant associations while masking inflated associations[29,30]. We repeated the association analysis after



© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

Table 2 *Cis* **associations in single populations and multipopulation subsets**

	0.001 permutation threshold						
Population	1 LR: significant genes	2 CEU-CHB-JPT-YRI multipop	3 CEU-CHB-JPT multipop	4 CHB-JPT multipop	Overlap, 1 and 2	Overlap, 1 and 3	Overlap, 1 and 4
CEU	299	803	735	651	204	215	180
CHB	318	803	735	651	230	255	276
JPT	341	803	735	651	244	260	287
YRI	394	803	735	651	213	183	174
Nonredundant	831						
4 populations	62						
≥2 populations	310						

	0.01 permutation threshold						
Population	1 LR: significant genes	2 CEU-CHB-JPT-YRI multipop	3 CEU-CHB-JPT multipop	4 CHB-JPT multipop	Overlap, 1 and 2	Overlap, 1 and 3	Overlap, 1 and 4
CEU	606	1186	1149	1071	356	373	312
CHB	634	1186	1149	1071	373	406	430
JPT	679	1186	1149	1071	417	436	473
YRI	742	1186	1149	1071	365	339	312
Nonredundant	1746						
4 populations	114						
≥2 populations	533						

pooling unrelated individuals of all four populations, a subset of three (CEU-CHB-JPT) populations, and two (CHB-JPT) populations. The rationale for the choice of population combinations was to pool those sets of populations that are more closely related. To correct for inflation of the *P* values, we performed conditional permutations, such that expression values from an individual of a given population were assigned only to another individual of the same population. This process corrects for the *P* value inflation because *P* values from permuted data sets are also inflated. We detected 803, 735 and 651 genes as significant for the four-population, three-population and two-population pools, respectively, corresponding to 1,083 distinct genes and an FDR of 1–2%. The overlap between the multipopulation and the single-population analysis (**Fig. 2a**) demonstrates that this methodology captures most of the population-shared associations that were detected in the single-population analysis, as expected, but it also captures several additional *cis* effects (see example in **Fig. 2b**). Most of the effects detected in the multipopulation analysis are smaller (R^2 = 0.08–0.41) than those detected in single populations (**Fig. 2c**), which is a direct function of the increase in the sample size analyzed.

Most previous studies have used a linear regression model to associate SNP genotypes with gene expression. We used an alternative nonparametric method to evaluate the sensitivity of our results to the statistical methodology used. We used Spearman rank correlation (SRC) to perform the same *cis* analysis described above. We detected 293, 274, 326 and 363 *cis* associations for CEU, CHB, JPT and YRI, respectively, corresponding to 783 distinct genes and an FDR of 4–5%. Of these genes, 283 were detected in at least two populations and 57 in all four. The overlap of SRC with LR was between 77% and 86% of the genes, depending on the population (**Table 1**). We conclude that SRC performs at a level equivalent to LR.

Allelic effects between populations

We have reported that many genes showing *cis* associations at the 0.001 permutation threshold are shared (about 37%) in at least two populations (**Table 2** and ref. 19). This comparison refers to the across-population association of the same gene and, in most cases, the same SNPs. The gold standard for association replication requires that the same SNP is associated with the same phenotype, and that the allelic effects are in the same direction across multiple independent populations. We compared the allelic directions of SNP-gene associations shared in all pairs of populations. In 95–97% of the shared associations, the direction of the allelic effect was the same across populations (**Fig. 3**), and the discordant 3–5% was of the same order as the FDR. This finding further corroborates the view that the associations that we observe represent real genetic effects on gene expression.

We also investigated whether the extent to which those associations that are not shared between populations could be attributed to differential allele frequencies across populations, as has been reported[16]. For each pair of populations, we split the associated SNPs (0.001 permutation threshold) into three categories: SNPs significant for the same gene in both populations (SNP-shared associations); gene associations in both populations but with different SNPs (gene-shared associated SNPs); and population-specific associations (unshared associations). For these three categories of SNPs, we computed the difference in expected heterozygosity (2pq) in the same direction (for example, $\mathrm{Het}_{\mathrm{population1}} - \mathrm{Het}_{\mathrm{population2}}$) and compared the distributions of the differences among the three categories. As expected, median difference in heterozygosity was the lowest for SNP-shared associations, and gene-shared associated SNPs

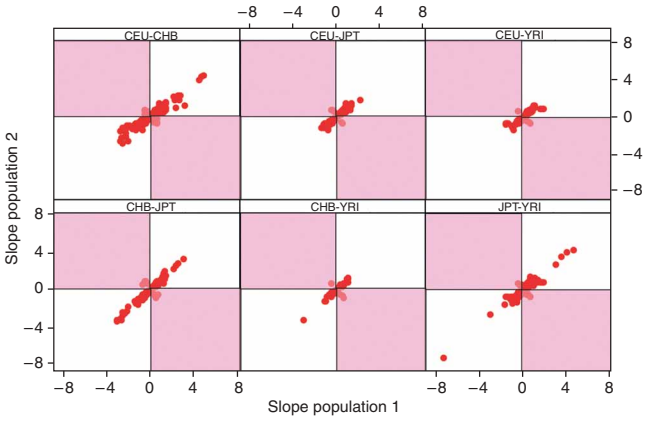


Figure 3 Comparison of the direction of shared SNP-gene allelic effects across all pairs of populations. The permutation threshold was 0.001. White squares indicate effects in the same direction.

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

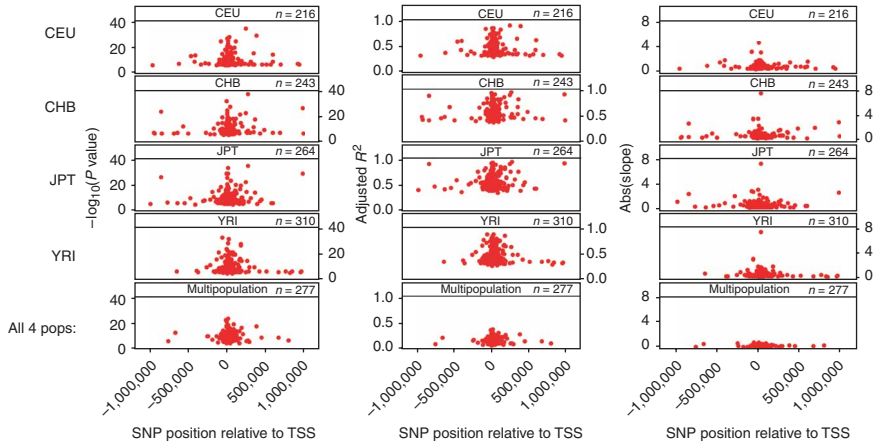


Figure 4 Properties of significant *cis* associations as a function of SNP distance from the transcription start site. Statistical significance, adjusted R^2 of the association (proportion of the expression variance explained by the linear relationship between genotype and phenotype), and absolute value of the slope of the linear regression, as a function of distance from the TSS, of the most significantly associated SNP per gene in each of the four populations (order CEU, CHB, JPT and YRI) and the pooled sample of all four populations.

showed the second lowest difference (**Supplementary Fig. 5**). Our results are consistent with previous observations[16]. One small caveat is that, because of small sample sizes, there could be slight fluctuations in allele frequencies simply due to sampling variance that might affect the detection of associations above a certain threshold.

Associations with respect to genome annotation and conservation

The high SNP density of the HapMap makes it possible that some of the SNPs interrogated are the causal variants (it is estimated that 30–50% of SNPs with MAF > 5% are represented in the HapMap[22,23]), which means that evaluation of the genomic annotation where associated SNPs are found may be informative. For each of the four populations, we mapped the most significant SNP for each of the genes with significant *cis* associations from the single-population analysis relative to the TSS of genes with annotated 5′ UTRs. We found that a strong signal for the SNPs was located very close to the TSS (**Fig. 4**), with no discernable trend in the 3′ or 5′ direction. This symmetrical trend is also evident in the analysis of the ENCODE Consortium[31]. When we considered the most significant SNP of the 341 additional genes detected in the four-population multipopulation analysis, the signal was even tighter around the TSS. Three of the associated SNPs (rs10998076, rs869736 and rs1010167) have been shown in promoter transfection assays to have a direct effect on transcriptional activation[32] in kidney and brain cell lines.

We mapped the location of the most significant SNP of the 831 *cis*-associated genes from the single-population analysis with respect to gene promoters, coding sequences and conserved noncoding sequences (CNCs). We observed a significant excess of associated SNPs (relative to those tested) in promoters and coding sequences (Fisher's exact test, $P = 8.94 \times 10^{-24}$ and $P = 4.52 \times 10^{-12}$, respectively), and under-representation in CNCs ($P = 0.00358$). The first two signals are expected and partly confounded because most of the causal variants are found in the genic regions; thus, SNPs in linkage disequilibrium with causal variants may also map to coding sequences. We also observed enrichment of associated SNPs in regions that align in several mammals and as far as fish or chicken (Fisher's exact test, $P = 10^{-5}$). The apparent contradiction between enrichment in conserved nucleotides and deficit in CNCs probably arises because most of the signal for conserved nucleotide comes from exonic sequences that are close to the TSS where the associations are mainly found. The deficit of associated SNPs in CNCs is unexpected because previous reports have suggested that these SNPs are selectively constrained[31,33,34] and in some studies they have been shown to have a role in gene regulation[35–37]. It is possible that the skew toward large effect sizes also skews the distribution of causal regulatory variants.

Trans associations of gene expression with SNPs

The availability of whole-genome expression and SNP data allows the elucidation of genetic effects acting in *trans*. The number of 2.2 million SNPs per population is large (MAF > 0.05); therefore, testing all SNPs against all genes becomes computationally and statistically challenging (correcting for millions of tests). We took a candidate variant approach by testing only putatively functional SNPs. The goal of the analysis is not to compare the numbers of *cis* and *trans* effects, which is an irrelevant issue in the genome-wide context, especially given the differential power in detection; instead, the goal is to assess the relative contribution of primary molecular variants in *trans*. We therefore selected four categories of SNPs to analyze for *trans* effects: SNPs with functional effects on gene expression in *cis* (as determined above in the single-population analyses); nonsynonymous SNPs (Ensembl v41 annotation); SNPs influencing splicing (Ensembl v41 annotation); and SNPs within microRNAs (as annotated in miRBase). These correspond to ~25,000 SNPs (MAF > 0.05) per population (see **Table 3** for counts in each category).

For each population, we used a linear regression model as described above to test for association between SNP variation and expression variation. We confined the *trans* analysis to those SNP-gene combinations where the genomic distance between probe midpoint and SNP was greater than 1 Mb (or where probe and SNP were on different chromosomes). Significance was evaluated through 10,000 permutations as described above. We identified 43, 37, 38 and 23 genes in CEU, CHB, JPT and YRI, respectively, with significant *trans* associations (0.001 permutation threshold). In total, 108 genes show a significant *trans* association in at least one population (16 genes or 15% show a significant *trans* association in at least two populations and five in all four populations). We also performed analysis in pooled populations as described above and detected 44, 52 and 39 genes for the four-, three- and two-population

Table 3 Number and source category of SNPs used in the *trans* analysis

SNP category	CEU	CHB	JPT	YRI
Cis-associated (rSNPs)	13,221	13,133	13,191	13,375
Nonsynonymous	9,904	9,383	9,378	10,727
Splicing	1,756	1,585	1,594	1,950
miRNA	34	34	32	37
Nonredundant[a]	24,635	23,854	23,907	25,797

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

Table 4 *Trans* **associations in single-population and multipopulation analysis**

	0.001 permutation threshold						
	1	2	3	4			
Population	LR: significant genes	CEU-CHB-JPT-YRI multipop	CEU-CHB-JPT multipop	CHB-JPT multipop	Overlap, 1 and 2	Overlap, 1 and 3	Overlap, 1 and 4
CEU	43	44	52	39	9	12	12
CHB	37	44	52	39	10	14	15
JPT	38	44	52	39	10	14	16
YRI	23	44	52	39	7	7	7
Nonredundant	108						
4 populations	5						
≥2 populations	16						

	0.01 permutation threshold						
	1	2	3	4			
Population	LR: significant genes	CEU-CHB-JPT-YRI multipop	CEU-CHB-JPT multipop	CHB-JPT multipop	Overlap, 1 and 2	Overlap, 1 and 3	Overlap, 1 and 4
CEU	247	171	329	208	14	28	18
CHB	210	171	329	208	11	22	20
JPT	196	171	329	208	15	20	20
YRI	159	171	329	208	15	17	15
Nonredundant	756						
4 populations	9						
≥2 populations	32						

pools, respectively. Overall, there seems to be low power to detect *trans* effects in these cell lines and sample sizes (**Table 4**), as indicated by the small number of genes discovered and consequently the high FDR.

At the 0.001 threshold, most *trans* associations are caused by SNPs from the first category, that is, those with *cis*-regulatory effects, showing three- to sixfold enrichment relative to the total SNPs tested (Fisher's exact test, $P = 10^{-10}$–10^{-24} for CEU, CHB and JPT; not significant for YRI). Some SNPs were significantly associated with expression of multiple genes (up to six genes for a single SNP). The numbers of SNPs that are associated with more than one gene are 29 (CEU), 13 (CHB), 7 (JPT) and 4 (YRI). Eight genes had a *trans* association on the same chromosome (distance >1 Mb) with distances ranging from 1,003,413 bp (potential *cis* effect) to 187,659,746 bp. If gene expression perturbations are similar in underlying genetic effects to whole-organism perturbations and disease, then this last result suggests that most of the common phenotypic variation in humans is driven by variants in regulatory elements, rather than variants in protein-coding sequences, providing some potential answers to the longstanding question of the relative contribution of regulatory and coding causal variants to complex phenotypes.

DISCUSSION

We have performed a comprehensive analysis of genetic effects on gene expression variation in human lymphoblastoid cell lines, presenting evidence for *cis*-regulatory effects of 1,348 genes and their biological properties by adopting a candidate region approach. The limited power of our analysis means that we detect only a subset of the existing functional regulatory effects in these populations. In addition, because we have interrogated only a single cell type, variation manifested only in other cell types is not represented here. These two facts argue for an abundance of *cis*-regulatory variants segregating in human populations, some of which may be responsible for higher-order phenotypic variation and susceptibility to disease.

Our analysis goes beyond the detection of *cis*-regulatory effects. To our knowledge, we have performed the most comprehensive analysis so far of the properties of the *cis*-association signals, and we have systematically described characteristics of the expression data. Together, these analyses provide us with confidence in the signals detected. In addition, we have demonstrated that the association signals detected replicate very well across populations, even though the populations are divergent and the sample sizes are small. We have also detected the effect of population differentiation on gene expression: we have confirmed what has been documented in smaller scale studies[16,28]; that is, among-population allele frequency differences exist and provide a framework for the study of phenotypic differences among populations.

We have provided methodological insights into the analysis of gene expression variation. By using pooling of divergent populations and conditional permutation schemes, we have increased the sensitivity of our analysis, detecting smaller regulatory effects that are shared across populations. We can imagine a more sophisticated conditional permutation scheme that would permit pooling of any set of populations for which the population identities or relatedness metrics are known. We have also used a nonparametric test, namely SRC, and demonstrated that it has enough power to be used in such studies. In addition, SRC has some advantages over linear regression because, contrary to linear regression in which outliers can have a large impact on the *P* values, SRC is not sensitive to outliers and therefore the nominal *P* values can be used directly in methods that estimate the FDR (see **Supplementary Fig. 6** online for an example).

The evolutionary and annotation properties of *cis*-regulatory associations are very relevant because the density of the phase II HapMap facilitates a fine-scale analysis of the association signal. Most of the *cis*-regulatory effects detected map very close to the TSS and are enriched in regions of high sequence conservation. This information provides a useful framework to search for *cis*-regulatory variants in the human genome and suggests that most of the large effect variants are in the genic and immediate intergenic regions. The association data will become available on the Ensembl website in October 2007 as 'Distributed Annotation System' tracks to allow browsing and downloading.

We have attempted to analyze effects in *trans* by adopting a candidate variants approach, which assigns prior relevance to those SNPs known to be associated with *cis* regulation, protein sequence variation (amino acid or splicing variation) or miRNA structure; this approach made correction by permutation feasible. Fewer genes showed significant *trans* effects than showed *cis* effects. This observation arises because *trans* effects are often more indirect and therefore weaker; thus, our sample size does not provide us with enough power in conjunction with the much larger number of tests for which we

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

have to correct. In general, the detection of *trans* effects has been less successful in humans than in yeast[38,39]. Because the yeast cell comprises the whole organism, study of the biological interactions in a yeast cell has the potential to detect all of the interactions. By contrast, the human cell is just a small part of the organism; thus, many of the intercellular effects mediating *trans* effects cannot be discovered. Lastly, we have provided evidence that, among a set of potential variants that could have effects in *trans*, there is a large enrichment in the contribution of *cis*-regulatory variants, which may suggest that *cis*-regulatory variation explains much of the complex phenotypic variation in humans, at least at the molecular level.

We have described a comprehensive analysis of gene expression variation in human populations, and have provided a detailed characterization of both the genetic and the positional effects in the genome. This detailed analysis provides a robust and useful framework for the future analysis of gene expression variation in large cohorts with larger sample sizes but lower SNP densities and potentially multiple cell types. It will also greatly facilitate the interpretation and follow up of disease association studies by allowing the dissection of biological effects in regions that carry strong statistical signals of association. This and future studies will lead to a detailed map of functional variation in the human genome that will complement functional and variation studies toward the complete understanding of phenotypic variation in human populations.

METHODS

RNA preparation. Total RNA was extracted from lymphoblastoid cell lines of the 270 individuals of the HapMap Consortium[22] (Coriell). Two, one-quarter scale Message Amp II reactions (Ambion) were performed for each RNA extraction using 200 ng of total RNA as described[18]. We hybridized 1.5 μg of the cRNA (amplified antisense RNA) to an array[19].

Gene expression quantification. To assay transcript abundance in the cell lines, we used a commercial whole-genome expression array[40] (Sentrix Human-6 Expression BeadChip version 1, Illumina).

Post-experimental normalization of raw data. 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[25]. To combine data from our multiple replicate hybridizations, raw data were read by the beadarray R package[24] and subsequently normalized first on a log scale using a quantile normalization method[41] across replicates of a single individual, and then by a median normalization method across all 270 individuals.

Association analyses. Of the 47,294 probes for which we collected expression data, 14,925 probes were initially selected for analysis as described[19]. 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 process resulted in a set of 14,456 probes, corresponding to 13,643 unique autosomal genes, that were analyzed in the association analyses.

Association and multiple-test correction (individual populations). For each of the probes selected to interrogate expression and for each SNP, we fitted a linear regression model as described[18,19]. We also performed SRC. Both of these analyses were applied to each population separately, considering only the unrelated individuals.

For the *cis* association, we limited the analysis to those probes and SNPs (MAF $>$ 5%) in which the distance from probe genomic midpoint to SNP genomic location was $\leq$1 Mb. For the *trans* association, we selected a subset of phase II HapMap SNPs that have a higher probability of being functional as compared with randomly selected SNPs of the genome. We selected SNPs of four categories: all SNPs with significant *cis* associations; all nonsynonymous SNPs (rs numbers from Ensembl v41, genotypes extracted from HapMap v21); all splice SNPs (rs numbers from Ensembl v41, genotypes extracted from HapMap v21); and microRNA SNPs (as annotated in miRBase; genotypes from HapMap v21). Together, these categories comprised a set of ~29,000 SNPs with MAF $>$ 5% in each of the four populations.

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[42,43].

Association and multiple-test correction (multiple population panels). With the aim of increasing the power of our *cis*-association analysis, data (normalized expression values and SNP genotypes) were combined for unrelated individuals of multiple populations to comprise three different multipopulations: CEU-CHB-JPT-YRI, CEU-CHB-JPT, and CHB-JPT. The *cis* association was performed separately for each of these multipopulations using linear regression as described above, only considering those SNPs located $<$1 Mb away from the probe midpoint. Conditional permutations were performed to assess the significance of the nominal *P* values[29,30].

Significant associations. All significant expression-SNP associations detected in this paper are listed in **Supplementary Tables 2–6** online.

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; ref. 19).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the HapMap Consortium for data availability; M. Smith for assistance with software development; and M. Gibbs, J. Orwick and C. Gerringer for technical support. Funding was provided by the Wellcome Trust (to E.T.D. and P.D.), the US National Institutes of Health ENDGAME (to E.T.D. and S.T.), Cancer Research UK (to S.T.), and the Medical Research Council (to M.D.). S.T. is a Royal Society Wolfson Research Merit Award holder.

AUTHOR CONTRIBUTIONS

B.E.S. performed the majority of the analysis, coordinated the efforts on the project, performed part of the experimental work, and wrote part of the manuscript. E.T.D. and P.D. helped with the analysis, wrote part of the manuscript, and led the project. S.T. and M.D. performed the normalization and helped with statistical analysis. A.C.N., A.D., C.P.B., P.F. and S.M. performed specific parts of the analysis. M.S.F. helped with the analysis and performed part of the experimental work. C.E.I. performed most of the experimental work. C.B. wrote some of the scripts and performed part of the analysis. D.K. provided advice on the permutation analysis.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Ferrari, S.L. *et al.* Polymorphisms in the low-density lipoprotein receptor-related protein 5 (LRP5) gene are associated with variation in vertebral bone mass, vertebral bone size, and stature in whites. *Am. J. Hum. Genet.* **74**, 866–875 (2004).
2. Bansal, A. *et al.* Association testing by DNA pooling: an effective initial screen. *Proc. Natl. Acad. Sci. USA* **99**, 16871–16874 (2002).
3. Mahley, R.W. & Huang, Y. Apolipoprotein E: from atherosclerosis to Alzheimer's disease and beyond. *Curr. Opin. Lipidol.* **10**, 207–217 (1999).
4. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
5. Valentonyte, R. *et al.* Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nat. Genet.* **37**, 357–364 (2005).
6. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
7. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
8. Fanciulli, M. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics

9. Stankiewicz, P. & Lupski, J.R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).
10. Merla, G. *et al.* Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am. J. Hum. Genet.* **79**, 332–341 (2006).
11. Li, M. *et al.* CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat. Genet.* **38**, 1049–1054 (2006).
12. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
13. Stranger, B.E. & Dermitzakis, E.T. The genetics of regulatory variation in the human genome. *Hum. Genomics* **2**, 126–131 (2005).
14. Stranger, B.E. & Dermitzakis, E.T. From DNA to RNA to disease and back: the 'central dogma' of regulatory disease variation. *Hum. Genomics* **2**, 383–390 (2006).
15. Doss, S., Schadt, E.E., Drake, T.A. & Lusis, A.J. *Cis*-acting expression quantitative trait loci in mice. *Genome Res.* **15**, 681–691 (2005).
16. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
17. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
18. Stranger, B.E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
19. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
20. Knight, J.C. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **83**, 97–109 (2005).
21. Monks, S.A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
22. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
23. The International HapMap Consortium. The phase II haplotype map of the human genome. *Nature* (in the press).
24. Dunning, M.J., Smith, D.R., Thorne, N.P. & Tavare, S. beadarray: an R Package to analyse Illumina BeadArrays. *R News* **6**, 17 (2006).
25. Dunning, M.J., Thorne, N.P., Camilier, I., Smith, M.L. & Tavare, S. Quality control and low-level statistical analysis of Illumina BeadArrays. *Rev. Stat.* **4**, 1–30 (2006).
26. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
27. Camon, E., Barrell, D., Lee, V., Dimmer, E. & Apweiler, R. The Gene Ontology Annotation (GOA) Database—an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.* **4**, 5–6 (2004).
28. Storey, J.D., Akey, J.M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* **3**, e267 (2005).
29. 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. USA* **103**, 14062–14067 (2006).
30. Koren, M. *et al.* ATM haplotypes and breast cancer risk in Jewish high-risk women. *Br. J. Cancer* **94**, 1537–1543 (2006).
31. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
32. Hoogendoorn, B. *et al.* Functional analysis of polymorphisms in the promoter regions of genes on 22q11. *Hum. Mutat.* **24**, 35–42 (2004).
33. Dermitzakis, E.T. *et al.* Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033–1035 (2003).
34. Drake, J.A. *et al.* Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38**, 223–227 (2006).
35. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
36. Abbasi, A.A. *et al.* Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. *PLoS ONE* **2**, e366 (2007).
37. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
38. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* **102**, 1572–1577 (2005).
39. Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* **35**, 57–64 (2003).
40. Kuhn, K. *et al.* A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res.* **14**, 2347–2356 (2004).
41. 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**, 185–193 (2003).
42. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
43. Doerge, R.W. & Churchill, G.A. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294 (1996).

© 2007 Nature Publishing Group http://www.nature.com/naturegenetics