Genetic analysis of genome-wide variation in human gene expression

Michael Morley[1,3]*, Cliona M. Molony[2]*, Teresa M. Weber[1,3], James L. Devlin[2], Kathryn G. Ewens[2], Richard S. Spielman[2] & Vivian G. Cheung[1,2,3]

[1]*Department of Pediatrics and* [2]*Department of Genetics, University of Pennsylvania,*
[3]*The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA*

* These authors contributed equally to this work

Natural variation in gene expression is extensive in humans and other organisms, and variation in the baseline expression level of many genes has a heritable component. To localize the genetic determinants of these quantitative traits (expression phenotypes) in humans, we used microarrays to measure gene expression levels and performed genome-wide linkage analysis for expression levels of 3,554 genes in 14 large families. For approximately 1,000 expression phenotypes, there was significant evidence of linkage to specific chromosomal regions. Both *cis*- and *trans*-acting loci regulate variation in the expression levels of genes, although most act *in trans*. Many gene expression phenotypes are influenced by several genetic determinants. Furthermore, we found hotspots of transcriptional regulation where significant evidence of linkage for several expression phenotypes (up to 31) coincides, and expression levels of many genes that share the same regulatory region are significantly correlated. The combination of microarray techniques for phenotyping and linkage analysis for quantitative traits allows the genetic mapping of determinants that contribute to variation in human gene expression.

The expression level of many genes shows abundant natural variation in species from yeast to humans[1–6]. This trait, the 'gene expression phenotype'[7], also shows familial aggregation[5,6,8] and simple segregation patterns in yeast[2], suggesting an inherited contribution. Here, we extend our analysis[6] by genetic mapping of regulatory elements that influence the baseline level of gene expression in human cells. Our goal is to identify determinants whose 'targets' are the genes with regulated expression.

We used microarrays to measure the baseline expression levels of genes in immortalized B cells from members of Centre d'Etude du Polymorphisme Humain (CEPH) Utah pedigrees[9]. For each of the ~8,500 genes on the array, we estimated the variance of expression level among unrelated individuals (94 CEPH grandparents) and the mean of variance of array replicates (two array replicates per individual). We restricted our analysis to genes with greater expression variation between individuals than between replicates (within individuals); these 3,554 most variable expression phenotypes are the quantitative traits that we mapped to chromosomal locations by genome scans.

Genotypes for genetic markers (single nucleotide polymorphisms; SNPs) were obtained from The SNP Consortium[10]. We used the computer program S.A.G.E. v. 4.5 (ref. 11) to carry out genome-wide linkage analysis for the 3,554 expression phenotypes in 14 CEPH families. The analysis gives the strength of the evidence for linkage at each map position in the form of a *t*-value[12], with associated point-wise significance level.

We selected expression phenotypes for further analysis using two different levels of stringency for evidence of linkage (that is, for a regulator of expression phenotype). For the more stringent level, we used a threshold of $t > 5$ from the S.A.G.E. analysis; in our sample of families, this corresponds to a point-wise *P*-value of $<4.3 \times 10^{-7}$ (a logarithm of odds (lod) score of ~5.3). For such a finding, the corresponding genome-wide significance level[13] (see Methods) is approximately 0.001. Applying this genome-wide threshold to 3,554 scans we would expect only 3.5 genome scans to show any linkage evidence with a *P*-value this extreme by chance. Instead we found 142 expression phenotypes with evidence for linkage beyond the *P*-value threshold, and in some cases far beyond, so we conclude that false-positive linkage findings are at most a small fraction of the significant results. The expression phenotypes with the most significant evidence of linkage are shown in Table 1.

In order to include additional phenotypes, we relaxed the stringency in some of the analyses by lowering the threshold to $t > 4$, which corresponds to a point-wise *P*-value of $<3.7 \times 10^{-5}$ (lod ~3.4) and approximately $P = 0.05$ genome-wide. There are 984 expression phenotypes that exceed this threshold, far more than the ~178 false positives expected by chance.

Cis and *trans* regulators of expression phenotypes

We consider the regions that are linked to the expression levels to be regulatory regions or 'regulators' of the expression phenotypes (of the target genes). We examined the regulatory regions for the 142 expression phenotypes with the most significant evidence of linkage, and for each quantitative phenotype we distinguished between apparently *cis*- and *trans*-acting regulators. We restricted the category of *cis* regulators to those that mapped within 5 megabases (Mb) of the target gene. This relatively large region was chosen to allow for imprecision of linkage and to include long-range regulators, as some *cis*-acting regulators act over megabase distances[14,15]. All other significant linkage represents *trans* regulators. By these definitions of *cis* and *trans*, we found the following distribution of phenotypes: 27 (19%) have only a *cis*-acting transcriptional regulator, 110 (77.5%) have only a *trans*-acting regulator, and 5 (3.5%)

Table 1 **Expression phenotypes with the strongest evidence of linkage from genome scans**

P-value	Gene	Location	*Cis*/*trans*
$<10^{-11}$	*ICAP-1A*	2q25	*Cis**
$<10^{-11}$	*TM7SF3*	12p11	*Cis**
$<10^{-10}$	*HSD17B12*	11p11	*Cis*
$<10^{-10}$	*CHI3L2*	1p12	*Cis*
$<10^{-10}$	*PSPHL*	7p11	*Cis*
$<10^{-10}$	*DSCR2*	21q22.2	*Trans*
$<10^{-10}$	*CBR1*	21q22.1	*Trans*
$<10^{-10}$	*HOMER1*	5q14	*Trans*
$<10^{-9}$	*DDX17*	22q13	*Cis*
$<10^{-9}$	*ZP3*	7q11	*Cis*
$<10^{-9}$	*IL16*	15q25	*Cis*
$<10^{-9}$	*ALG6*	1p31	*Trans*
$<10^{-9}$	*TNFRSF11A*	18q22	*Trans*

*The most extreme *P*-values occurred at SNPs located >5 Mb from the gene, but linkage evidence within 5 Mb of target gene was also highly significant ($P < 4.3 \times 10^{-7}$).

©2004 Nature Publishing Group

have two regulators (two phenotypes with a *cis*- and a *trans*-acting regulator, and three phenotypes with two *trans*-acting regulators).

Examples of the genome scan results for several expression phenotypes are shown in Fig. 1. We detected multiple regulators for only a small proportion of expression phenotypes, possibly because individual regulators make a smaller contribution when several regulators influence the expression level of a gene. Thus, some true regulators would not meet our criterion of $P < 4.3 \times 10^{-7}$ for evidence of linkage. We therefore also examined the 984 expression phenotypes with at least one marker significant at the reduced stringency of $P < 3.7 \times 10^{-5}$. Among these, we found 164 (16%) with multiple regulators of expression level, an appreciably higher percentage than the 3.5% found with the more stringent threshold. Multiple *trans*-acting regulators were found for 152 phenotypes, with both *cis*- and *trans*-acting regulators for the remaining 12.

Master regulators of transcription

In addition to genomic regions with regulators that affect single phenotypes *in cis* or *in trans*, we found genomic regions containing transcriptional regulators that influence multiple expression phenotypes. We divided the autosomal genome into 491 windows of 5 Mb and determined the number of regulators mapping to each window. We began by examining the regulators for the 142 expression phenotypes with $P < 4.3 \times 10^{-7}$. We found windows that contained many more 'hits' than expected by chance. If regulators for these phenotypes were distributed at random across the genome, the probability of six or more hits per window would be less than 6×10^{-5} and we would not expect to see any windows with more than four hits. Instead, we found two hotspots with six or more hits ($P < 0.03$ after Bonferroni correction): seven phenotypes mapped to one window on chromosome 14, and six phenotypes mapped to one window on chromosome 20.

When we relaxed our linkage criterion to include the 984 regions with $P < 3.7 \times 10^{-5}$, we found many more expression phenotypes whose regulation mapped to shared hotspots. The two regions indicated above contain *trans*-acting regulators for the most expression phenotypes (Fig. 2a). Regulation for 31 of the 984 expression phenotypes mapped to the 5-Mb window on chromosome 14 (14q32), and regulation for 25 phenotypes mapped to the window on chromosome 20 (20q13).

We consider the existence of hotspots to be evidence for master regulators of the baseline level of gene expression. The mapping was done without considering possible relationships among phenotypes, but the shared expression control regions suggest co-regulation. We therefore examined the correlation in expression levels of the 31 and 25 target genes corresponding to the two master regulatory regions. The expression levels in 94 CEPH grandparents were used. In permutation tests with 1,000 replications, we found that the pairwise correlation between any two expression phenotypes did not exceed 0.52. We therefore set 0.52 as the threshold for correlation by chance (nominal $P = 0.001$). Hierarchical clustering was used to summarize these results graphically and group genes by the correlations of their expression levels. We looked for clusters of expression phenotypes whose members have pairwise correlations that all exceed 0.52. Among the 31 target genes whose regulators

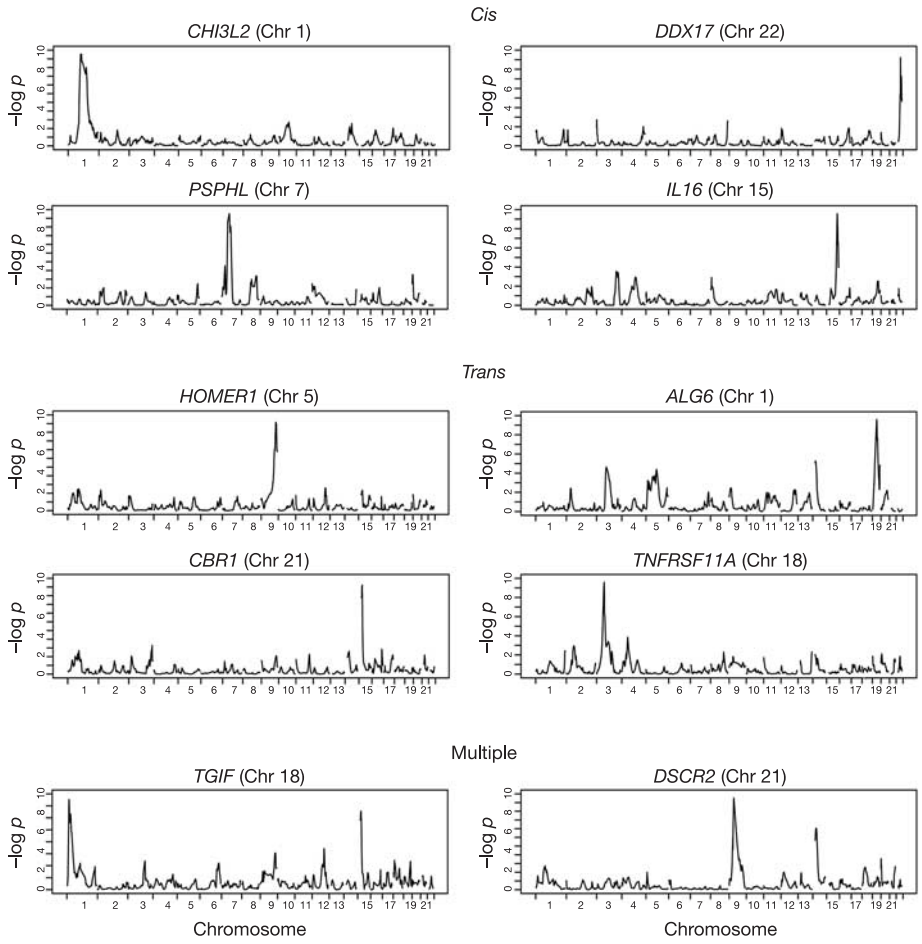

Figure 1 Genome scans for ten expression phenotypes. Chromosomal location of the regulated (target) gene is given in parentheses. The top eight panels show examples of linkage with *cis*- or *trans*-acting transcriptional regulators. The bottom two panels show examples of phenotypes regulated by several unlinked genetic determinants.

©2004 Nature Publishing Group

mapped to the chromosome 14 hotspot, we found one such regulated cluster with 14 genes, and three additional clusters each with two genes (Fig. 2b). Similarly, among the 25 phenotypes whose regulators mapped to the chromosome 20 region, we found one cluster of four and two clusters of two genes whose members have pairwise correlations that all exceed 0.52. The correlation in expression level of these genes supports the observation that they share common transcriptional regulators. However, the regulatory regions defined by mapping are still large, and there might be subgroups of co-regulated phenotypes that are influenced by distinct, but very closely linked, regulators.

Some sets of closely linked genes are influenced by the same *cis* regulators, and have correlated expression profiles[16–18]. In our data, some target genes whose expression levels map to a *trans*-acting master regulatory region are also located very close to each other. For example, among the target genes whose expression maps to the regulatory region on chromosome 14 are four genes (*MMP24*, *C20orf24*, *RPN2* and *TOP1*) found in a 6-Mb region of chromosome 20 (UCSC Genome Browser, version hg15). In addition, among the target genes of the regulatory region on chromosome 20 are two genes (*ITM2B*, *RB1*) separated by less than 50 kilobases (kb) on chromosome 13. These observations reflect a complex regulatory network where master transcriptional regulators affect baseline expression levels of many genes that have similar expression profiles, and in some cases, reside close together on human chromosomes.

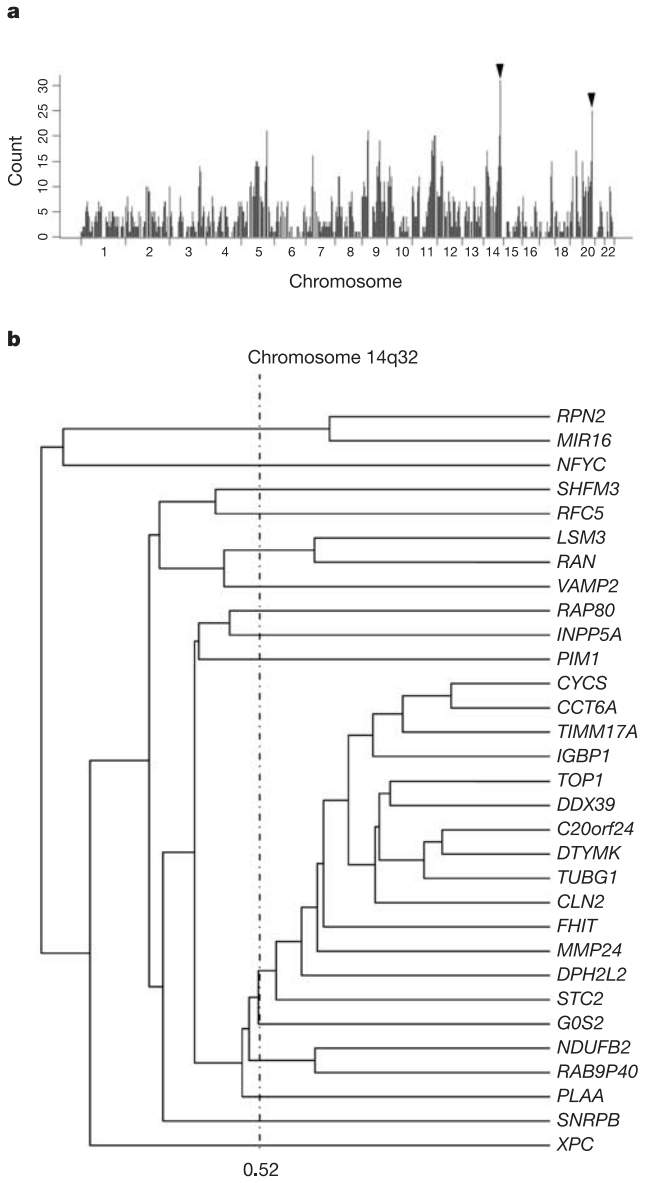

Figure 2 Master transcriptional regulators. **a**, Distribution of significant linkage peaks ($P < 3.7 \times 10^{-5}$) in 5-Mb windows across the autosomal genome. Arrowheads indicate the two windows (located on chromosomes 14 and 20) that contain regulators for the most expression phenotypes. **b**, Dendrogram representing hierarchical clustering of genes whose transcriptional regulators map to one 5-Mb window. Target genes for the hotspot of regulation on chromosome 14q32 are shown. Expression levels of genes with branches connected to the right of the dotted line are correlated at $P < 0.001$ (see text).

Family and population association analysis

Unlike linkage *in trans*, *cis* linkage of phenotypes immediately suggests a small region containing the regulatory element. This expectation led us to carry out follow-up studies with markers at several of the regulated genes. Among the 27 phenotypes with *cis* regulators (at $P < 4.3 \times 10^{-7}$), 17 were followed up by typing two or more additional SNP markers within or near the target gene. In each case, the expression data were used for both family-based and population-based analysis (Table 2). Analysis of the members of the 14 CEPH pedigrees by the Quantitative Transmission Disequilibrium Test (QTDT)[19] showed significant evidence ($P < 0.01$) for the combined presence of linkage and association at 14 (82%) of these 17 genes, strengthening our conclusions in several ways. First, the QTDT results confirm the mapping in these cases to the target genes. Second, the results therefore support the inferred regulation *in cis*. Finally, the results also imply differential allelic expression (see below).

These conclusions were extended by regression analysis of data from 94 unrelated CEPH grandparents. Marked associations were found for many genes between expression phenotype and closely linked SNPs. Figure 3 shows examples with leukocyte-derived arginine aminopeptidase (LOC64167 or *LRAP*) and 3-ketoacyl-CoA reductase (*HSD17B12*). The corresponding results for linear regression analysis are given in Table 2. Among the 17 phenotypes tested, the same 14 found to be significant by QTDT showed highly significant evidence ($P < 0.005$) for population association between gene expression level and a SNP located within or near the gene (Table 2), directly demonstrating differential allelic expression. For two genes (*TM7SF3*, *ICAP-1A*) that did not show significant association, the linkage peaks were exceptionally broad as well as high, and spanned more than 10 Mb. Although evidence at the target gene itself was also statistically significant, the highest linkage peak was located >5 Mb away.

Differential allelic expression

The degree of differential allelic expression detected varies considerably. The largest effect was found for the phosphoserine phosphatase-like (*PSPHL*) gene, where there was an approximately eightfold difference in mean expression level between individuals homozygous for different alleles of a SNP marker (rs6700). In contrast, for a SNP (rs7176604) in the coding region of cathepsin H (*CTSH*), the fold difference between CC and TT homozygotes was 1.44 (Table 2). To follow up this latter finding, we used allele-specific quantitative real-time polymerase chain reaction (qRT–PCR) to compare the expression of the two alleles of that marker in 30 heterozygous individuals. We found similar allelic differences in expression level (mean fold difference = 1.6; s.d. = 0.45). The QTDT results for *CTSH* strongly support this finding ($P = 3 \times 10^{-6}$). Thus, several related approaches confirm the allelic differences *in cis*, and imply linkage disequilibrium between a SNP genotype and a nearby determinant that influences expression phenotype.

The region of linkage disequilibrium associated with differential allelic expression of some phenotypes extends for large distances, up

©2004 Nature Publishing Group

to several hundred kilobases. For example, two SNPs at *HSD17B12*, separated by ~172 kb, show nearly identical correlations with gene expression level (Fig. 3). This strong linkage disequilibrium makes it difficult to narrow down the region that contains the sequence variant(s) responsible for variation in the expression level of *HSD17B12*. In order to determine whether studying other populations (which may have different linkage disequilibrium structure) can solve this problem, we examined the linkage disequilibrium pattern for the same genomic region in African-Americans and found much less linkage disequilibrium for the SNPs shown. Standardized linkage disequilibrium coefficient D' is estimated as 1.0 in CEPH, but only 0.116 in African-Americans. In general, smaller linkage disequilibrium will make it easier to localize determinants[20].

For the phenotypes listed in Table 2, we estimated how much of the variation in expression phenotype could be attributed to *cis*-acting regulators. We used the results of the linear regression analysis, and calculated R^2 (the customary estimate of variation explained by regression; see Methods). The last column of Table 2 shows the estimates obtained. These may be thought of as the 'heritability' attributable to the chromosomal region that is in linkage disequilibrium with the SNP tested, and therefore indicate what part of the variation in expression is influenced by *cis*-acting genetic determinants. For four of the genes in Table 2, this fraction is large—greater than 0.50. On the other hand, the fraction not explained in this way is also of interest, as it includes non-genetic causes (environmental differences, measurement variation) and possibly other genetic differences not in linkage disequilibrium with the SNP.

Discussion

Our study combined microarray expression data with publicly available SNP genotype data, and applied genome-wide mapping techniques to identify the chromosomal regions linked to the gene expression phenotypes. Level of gene expression is thus a trait like many others, and is amenable to genetic analysis. The classical

Table 2 **Properties of genes whose expression level is regulated by *cis*-acting determinants**

Gene	Location	Peak (*cis*) *P*-value (genome scan)	SNP (rs or ABI hCV identifier)	Highest/lowest ratio†	Population association‡ *P*-value	QTDT *P*-value	Variation explained (R^2)
LOC64167	5q15	1×10^{-7}	4869311	7.02	2.3×10^{-19}	6×10^{-24}	0.60
HSD17B12	11p11	2×10^{-11}	1061810	1.68	5.9×10^{-18}	6×10^{-22}	0.57
RPS26	12q13	2×10^{-9}	1506440	1.47	1.7×10^{-17}	1×10^{-15}	0.55
IRF5	7q32	2×10^{-8}	2280714	2.33	2.0×10^{-16}	2×10^{-19}	0.52
CSTB	21q22	2×10^{-9}	26539999	1.74	2.9×10^{-12}	1×10^{-13}	0.42
PSPHL	7p11	3×10^{-11}	6700	8.43	3.5×10^{-12}	3×10^{-14}	0.41
CHI3L2	1p12	3×10^{-11}	755467	3.69	1.7×10^{-11}	8×10^{-12}	0.39
CPNE1	20q11	1×10^{-7}	6060516	2.57	6.8×10^{-11}	4×10^{-13}	0.38
CTBP1	4p16	2×10^{-9}	2279282	1.70	6.0×10^{-10}	2×10^{-13}	0.34
PPAT	4q12	2×10^{-7}	2030364	1.56	3.1×10^{-6}	8×10^{-5}	0.21
VAMP8	2p11	9×10^{-8}	6547625	1.38	4.8×10^{-5}	2×10^{-10}	0.17
CTSH	15q25	7×10^{-9}	7176604	1.44	1.5×10^{-4}	3×10^{-6}	0.15
IL16	15q26	3×10^{-10}	4128767	1.43	5.0×10^{-4}	0.0029	0.13
ZP3	7q11	9×10^{-10}	306191*	2.70	2.3×10^{-3}	0.0011	0.10
GSTM2	1p13	3×10^{-8}	668413	NA	0.092	>0.5	0.03
TM7SF3	12p11	$<10^{-11}$	3134726*	NA	0.42	>0.5	<0.01
ICAP-1A	2p25	$<10^{-11}$	434836*	NA	0.79	>0.5	<0.01

*ABI hCV SNP identifier.
†Ratio of mean expression levels of homozygotes for SNPs. NA indicates that the regression was not significant.
‡*P*-value for regression of expression level on genotype.

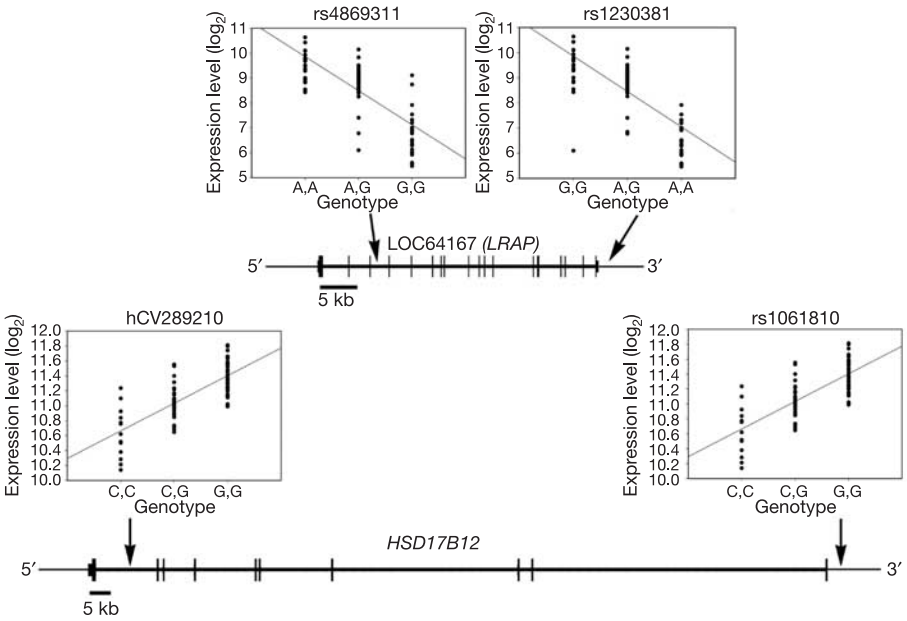

Figure 3 Regression of expression phenotype of LOC64167 and *HSD17B12* on nearby SNPs. For LOC64167, the distance between SNP markers rs4869311 and rs1230381 is 30 kb. For *HSD17B12*, the distance between hCV289210 and rs1061810 is 172 kb.

©2004 Nature Publishing Group

linkage strategy allowed us to identify numerous transcriptional regulatory loci, without any prior knowledge of the regulatory mechanism. It uncovered a complex network of regulation that includes determinants that influence expression of nearby genes (*cis*-acting), determinants located on other chromosomes (*trans*-acting), and hotspots of genetic determinants that affect the expression of many genes. The approach is reliable and accurate; results from association and differential allelic expression support the linkages *in cis*, suggesting that findings of linkages *in trans* are also valid. Our approach detects differential allelic expression without requiring that sequence variants be located in coding regions.

Many studies have shown that gene expression levels differ according to developmental stages, health and disease, and physiological or other biologically relevant states. However, little is known about natural variation in human gene expression, especially as a result of germ-line differences. Normal variation in gene expression is likely to account for a substantial part of human variation. It will therefore contribute to differences that are important for understanding essential aspects of human biology, including networks of interacting genetic effects, evolution, and susceptibility to complex diseases.

Mapping quantitative traits and unravelling transcriptional control are challenging, even when applied to one phenotype at a time. In studies of typical quantitative human traits like blood pressure or serum levels of metabolites, strong effects are rarely found. Here we have coupled genomic technologies for expression profiling with genome-wide genetic mapping using SNP markers, and shown that specific chromosomal regions contain germ-line determinants that regulate gene expression. These approaches and results show how it will be possible to dissect the genetic contribution to natural variation in human gene expression. □

Methods

CEPH samples and expression phenotyping

The data were from members of 14 CEPH families (CEPH 1333, 1340, 1341, 1345, 1346, 1347, 1362, 1408, 1416, 1418, 1421, 1423, 1424 and 1454). Expression and marker genotype data were available for all parents and a mean of eight offspring per sibship (range 7–9). (Data from grandparents are not used in SIBPAL.)

For the expression analysis, RNA was extracted from lymphoblastoid cells of each individual and hybridized onto Affymetrix Genome Focus Arrays per the manufacturer's protocol. Expression intensity was scaled to 500 and transformed by $\log_2$.

Genotypes

SNP genotypes for 2,756 autosomal SNP markers for individuals whose lymphoblastoid cells were phenotyped were downloaded from The SNP Consortium database of the SNP Consortium Linkage Map Project (http://snp.cshl.org/linkage_maps/). Most SNP Consortium SNPs are clustered in very closely linked sets (two or three SNPs within 100 kb) with average intercluster distance approximately 3 Mb. We used PedStat[21] to check for mendelian inconsistencies. This resulted in the removal of 815 genotypes at 237 distinct SNP markers.

Analysis of linkage and association

Multipoint genome-wide linkage analysis was done by SIBPAL in S.A.G.E[11]. We used the recommended option ('W4' SIBPAL) for weighting pairwise phenotypic differences between siblings[22]. SIBPAL determines evidence for linkage at each SNP from regression of the phenotype difference between siblings on the estimated proportion of marker alleles shared identical-by-descent between siblings; the result is reported as a *t*-value with corresponding significance, as given in the text. Point-wise significance was converted to genome-wide significance by use of the expression in ref. 13 (see http://www.imbs.mu-luebeck.de/pub/silcLOD/). In permutation analysis by S.A.G.E. of the results for the 142 phenotypes ($t > 5$, $P < 4.3 \times 10^{-7}$), we found one phenotype with one *t*-value >5 among 1,000 replicates; 100,000 additional permutations of this phenotype yielded two more *t*-values > 5. Further testing with 100,000 replicates for eight phenotypes (three with $4.1 \times 10^{-7} < P < 4.3 \times 10^{-7}$) yielded no *t*-values >5.

For association analysis, the $\log_2$-transformed expression level of 94 unrelated individuals (CEPH grandparents), as the dependent variable, was regressed on SNP genotype (coded 0, 1, 2). Conventional analysis of linear regression was carried out. R^2 was estimated for each phenotype/SNP combination as the ratio of regression sum of squares to total sum of squares.

Master regulator probability

The autosomal genome was divided into 491 windows of 5 Mb each (with smaller windows at the ends of chromosomes). For each of 142 phenotypes, we considered all SNPs with $t > 5$, $P < 4.3 \times 10^{-7}$. Any window with one or more such SNP was counted as having one 'hit' for that phenotype. Some phenotypes have more than one hit in the genome because some have multiple linkage peaks, or because peaks for some phenotypes are broad and span adjacent 5-Mb windows, in which case, each window is counted as having one hit. There were 318 hits defined this way, representing linkages for the 142 phenotypes. We assumed that if the hits were distributed randomly across the genome, their distribution over windows would be approximately Poisson, with mean 0.65 (318 out of 491).

Clustering

The similarity of the expression phenotypes that mapped to the hotspots of transcriptional control on chromosomes 14 and 20 was assessed by Pearson's correlation (absolute value), and the phenotypes were grouped by hierarchical clustering using the average linkage method. The significance of the correlation between genes was assessed by permutation. For each permutation, the expression levels of the 3,554 genes for each individual were permuted, and all $3{,}554 \times 3{,}553/2$ pairwise correlations were calculated. Among the 1,000 permuted sets, the highest pairwise correlation coefficient was 0.52.

SNP genotyping and allele-specific RT–PCR

SNPs in the region of the genes with *cis* regulators were identified using NCBI dbSNP or Applied Biosystems (ABI)/Celera Discovery System. DNA samples were genotyped using ABI TaqMan technology and the 7900 HT Sequence Detection System. PCR was carried out with primer and probe sets (ABI Assay-on-Demand and Assay-by-Design) according to the manufacturer's protocols. Allele-specific RT–PCR was performed using similar protocols with complementary DNA samples as template.

Received 22 March; accepted 5 July 2004; doi:10.1038/nature02797.
Published online 21 July 2004.

1. Oleksiak, M. F., Churchill, G. A. & Crawford, D. L. Variation in gene expression within and among natural populations. *Nature Genet.* **32,** 261–266 (2002).
2. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296,** 752–755 (2002).
3. Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **3,** 57–64 (2003).
4. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297,** 1143 (2002).
5. Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422,** 297–302 (2003).
6. Cheung, V. G. *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genet.* **33,** 422–425 (2003).
7. Cheung, V. G. & Spielman, R. S. The genetics of variation in gene expression. *Nature Genet.* **32,** 522–525 (2002).
8. Cheung, V. G. *et al.* Genetics of quantitative variation in human gene expression. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 403–407 (2003).
9. Dausset, J. *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6,** 575–577 (1990).
10. Matise, T. C. *et al.* A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.* **73,** 271–284 (2003).
11. S.A.G.E. Statistical Analysis for Genetic Epidemiology. (Statistical Solutions Ltd, Cork, Ireland, 2003).
12. Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2,** 3–19 (1972).
13. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11,** 241–247 (1995).
14. Nobrega, M., Ovcharenko, I., Afzal, V. & Rubin, E. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
15. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99,** 7548–7553 (2002).
16. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26,** 183–186 (2000).
17. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289–1292 (2001).
18. Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1,** 5 (2002).
19. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66,** 279–292 (2000).
20. McKenzie, C. A. *et al.* Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* **10,** 1077–1084 (2001).
21. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30,** 97–101 (2002).
22. Shete, S., Jacobs, K. B. & Elston, R. C. Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. *Hum. Hered.* **55,** 79–85 (2003).

Acknowledgements We thank T. Matise and W. Ewens for discussions and advice, and J. Burdick for technical help. Some analyses for this paper were carried out by using the program package S.A.G.E., which is supported by a grant from the National Center for Research Resources. This work is supported by grants from the National Institutes of Health (to R.S.S. and V.G.C.) and the W.W. Smith Endowed Chair (to V.G.C.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.G.C. (vcheung@mail.med.upenn.edu) or R.S.S. (spielman@pobox.upenn.edu). The GEO accession number for the microarray data is GSE1485.

©2004 Nature Publishing Group