nature genetics

Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes

Harald H H Göring[1], Joanne E Curran[1], Matthew P Johnson[1], Thomas D Dyer[1], Jac Charlesworth[1], Shelley A Cole[1], Jeremy B M Jowett[2,3], Lawrence J Abraham[4], David L Rainwater[1], Anthony G Comuzzie[1], Michael C Mahaney[1], Laura Almasy[1], Jean W MacCluer[1], Ahmed H Kissebah[5], Gregory R Collier[3,6], Eric K Moses[1] & John Blangero[1,3]

Quantitative differences in gene expression are thought to contribute to phenotypic differences between individuals. We generated genome-wide transcriptional profiles of lymphocyte samples from 1,240 participants in the San Antonio Family Heart Study. The expression levels of 85% of the 19,648 detected autosomal transcripts were significantly heritable. Linkage analysis uncovered >1,000 *cis*-regulated transcripts at a false discovery rate of 5% and showed that the expression quantitative trait loci with the most significant linkage evidence are often located at the structural locus of a given transcript. To highlight the usefulness of this much-enlarged map of *cis*-regulated transcripts for the discovery of genes that influence complex traits in humans, as an example we selected high-density lipoprotein cholesterol concentration as a phenotype of clinical importance, and identified the *cis*-regulated vanin 1 (*VNN1*) gene as harboring sequence variants that influence high-density lipoprotein cholesterol concentrations.

Phenotypic differences among individuals are partly the result of quantitative differences in transcript abundance. Although environmental stimuli may influence the location, timing, and/or level of transcription of specific genes, genetic differences among individuals are also known to have a significant role. Transcript levels may be thought of as quantitative endophenotypes that can be subjected to statistical genetic analyses in an effort to localize and identify the underlying genetic factors, an approach that is sometimes referred to as genetical genomics[1]. Using microarray technology, it is now possible to assess the abundance of many transcripts—and, indeed, of the entire known transcriptome—simultaneously. Studies that attempt to localize the genetic regulators of gene expression have been carried out in several species, including yeast[2–5], plants (maize and eucalyptus)[6,7], fly[8], mouse[6,9,10] and rat[11]. Several recent investigations have also focused on humans. In most of these studies, microarray-based gene expression profiles were generated for transformed cell lines derived from lymphocytes from members of the Centre d'Etude du Polymorphisme Humain (CEPH) families[12], and linkage and/or linkage disequilibrium approaches were used to map the genetic determinants that regulate the expression of individual transcripts[13–17]. These studies, reviewed in refs. 18,19, uncovered many interesting aspects of gene expression regulation in humans, but were limited in sample size (57–195 individuals). Other studies[20,21] focused on white blood cells, but also used small sample sizes (12–75 individuals). Two recently published studies focused on the relative importance of nucleotide versus copy number variation[22] and on the effect of ancestry[23] on gene expression variation in samples from the International HapMap Project. Here we report on genetic analyses of genome-wide transcriptional profile data from lymphocytes obtained from participants in the San Antonio Family Heart Study (SAFHS). Apart from using natural tissue rather than clones of transformed cells, our study differs by its much larger scale, with lymphocytes available from 1,240 individuals and transcript-specific expression levels detected with 20,413 unique oligonucleotide probes, allowing us to build a much more detailed map of *cis*-regulated transcripts. To highlight the usefulness of this resource for the discovery of the genetic factors that influence complex diseases in humans, we focused on high-density lipoprotein cholesterol (HDL-C) concentrations as a phenotype serving as an example and identified a promoter variant of the vanin 1 (*VNN1*) gene as influencing HDL-C concentrations.

RESULTS

Gene expression profiles were generated on lymphocyte samples from randomly ascertained participants in the SAFHS—a study that investigates the genetics of cardiovascular disease in Mexican Americans[24]. Peripheral blood samples, with subsequent lymphocyte preparation,

[1]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas 78245, USA. [2]International Diabetes Institute, Caulfield, Victoria 3162, Australia. [3]ChemGenex Pharmaceuticals, Geelong, Victoria 3216, Australia. [4]School of Biomedical, Biomolecular and Chemical Sciences, University of Western Australia, Crawley, Western Australia 6009, Australia. [5]TOPS Center for Obesity and Metabolic Research, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. [6]Metabolic Research Unit, School of Health Sciences, Deakin University, Waurn Ponds, Victoria 3216, Australia. Correspondence should be addressed to H.H.H.G. (hgoring@sfbrgenetics.org).

Received 7 December 2006; accepted 3 August 2007; published online 16 September 2007; doi:10.1038/ng2119

Table 1 Counts and proportions of heritable transcripts by FDR and heritability estimate threshold

		All transcripts		RefSeq transcripts		Non-RefSeq transcripts	
FDR	Heritability estimate threshold	Number	Percentage	Number	Percentage	Number	Percentage
0.1		17,361	88.4	12,374	92.8	4,941	78.3
0.05		16,678	84.9	12,086	90.6	4,527	71.8
0.01		15,310	77.9	11,493	86.1	3,740	59.3
0.001		13,804	70.3	10,773	80.8	2,941	46.6
	0.1	15,142	77.1	11,397	85.4	3,745	59.4
	0.2	10,880	55.4	9,045	67.8	1,835	29.1
	0.3	6,370	32.4	5,539	41.5	831	13.2
	0.4	2,606	13.3	2,319	17.4	287	4.6
	0.5	826	4.2	728	5.5	98	1.6

FDR, false discovery rate.

were obtained in the morning after an overnight fast. Although 1,280 frozen tissue samples were available, high-quality RNA could be obtained from only 1,240 individuals. Most of these individuals (89%) are members of 30 extended families, with 10–87 phenotyped individuals spanning up to four generations of each family. The average number of phenotyped sibs per sibship is ~2.6 (with a maximum of 11). The sample contains 506 men (40.8%) and 734 women (59.2%), with a mean age of 39.3 years (ranging from 15 to 94).

Genome-wide transcriptional profiles were generated using Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChips, which contain 47,289 unique 50-mer oligonucleotides in total, with hybridization to each probe assessed at ~30 different beads on average. 22,151 probes (47%) are targeted at Reference Sequence (RefSeq)[25] transcripts, and the remaining 25,138 probes (53%) are for other, generally less well characterized transcripts (including predicted transcripts). A χ^2 'tail' test was used to assess whether there was a significant excess of samples with transcript-specific expression values above the 95th percentile of the null distribution based on manufacturer-provided negative control samples. This allowed the 'detection' of even those RNA molecules that are clearly present above baseline levels in some individuals. Using this test and a false discovery rate (FDR)[26] of 5%, we identified 20,413 phenotypes (43.2%) with significant expression. The proportion of detected transcripts was substantially higher among RefSeq genes (62.5%) than non-RefSeq genes (26.1%), reflecting the greater degree of knowledge and certainty about the existence of RefSeq transcripts[25]. Based on information about the physical location, we were able to place 19,648 probes at specific positions on autosomal chromosomes, and the remainder of our investigations focused on this subset. Some genes are represented by >1 oligonucleotide (with a maximum of 4 probes per gene), such that the 19,648 unique probes represent transcripts from ~18,519 different genes. For simplicity, and in agreement with most other genetic investigations of microarray-based expression profiles, all probes were treated as distinct entities in this examination. **Supplementary Table 1** online provides a summary of the expression phenotype count and density by chromosome. Significantly detected transcripts cover the autosomes at an average density of ~6.9 transcripts per 1 Mb, or 1 transcript every ~150 kb.

Estimation of heritability

A series of standardization steps were used to make the expression phenotypes comparable across individuals and across transcripts, resulting in normally distributed expression phenotypes (see Methods). To assess the influence of aggregate genetic variability among individuals on message-specific expression levels, we carried out variance components-based heritability analyses. To account for the effects of sex and age, which influence the expression of many genes in flies[27] and humans (data not shown), sex and sex-specific age and age-squared terms were included in the analytical model as linear predictors of transcript-specific abundance. All expression phenotypes were treated in this manner, irrespective of whether significant covariate effects were found, and the presented heritability estimates are indicative of the importance of genetic factors after accounting for sex and age effects. **Table 1** shows the proportions of transcripts found to be heritable. At an FDR of 5%, 16,678 transcripts (84.9%) were found to be significantly heritable. A histogram of the estimated additive heritabilities is shown in **Figure 1**. The median heritability estimate among all expressed transcripts is 22.5%. It seems that the abundance of most transcripts is significantly influenced by the genetic constitution of an individual. However, many expression phenotypes have modest heritability estimates, suggesting the substantial influence of the environment and the physiological state of an individual at the time of blood draw (for example, the influence of the time of day at tissue sampling on gene expression[21]) and/or measurement error.

Identification of *cis*-regulated transcripts

We next carried out linkage analysis to map the genetic factors that influence the expression level of individual transcripts (which are often referred to as expression quantitative trait loci, or eQTLs). Genotypes were available for 1,345 individuals (including all those with expression profiles) at 432 highly polymorphic microsatellite

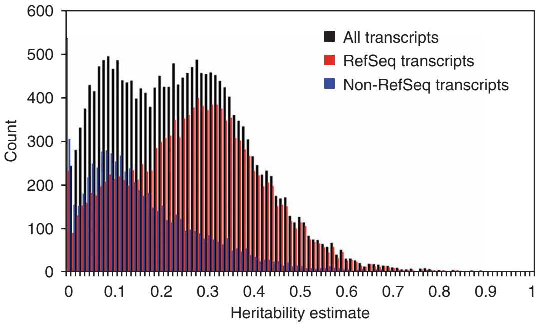

Figure 1 Histogram of heritability estimates for significantly detected transcripts after adjusting for the average expression level of all transcripts in an individual and for the effects of sex and age.

Table 2 Counts and proportions of *cis*-regulated transcripts under different criteria

				All transcripts		RefSeq transcripts		Non-RefSeq transcripts	
FDR	Lod score threshold	Pointwise *P* value	Locus-specific heritability estimate threshold	Number	Percentage	Number	Percentage	Number	Percentage
0.1				1,706	8.7	1,533	11.5	236	3.7
0.05				1,345	6.8	1,199	9.0	205	3.3
0.01				935	4.8	795	6.0	146	2.3
	0.588	0.05		3,222	16.4	2,581	19.3	641	10.2
	1.175	0.01		1,786	9.1	1,468	11.0	318	5.0
	2.074	0.001		1,071	5.5	882	6.6	189	3.0
	3	0.0001		750	3.8	618	4.6	132	2.1
	4	9×10^{-6}		571	2.9	475	3.6	96	1.5
	5	8×10^{-7}		460	2.3	385	2.9	75	1.2
			0.1	3,043	15.5	2,484	18.6	559	8.9
			0.2	979	5.0	818	6.1	161	2.6
			0.3	464	2.4	388	2.9	76	1.2
			0.4	245	1.2	207	1.6	38	0.6
			0.5	128	0.7	111	0.8	17	0.3

FDR, false discovery rate.

markers across all autosomes, with an average intermarker spacing of <10 cM. Joint genotype data from several markers were used to estimate the multipoint identity-by-descent sharing probabilities for all relative pairs at 1-cM intervals using the Monte Carlo multipoint method implemented in the Loki computer program[28]. Linkage analysis was carried out using a variance components model and the SOLAR computer package[29].

We first sought to identify the transcripts that contain, within (or near) their own locus, sequence variants that substantially influence their own abundance. We will refer to this phenomenon as *cis* regulation. To impute the genetic locations that correspond to the physical locations of individual transcripts, we used the genetic map developed by deCODE genetics[30] and a linear interpolation procedure to place transcripts between their nearest flanking markers, based on the physical locations of the markers and the transcripts. In order to assess *cis* regulation, we computed the multipoint lod score at a single location, namely at the (integer) centimorgan location nearest the underlying structural gene (this contrasts with the approach taken by most other studies, which have focused on the highest lod score peak in some interval around a gene). As a result, only a single pointwise test was carried out, and no correction for multiple testing was required to interpret the obtained lod scores. All transcripts, including the small proportion of transcripts with insignificant heritability estimates, were included in the analysis, because *cis*-regulatory effects may still be detectable even if the overall heritability estimate is insignificant. At an FDR of 5% we identified 1,345 *cis*-regulated transcripts (6.8%) (**Table 2**). These results substantially increase the number of known *cis*-regulated transcripts over earlier studies that

Table 3 Counts and proportions of expression phenotypes with significant *trans* eQTLs and comparison of *cis* and *trans* eQTL frequencies using the lod score threshold

Lod score threshold	Approx. *trans* FDR (genome-wide analysis)	Number of transcripts with *trans* eQTLs	Transcripts with *trans* eQTLs (%)	Number of *trans* eQTLs	*Cis* FDR (pointwise analysis)	Number of *cis* eQTLs	eQTLs located in *cis* (%)
All transcripts							
3	~0.67	1,072	5.5	1,138	2.6×10^{-3}	750	39.7
4	~0.59	108	0.5	109	3.0×10^{-4}	571	84.0
5	~0.22	25	0.1	25	3.4×10^{-5}	460	94.8
10	$\sim 4.6 \times 10^{-6}$	6	0.03	6	5.1×10^{-10}	221	97.4
RefSeq transcripts							
3	~0.56	873	6.5	929	2.2×10^{-3}	618	39.9
4	~0.49	88	0.7	89	2.5×10^{-4}	475	84.2
5	~0.21	18	0.1	18	2.8×10^{-5}	385	95.5
10	$\sim 3.7 \times 10^{-6}$	5	0.04	5	4.2×10^{-10}	183	97.3
Non-RefSeq transcripts							
3	→1	199	3.2	209	4.8×10^{-3}	132	38.7
4	→1	20	0.3	20	5.8×10^{-4}	96	82.8
5	~0.25	7	0.1	7	6.7×10^{-5}	75	91.5
10	$\sim 8.8 \times 10^{-6}$	1	0.02	1	9.6×10^{-10}	38	97.4

Cis lod scores were obtained in a pointwise multipoint linkage test at the (integer) centimorgan location nearest the genetic location of the structural locus of a given transcript. *Trans* lod scores were recorded from multipoint linkage analysis across all autosomal *trans* chromosomes (in other words, all chromosomes other than the chromosome on which the structural locus resides), counting no more than one lod score peak per chromosome. Although the comparison of *cis* and *trans* eQTLs (expression QTLs) is carried out on the same scale, the same lod score thresholds have different FDRs (false discovery rates) associated with them, as indicated, depending on whether the linkage analysis is genome-wide (as in the discovery of *trans* eQTLs) or pointwise (as in the discovery of *cis* eQTLs). In the San Antonio Family Heart Study, the approximate genome-wide significance levels corresponding to lod scores of 3, 4, 5 and 10 are 0.039, 0.0033, 0.00028 and 1.4×10^{-9}.

Table 4 Consistency of linkage findings for *cis* and *trans* regulation across studies

Gene symbol	Probe ID	Chr.	Genetic location (cM)	*P* value (from ref. 13)	Location of linkage peak (from ref. 13): *cis* or *trans* (chr.)	*P* value (this study)
ITGB1BP1 (*ICAP-1A*)	GI_20143949-A	2	24	$<10^{-11}$	*Cis*	3×10^{-32}
TM7SF3	GI_7706574-S	12	50	$<10^{-11}$	*Cis*	6×10^{-9}
HSD17B12	GI_7705854-S	11	60	$<10^{-10}$	*Cis*	9×10^{-3}
CHI3L2	GI_11993934-S	1	134	$<10^{-10}$	*Cis*	6×10^{-4}
DDX17	GI_38201711-I	22	49	$<10^{-9}$	*Cis*	3×10^{-3}
POMZP3 (*ZP3*)	GI_23510405-A	7	90	$<10^{-9}$	*Cis*	2×10^{-21}
IL16	GI_27262656-I	15	88	$<10^{-9}$	*Cis*	4×10^{-22}
DSCR2	GI_44680112-S	21	50	$<10^{-10}$	*Trans* (9)	0.07
CBR1	GI_4502598-S	21	43	$<10^{-10}$	*Trans* (15)	0.06
HOMER1	GI_20127465-S	5	96	$<10^{-10}$	*Trans* (9)	0.20
ALG6	GI_38026891-S	1	91	$<10^{-9}$	*Trans* (19)	0.16

This table is based on Table 1 in ref. 13, which reported pointwise *P* values (in other words, not corrected for multiple testing) for the 13 genes with the highest evidence of linkage in their study. Transcripts for 11 out of the 13 genes were significantly detected in this study and the comparative pointwise *P* values are shown here. (For three *cis*-regulated genes, hybridization to more than one oligonucleotide probe was significantly detected in this study. The results for the probe giving the most significant *P* value is shown here.) In ref. 13, *cis P* values were based on the maximum obtained lod score within a 5-Mb interval on either side of the structural locus of a given transcript, whereas *cis P* values in this study are based on a single pointwise multipoint linkage test at the nearest integer centimorgan location. The pointwise significance level of evidence for *trans* regulation in this study are based on the maximum lod score on the *trans* chromosome where the highest evidence of linkage was observed, as taken from Figure 1 in ref. 13. The genetic map locations are based on the genetic map developed by deCODE genetics[30], using an interpolation procedure based on physical locations to place markers that were genotyped in the San Antonio Family Heart Study but that were not on the map from deCODE genetics. Chr. chromosome.

were of a smaller scale. Many *cis* lod scores are of substantial magnitude (**Table 3**). **Supplementary Table 2** online lists the transcripts that give the 20 highest *cis* lod scores. The maximum *cis* lod score (obtained with the probe Hs.379903-S) was 52.5, with a corresponding eQTL-specific heritability estimate of >77% (although this estimate is probably biased upward because of the so-called 'winner's curse'[31], which in this case results from maximizing the *cis* lod score and its associated locus-specific heritability estimate over the different transcripts analyzed). For many of these significant *cis* eQTLs, genetic variants in or near the gene are estimated to explain a large portion of the variance in message abundance (for example, 128/56/42 *cis* eQTLs have locus-specific heritabilities of ≥50/≥60/≥70%), which is not inconsistent with these expression phenotypes being essentially monogenic. Among all detected transcripts, the median (mean) *cis* eQTL effect size is 1.8% (5.0%), which suggests that *cis* regulation accounts for about 5% of variation in expression levels overall, and at least 2% of variation in nearly half of all transcripts. Among the 1,345 transcripts that are *cis* regulated at an FDR of 5%, the median (mean) *cis* eQTL effect size is 24.6% (29.1%). These numbers highlight the importance of *cis*-regulatory effects on gene expression in general. **Supplementary Figure 1** online shows the relationship between overall heritability and *cis* eQTL-specific heritability, indicating that the phenomenon and strength of *cis* regulation increases in frequency with an increase in transcript heritability. An investigation of the distribution of *cis* regulation throughout the genome failed to yield evidence for clusters of *cis*-regulated transcripts, which may indicate the existence of mid- or long-range enhancer/repressor elements that influence the expression levels of nearby genes (data not shown).

Examination of *trans* regulation

One might expect that the expression levels of individual transcripts are simpler in etiology and closer to the action of individual genes than are the vast majority of complex disease phenotypes that are the primary target of gene mapping experiments nowadays. If that assumption holds, then major regulatory loci ought to exist and be individually mappable using statistical genetic approaches. To examine this hypothesis, we carried out genome-wide linkage analysis in an effort to localize *trans*-acting genetic factors that influence gene expression. For conceptual clarity, we will use the term *trans* only for factors that are located on a chromosome other than that on which the gene for a given message resides, but not for factors that are located far away on the same chromosome. Despite having the power to detect linkage at a lod score threshold of 3 in our large sample of extended families (with 80% and 90% power to detect loci explaining roughly 22% and 24% of the phenotypic variance, respectively, depending also slightly on the overall heritability[32]), we were able to map *trans*-acting regulators for only a small proportion of the analyzed phenotypes (**Table 3**). The proportions of linkage peaks above lod score thresholds of 3, 4 and 5 are elevated over what would be expected by chance, but the FDRs that correspond to these lod score thresholds in the entire set of analyzed autosomal transcripts are high (**Table 3**). Hence, many of these *trans* eQTLs are expected to be false positives. We obtained two lod score peaks of ≥3 for 58 expression phenotypes (0.3%) and three peaks of this magnitude for 4 transcripts (0.02%). The small number of significant *trans* eQTLs suggests that expression phenotypes in natural tissues such as lymphocytes may not be as simple as might be naively expected. Given the strong heritability of most transcripts, it may be the case that their individual expression is controlled by many *trans*-acting factors of individually small to modest effect, such that these *trans* regulators are individually difficult to localize using linkage analysis. No notable evidence of master regulators located in *trans* was obtained using several approaches (data not shown).

A comparison between *cis* and *trans* lod scores shows the marked differences in frequency and magnitude, especially if one takes into account the fact that *cis* and *trans* lod scores of the same numerical value have entirely different FDRs (because of the different multiple testing situations: pointwise testing to assess *cis* regulation and genome-wide testing for *trans* regulation). Even if lod scores are compared at their nominal values, thus ignoring the difference in the multiple testing situation and thereby favoring *trans* effects, many of the most significant linkages are located in *cis* (**Table 3**). Nearly half (44.2%) of all lod scores above the traditional lod score threshold of 3

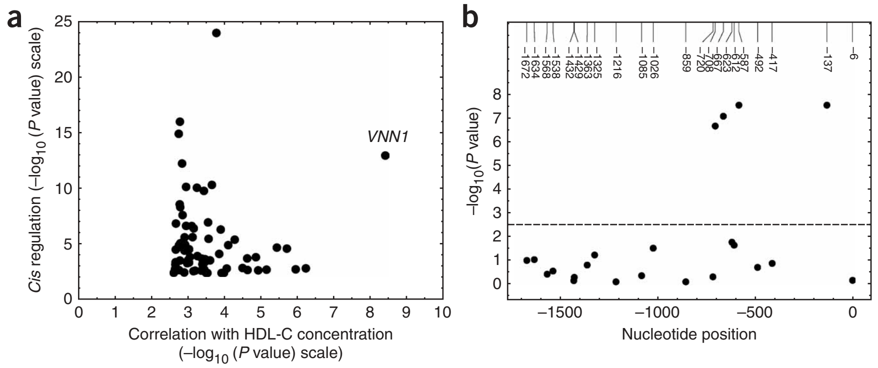

Figure 2 Identification of *VNN1* (vanin 1) as an HDL-C (high-density lipoprotein cholesterol) candidate gene and association analysis of *VNN1* promoter variants and *VNN1* transcript abundance. (**a**) The scatter plot includes all those transcripts that are significantly *cis* regulated (at a false discovery rate (FDR) of 0.05) and also significantly correlated with HDL-C concentrations (FDR of 0.05). *VNN1* (labeled in the figure) stands out as being highly significant in both dimensions and was therefore selected as an HDL-C candidate gene for detailed follow-up. Both axes are on a $-\log_{10}$ scale. (**b**) Evidence for association between *VNN1* promoter variants and *VNN1* transcript levels in 96 unrelated individuals using an additive measured genotype model.

were located in *cis*, and the proportion of *cis* lod scores was further elevated at higher thresholds. These findings do not imply that *trans* regulation of gene expression is rare or unimportant, but that the regulators with the strongest effect tend to be located in *cis*. A list of the 20 expression phenotypes with the most significant *trans* eQTLs is provided in **Supplementary Table 3** online. **Supplementary Table 4** online contains information on the heritability estimate, *cis* lod score and maximum *trans* lod score, along with other information, for all of the 19,648 autosomal transcripts analyzed.

Consistency in eQTLs between studies

To examine whether the evidence for eQTLs is consistent between different studies, we compared our results with those from a previous study[13]. In their paper Table 1 contains a list of 13 expression phenotypes with the strongest linkage evidence in genome-wide analysis. We detected significant expression of transcripts for 11 out of the 13 genes in this study, and the comparative results are shown in **Table 4**. Despite numerous differences between the studies, the designation of transcripts as *cis* regulated is consistent. Pointwise multipoint linkage analysis at the location of each of the 7 *cis*-regulated transcripts yielded substantial evidence of *cis* effects, with all *P* values $<10^{-2}$. By contrast, linkage analysis of the four transcripts with significant *trans* eQTLs in the Morley study did not yield pointwise *P* values ≤ 0.05 (equivalent to a lod score ≥ 0.588) anywhere on the reported *trans*-regulating chromosome. Therefore, the findings of *cis* regulation are consistent between both studies, whereas the findings of *trans* regulation are not.

Utility of a *cis*-regulated transcript map for gene discovery

We sought to empirically evaluate the usefulness of the expression profile data and the map of *cis*-regulated transcripts for the rapid identification of novel candidate genes underlying complex traits. To show the usefulness of this resource, we used the concentration of HDL-C to serve as an example; HDL-C is almost universally accepted to increase the risk of cardiovascular disease at a low concentration[33]. HDL-C concentration was measured in plasma samples taken during the same visit to the clinic as the blood samples that were used for expression profiling[24].

As a first step, we examined the correlation between expression measures of all *cis*-regulated transcripts (at a 0.05 FDR) and the concentration of HDL-C, after making adjustments for sex and age as before. We focused exclusively on *cis*-regulated expression phenotypes because the underlying genes are expected to harbor genetic variants that substantially influence their own expression levels, and thus, among transcripts whose abundance is correlated with HDL-C concentration, these genetic polymorphisms are good candidates to influence HDL-C as well. Sixty-seven *cis*-regulated transcripts were found to be correlated with HDL-C at an FDR of 5% (**Fig. 2a**). Expression of the *VNN1* gene (labeled in the figure) provides particularly strong support both for *cis* regulation (*cis* lod = 11.7; *P* value = 1.1 × 10^{-13}) and for association with HDL-C concentration (*P* value = 4.0 × 10^{-9}). In bivariate analysis, the genetic correlation of *VNN1* expression level and HDL-C concentration was estimated as 0.28. *VNN1* codes for pantetheinase and produces cysteamine, a potent antioxidant that prevents lipid peroxidation[34]. Therefore, *VNN1* is a reasonable candidate gene for HDL-C concentration.

We thus proceeded to resequence the putative proximal promoter of this gene in 96 unrelated Mexican American founders of the SAFHS families, focusing on a region approximately 2 kb upstream from the transcriptional start site. Twenty-two SNPs were detected. Association analysis, using an additive model of allelic effect, was used to test for correlation between genotypes at individual promoter SNPs and transcript abundance[35]. Strong support for association was observed for several SNPs, with the SNP located at –137 bp relative to the transcriptional start site giving the most significant evidence ($P = 6.8 \times 10^{-7}$) (**Fig. 2b**). The observed association validated our initial linkage-based inference that *VNN1* was likely to be *cis* regulated.

We subsequently genotyped the six most highly associated SNPs in the entire sample of 1,240 individuals with transcriptional profiles. Association analysis, embedded within a variance components-based linkage model to account for the overall genetic relationship and pointwise genetic relationship at the *VNN1* gene among the individuals[35,36], provided overwhelming evidence for the association of five of these SNPs with *VNN1* transcript abundance (**Table 5**). Similarly, we observed strong evidence for the association of two of these SNPs with HDL-C concentration (**Table 5**). These findings

Table 5 Association between *VNN1* promoter variants and *VNN1* transcript abundance and HDL-C concentration in entire sample

VNN1 promoter variant	Association with *VNN1* transcript level (*P* value)	Association with HDL-C concentration (*P* value)
G–708A	3.0 × 10^{-34}	1.2 × 10^{-3}
G–667A	8.3 × 10^{-7}	2.7 × 10^{-5}
T–623C	0.11	0.84
C–612A	2.0 × 10^{-36}	0.13
A–587G	6.9 × 10^{-85}	1.9 × 10^{-3}
G–137T	5.7 × 10^{-83}	4.0 × 10^{-4}

HDL-C, high-density lipoprotein cholesterol; *VNN1*, vanin 1.

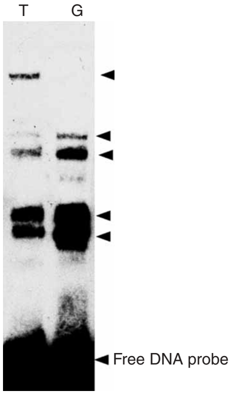

Figure 3 EMSA (electrophoretic mobility shift assay) analysis of the *VNN1* (vanin 1) –137 polymorphism and surrounding sequences. Approximately 20 fmol of probes were incubated with 3 μg of Jurkat nuclear extract for 30 min, analyzed by electrophoresis on a non-denaturing polyacrylamide gel, blotted to nylon membrane, and the DNA/protein complexes were visualized on X-ray film following incubation with streptavidin-horseradish peroxidase and development with luminal substrate. The lanes denoted T and G contained the *VNN1* –137T and –137G sequences, respectively. The position of the unbound probe is indicated. Arrowheads indicate the positions of complexes that preferentially bind to one of the alleles.

suggest that this approach of using transcriptional profile data, coupled with information on which transcripts are *cis* regulated, may be a suitable approach for the discovery of novel genes (and variants therein) that influence complex traits.

Bioinformatic analysis using P-Match[37,38] revealed that the *VNN1* –137T allele is embedded in a consensus Sp1 (stimulating protein 1) binding site[39] (TRANSFAC 6.0 Core similarity = 1.000, matrix similarity = 0.964) GGTAG sequence, whereas the –137G allele ablates the core GT box sequence (TRANSFAC 6.0 Core similarity = 0.985, matrix similarity = 0.952) to GGGAG. Sp1 is a member of the Kruppel-like factor (KLF) transcription factor family, all members of which recognize similar G-rich sequences[40]. Other ubiquitous and tissue-specific family members are also likely to bind the –137T and/or –137G sequence. KLF transcription factors interact with and stabilize binding of other transcription factors[41]. Therefore, this promoter variant has some indirect support for functionality.

We subsequently carried out comparative electrophoretic mobility shift assays (EMSA) of both alleles at the –137 promoter SNP (**Fig. 3**). Both the T and the G alleles preferentially form unique DNA-transcription factor complexes using nuclear extracts that are derived from the human T-cell line Jurkat, indicating that more than one factor is involved in binding to the sequence in an allele-specific way. Other lymphocyte cell line nuclear extracts show similar differential binding patterns (data not shown). These findings strongly support the contention that the T/G polymorphism at position –137 has functional consequences with respect to binding transcription factors.

DISCUSSION

The technology that allows the simultaneous assessment of the expression of virtually all known transcripts throughout the genome has only recently become available, and the analysis and use of transcriptional profile data are in their infancy. We still know very little about the role of quantitative gene expression variation in phenotype determination and about the genetic determinants of transcript levels. Genetic investigations of gene expression in humans have almost exclusively focused on lymphoblastoid cell lines from CEPH family members[13–17] and were severely limited in sample size. Here we have presented results from genetic analyses on transcriptional profiles from lymphocytes in a much larger sample of individuals.

There are many important differences between cell lines and natural tissues. The advantages of cell lines with regard to genetic investigations of transcriptional regulation include near complete control over the environment and the existence of only a single cell type. Both of these features undoubtedly reduce the range of factors that influence the expression phenotype, increasing power for genetic investigations. By contrast, natural tissues, including lymphocytes, contain a multitude of cell types (which probably vary in the expression levels of many transcripts[21]), and gene expression may be under the influence of many environmental factors that affect the study participant, thus creating noise for genetic studies. On the other hand, cells are subject to gross manipulation during the process of creating cell lines and are kept under artificial conditions; it is therefore not certain that gene expression profiles generated on cell lines are an accurate representation of the natural gene expression state *in vivo*, which complicates the interpretation of genetic findings and of analyses that correlate gene expression with physiological phenotypes of interest.

Despite these differences, some of our findings match those from earlier investigations carried out on lymphoblastoid cell lines. Our heritability estimates closely match those in ref. 14. At first glance, their reported median heritability estimate of 34% seems higher than our estimate of 22.5%. However, their estimate is based on only those transcripts found to be heritable at an FDR of 5%, and, for the most part, given the small sample size, heritability estimates were only significant if they were at least ~20% (see Fig. 1 in ref. 14). By contrast, our estimate is based on all transcripts. If we restrict estimation to the subset of transcripts that have estimated heritabilities ≥0.2, we obtain a median heritability estimate of 32%, which is similar to their estimate.

Our observation of the preponderance of *cis*-acting signals among significant linkages also agrees with most[13–16] but not all[17] studies. The designation of transcripts as being significantly *cis* regulated was found to be consistent between cell lines and lymphocytes (and across different technology platforms), with significant *P* values observed in this study for all seven *cis*-regulated expression phenotypes listed in ref. 13. This finding suggests that *cis* regulation is a stable characteristic of individual transcripts that is relatively immune to environmental influences and that may be consistent among different cell types and across tissues[11]. The *cis*-regulated transcripts identified here are therefore not primarily due to differences between individuals in lymphocyte cell type proportions[21], which may themselves be under genetic control. It seems unlikely that the many observations of *cis* regulation are artifacts that are due to the presence of polymorphisms that affect probe binding[10,11], although binding-site polymorphisms[14] cannot be ruled out for any given transcript without further examination. And variation in copy number rather than in (pointwise) sequence may be the cause of the *cis* regulation of a proportion of *cis*-regulated transcripts[22]. Future characterization of the allelic architecture of *cis* regulation should reveal whether different transcript levels are primarily due to different transcription levels or different RNA stability.

By contrast, we could not corroborate *trans* regulation, failing to find evidence of linkage on the listed *trans* chromosomes for even a single one of four transcripts that were reported to have significant *trans* eQTLs[13]. It seems unlikely that this disagreement can be easily explained either by the *trans* eQTLs listed in ref. 13 being false positives (given the listed significance levels of $<10^{-9}$, which survive adjustments for multiple testing that are due to genome scanning and analysis of many transcripts) or by the insufficient power of this study. Hence, *trans* regulation may be more sensitive to environmental influences and/or vary between different cell types[11]. Studies in yeast[2,3] and mice[6,9,10] have found evidence for the existence of master regulatory regions that are located in *trans*. The existence of such regulatory hubs in humans was suggested in ref. 13, but we and others[14] have not observed such clustering of *trans* regulation. It is not clear whether previous evidence for *trans*-acting hubs of gene expression regulation in humans is an artifact[19,42]. However, we did not look for such regulatory hubs among transcripts that belong to a specific class of genes.

If *cis* regulation is relatively stable across different cell types and tissues, then the detailed map of substantially *cis*-regulated transcripts identified in this paper ought to be of great potential value for the identification of candidate genes, irrespective of the tissues that are involved in a particular disease or other trait. Within positional candidate regions identified by linkage and/or association, gene expression data can be used to prioritize genes in several ways. Genes with an expression level that is significantly (genetically) correlated with the trait of interest are of higher priority than those that are not. Given that an observed correlation does not imply directionality (gene expression may be a consequence of the disease rather than a cause)[43], it may be advantageous to obtain gene expression data prospectively in order to increase the likelihood that correlated transcripts act upstream of the target phenotype. *Cis*-regulated genes are of interest because the underlying genes are expected to harbor genetic variants that influence their own expression level, which may also influence the physiological trait of interest if transcript abundance is correlated with the target phenotype (such as HDL-C concentration). Once functional regulatory sequence variants are identified, direct causal relationships between such genes and downstream transcriptional and physiological components can be identified. Schadt and colleagues[43] outline a sophisticated 'integrative genomics' approach to identify causal associations between transcript levels and disease. This area of research seems to be ripe for further methodological development and the use of transcriptional profile data for complex trait gene discovery seems to be promising. We have shown that this approach can now be used on an epidemiological scale for the rapid discovery of candidate genes that are involved in complex diseases.

METHODS

Study population. This study was based on participants in the SAFHS[24], which focuses on genetic aspects of cardiovascular health and disease in Mexican Americans. Individual families were ascertained through a single adult proband from the Mexican American community in San Antonio, Texas, independent of any specific phenotype, but the enrollment criteria enriched for large, multigenerational families with many relatives living in the local area. More than 1,400 individuals were recruited in total. The stated pedigree relationships were verified using the PREST computer program[44], based on the available genotypes at autosomal markers (see below), and appropriate corrections to intrafamilial relationships were made if necessary. All study participants provided informed consent. The study and all protocols presented here were approved by the Institutional Review Board at the University of Texas Health Science Center at San Antonio.

Isolation of lymphocytes from fresh blood. For this investigation, frozen lymphocyte samples were available for 1,280 individuals from their first visit to the clinic, carried out between 1991 and 1995. Blood samples were obtained in the morning after an overnight fast and collected in an EDTA tube. Lymphocytes were isolated from a 10-ml sample using Histopaque (Sigma Chemical Co.), following the suggested protocol of the manufacturer. The isolated and washed lymphocytes were frozen in 1 ml RPMI-C containing 30% FBS and 10% DMSO, stored at –80 °C overnight and then transferred into liquid nitrogen tanks.

Isolation of total RNA. Total RNA was isolated from 1,280 lymphocyte samples using a modified procedure of the QIAGEN RNeasy 96 protocol for isolating total RNA from animal cells using spin technology (QIAGEN Inc.). Total RNA yield (μg) and purity (260 nm:280 nm) were determined spectrophotometrically using the NanoDrop ND-1000 (Wilmington). The integrity of the resuspended total RNA was determined using the RNA 6000 Nano Assay Chip Kit on the Bioanalyzer 2100 and the 2100 Expert software (Agilent Technologies). A total of 500 ng total RNA was dried down using an Eppendorf Vacufuge Concentrator 5301 (Eppendorf) and stored at –20 °C.

Anti-sense RNA synthesis, amplification and purification. aRNA was synthesized, amplified and purified using the Ambion MessageAmp II Amplification Kit (Ambion) following the Illumina Sentrix Array Matrix 96-well expression protocol (Illumina Inc.). Synthesized cDNA samples were purified using QIAGEN's QIAquick 96 PCR purification supplementary protocol for spin technology (QIAGEN document QQ01.doc, October 2001). Biotin-16-UTP (Roche) labeled aRNA was synthesized using Ambion's proprietary MEGAscript *in vitro* transcription (IVT) technology and T7 RNA Polymerase. Purification of aRNA samples was carried out using QIAGEN's RNeasy 96 protocol for RNA cleanup using spin technology. Anti-sense RNA total yield (μg) and purity (260 nm:280 nm) were determined spectrophotometrically using the NanoDrop ND-1000 and a total of 1.5 μg aRNA was dried and stored at –20 °C before sample hybridization.

Sample hybridization to Illumina BeadChip. Hybridization of aRNA to Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChips and subsequent washing, blocking and detection were carried out using Illumina's BeadChip 6 × 2 protocol.

Sample scanning and detection. Samples were scanned on the Illumina BeadArray 500GX Reader using Illumina BeadScan image data acquisition software (version 2.3.0.13). Illumina BeadStudio software (version 1.5.0.34) was used for preliminary data analysis, with a standard background normalization, to generate an output file for statistical analysis. To assess quality metrics of each day's run, several quality control procedures were implemented. A total RNA control sample, supplied in the Ambion MessageAmpII Amplification Kit, was analyzed with each daily run. The Illumina BeadStudio software was used to view control summary reports, scatter plots of the Ambion total RNA control results from different days and scatter plots of daily run samples. The scatter plots compared control with control or sample with sample and calculated a correlation coefficient. Viewing the scatter plots determined whether controls across different days varied in quality, indicating a reduction in assay performance, and highlighted those samples that were of lesser quality. The control summary report is generated by the BeadStudio software, which evaluates the performance of the built-in controls of the BeadChips across a particular day's runs. This allows the user to look for variations in signal intensity, hybridization signal, background signal and the background to noise ratio for all samples analyzed that day.

Identification of expressed transcripts. To identify transcripts that had detectable quantitative expression in lymphocytes, the distribution of expression values for a given transcript was compared with the distribution of the expression values of the controls that are imbedded in each chip. For each transcript, we carried out a χ^2 tail test to establish whether there was a significant excess of samples with values above the ninety-fifth percentile of the control null distribution. This test was used because it allows the detection of even those transcripts that are clearly present above baseline levels in only a subset of individuals, while not being detectable above baseline levels in most individuals. Using an FDR of 0.05, we identified 20,413 transcripts that had significant expression according to this criterion.

Standardization of expression values. To minimize the influence of overall signal levels, which may reflect RNA quantity and quality rather than a true biological difference between individuals, abundance values of all 20,413 retained transcripts were first standardized by z-scoring within individuals (using decile percentage bins of transcripts, grouped by average log-transformed raw signals across individuals), followed by linear regression against the individual-specific average log-transformed raw signal and its squared value. For each transcript, we directly normalized these residual expression scores by using an inverse Gaussian transformation across individuals, to ensure that the assumptions underlying the variance components-based analyses were not violated. This conservative procedure results in normalized expression phenotypes that are comparable between individuals and across transcripts.

Resequencing. We resequenced ~2 kb upstream of the transcriptional start site of the *VNN1* gene in 96 unrelated founder individuals (providing ~90% power to detect a SNP with a minor allele frequency of 0.006). Sequencing primers were designed to cover the region at an average of 750 bp per fragment. PCR amplification of these segments was carried out using standard conditions. PCR amplicons were then purified using ExoSap-IT (USB) and used as templates in cycle-sequencing reactions using the BigDye Terminator v3.1 Cycle Sequencing Ready Reaction Kit (Applied Biosystems), according to the manufacturer's instructions. Cycle-sequencing reactions were carried out on both the sense and anti-sense strands and purified on 384-well filter plates (Edge Biosystems) to remove unincorporated BigDye and oligonucleotides. Capillary electrophoretic separation of DNA sequence fragments was carried out on an Applied Biosystems 3730 DNA Analyzer. Base calling quality assessment and comprehensive sequence alignment for polymorphism identification was carried out using Applied Biosystems' SeqScape analysis software (version 2.5.0).

SNP genotyping. Six *VNN1* promoter SNPs (G–708A, G–667A, T–623C, C–612A, A–587G and G–137T) were genotyped by resequencing an 871-bp amplicon, using the Applied Biosystems' cycle-sequencing platform described above.

Marker genotypes. The Human MapPairs Genome-Wide Screening Set Version 6 and 8 from Research Genetics was used. Genotyping occurred according to the manufacturer's instructions by PCR on lymphocyte-derived DNA samples from study participants. The PCR products were subsequently pooled into multiplex panels for genotype calling on an automated DNA sequencer (model 377 with Genescan 672 and Genotyper software programs; Applied Biosystems Inc.). Overall, 1,345 individuals were genotyped at 432 highly polymorphic microsatellite markers, distributed with an average inter-marker spacing of <10 cM across all 22 autosomes. The median distance between the detected autosomal transcripts and the nearest marker was ~2.7 cM, and 80% and 90% of transcripts were within 4.3 cM and 5.5 cM of the nearest marker, respectively. These genotypes were subjected to extensive data cleaning. Using the SimWalk2 software package[45], we estimated the probability that the genotype at a given marker locus in a given individual is erroneous[46], using all marker loci jointly. The computation was based on maximum likelihood marker allele frequencies[47] and deCODE genetics' genetic map[30]. This statistical procedure is designed to detect inconsistencies and unlikely genotypes. An iterative process was followed to eliminate genotypes that are likely to be erroneous until no more inconsistencies or possible errors remained. Overall, 1.4% of genotypes were discarded.

Statistical genetic analysis: Heritability, linkage, and association. All statistical analyses on related individuals were carried out using variance components-based methodology and the SOLAR v. 4.0 software package[29]. Heritability analysis was carried out under the classical approach that deconstructs the phenotypic variance into independent genetic and environmental components, assuming an additive model of gene action ('narrow sense' heritability) and expected kinship coefficients that is based on the observed intra-familial relationships. As preparation for linkage analysis, the probabilities of multipoint identity-by-descent allele sharing among pairs of related individuals were computed by the Monte Carlo Markov Chain multipoint approach implemented in the Loki software package[28], using the genotypes at all linked markers jointly in the computations. Maximum likelihood marker allele frequencies[47] were used, and the genetic map was based on the one developed by deCODE genetics[30], using an interpolation procedure that is based on physical locations to place markers that were genotyped in the SAFHS but were not on deCODE genetics' map. Covariates were included in the variance components framework as linear predictors of phenotype. Measured genotype analysis[48], embedded in a variance components-based linkage model, was used for association testing, assuming an additive model of allelic effect (in other words, the SNP genotypes AA, AB and BB were coded as –1, 0 and 1, respectively, and used as a linear predictor of phenotype)[35].

In our data set, the power to detect heritability at a significance level of 0.05 is approximately 80/90/95% for $h^2 = \sim 0.10/\sim 0.12/\sim 0.14$. At the same pointwise significance level, corresponding to a lod score of 0.588, the power to detect linkage is 80/90/95% for QTL effect sizes of approximately 10/12/14%, assuming a heritability of 22.5% (which is equal to the median estimate observed among all expression phenotypes).

HDL-C assay. HDL-C concentrations were measured in plasma obtained from peripheral blood drawn after an overnight fast. The assay protocol has been published[24]. Briefly, dextran sulfate-Mg^{2+} was used to precipitate apoB-containing particles from thawed plasma[49], and a commercial assay (Boehringer-Mannheim Diagnostics) was used to assay HDL-C levels.

Electrophoretic mobility shift assay (EMSA). The human T-cell derived Jurkat E6-1 (TIB-152) cell line obtained from the American Type Culture Collection was maintained in RPMI-1640, supplemented with 2 mM L-glutamine, 100 mg/ml each of streptomycin and penicillin, and 10% FBS, at 37 °C with 5% CO2. Complementary 5′ biotinylated oligonucleotides representing both allelic forms of the *VNN1* –137 promoter SNP under investigation were obtained commercially (Operon Biotechnologies) (**Supplementary Table 5b** online). Nuclear extracts were prepared from approximately 8×10^7 cells according to the method used in ref. 50. Extracts were frozen in liquid nitrogen and stored at –80 °C. The protein concentrations of extract preparations were determined using the Bio-Rad protein assay kit. For EMSA, nuclear proteins (3 μg) were pre-incubated for 10 min on ice with 1 μg of poly (dI-dC) (Pharmacia) in a binding buffer (4% Ficoll, 20 mM HEPES (pH 7.9), 1 mM EDTA, 1 mM DTT, 50 mM KCl) to give a final reaction volume of 20 ml. Nuclear proteins were then incubated with biotin-labeled double-stranded oligonucleotide probes (~20 fmol) for 30 min on ice and then analyzed on a 6% Novex DNA retardation gel (Invitrogen), and electroblotted onto a positively charged nylon membrane. Detection of protein–DNA complexes was achieved after incubation of the membrane with streptavidin-horseradish peroxidase and development with luminol substrate (Pierce; Lightshift Chemiluminescent EMSA kit). Light emission was captured on X-ray film.

URLs. Heritability estimates, *cis* lod scores, maximum *trans* lod scores and miscellaneous other information on all analyzed transcripts are available from the Southwest Foundation for Biomedical Research genetics website (http://www.sfbr.org/pages/genetics_SAFHS-transcriptional-profiling.php). SOLAR is available from http://www.sfbr.org/solar.

Accession numbers. ArrayExpress: raw expression values (of all transcripts on the microarray) and normalized expression values (of all 19,648 analyzed autosomal transcripts), along with information on sex, age and HDL-C concentration, are available under accession number E-TABM-305.

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

ACKNOWLEDGMENTS

We are grateful to the participants in the San Antonio Family Heart Study. Data collection was supported by a grant from the US National Institute for Heart, Lungs and Blood (HL045222). A donation from the Azar and Shepperd families paid for the transcriptional profiling. Additional funds for transcriptional profiling, sequencing, genotyping and statistical analysis were provided by ChemGenex Pharmaceuticals. The SOLAR statistical genetics computer package is supported by a grant from the US National Institute of Mental Health (MH059490). The supercomputing facilities used for this work at the AT&T Genetics Computing Center were supported in part by a gift from the SBC Foundation. The laboratory work was carried out in facilities that were constructed with support from the US National Center for Research Resources (RR013556).

AUTHOR CONTRIBUTIONS

J.B., E.K.M. and G.R.C. initiated the study. H.H.H.G. and J.B. performed or supervised all aspects of the statistical analysis and were aided by T.D.D. and J.C. E.K.M., J.E.C. and L.J.A. were responsible for all molecular analyses, including transcriptional profiles, resequencing, SNP typing and functional analysis, with aid from M.P.J. J.B., J.W.M., M.C.M., A.G.C., and L.A. were responsible for the Mexican American family samples. S.A.C. and J.B. were responsible for the 10-cM STR typing. D.L.R. was responsible for the HDL-C measurement. J.B.M.J. and J.C. performed bioinformatic analyses. A.H.K. and G.R.C. provided additional biological interpretation.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

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

1. Jansen, R.C. & Nap, J.P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
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. *Nat. Genet.* **35**, 57–64 (2003).
4. Storey, J.D., Akey, J.M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* **3**, e267 (2005).
5. 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).
6. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
7. Kirst, M. *et al.* Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiol.* **135**, 2368–2378 (2004).
8. Wayne, M.L. & McIntyre, L.M. Combining mapping and arraying: An approach to candidate gene identification. *Proc. Natl. Acad. Sci. USA* **99**, 14903–14906 (2002).
9. Bystrykh, L. *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.* **37**, 225–232 (2005).
10. Chesler, E.J. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* **37**, 233–242 (2005).
11. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**, 243–253 (2005).
12. Dausset, J. *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990).
13. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
14. Monks, S.A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
15. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
16. Stranger, B.E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
17. Deutsch, S. *et al.* Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. *Hum. Mol. Genet.* **14**, 3741–3749 (2005).
18. Pastinen, T., Ge, B. & Hudson, T.J. Influence of human genome polymorphism on gene expression. *Hum. Mol. Genet.* **15**, R9–R16 (2006).
19. de Koning, D.J. & Haley, C.S. Genetical genomics in humans and model organisms. *Trends Genet.* **21**, 377–381 (2005).
20. Pant, P.V. *et al.* Analysis of allelic differential expression in human white blood cells. *Genome Res.* **16**, 331–339 (2006).
21. Whitney, A.R. *et al.* Individuality and variation in gene expression patterns in human blood. *Proc. Natl. Acad. Sci. USA* **100**, 1896–1901 (2003).
22. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
23. Storey, J.D. *et al.* Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80**, 502–509 (2007).
24. Mitchell, B.D. *et al.* Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. *Circulation* **94**, 2159–2170 (1996).
25. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
26. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and poerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
27. Jin, W. *et al.* The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster. Nat. Genet.* **29**, 389–395 (2001).
28. Heath, S.C. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.* **61**, 748–760 (1997).
29. Almasy, L. & Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**, 1198–1211 (1998).
30. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
31. Göring, H.H., Terwilliger, J.D. & Blangero, J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* **69**, 1357–1369 (2001).
32. Williams, J.T. & Blangero, J. Power of variance component linkage analysis to detect quantitative trait loci. *Ann. Hum. Genet.* **63**, 545–563 (1999).
33. Young, C.E., Karas, R.H. & Kuvin, J.T. High-density lipoprotein cholesterol and coronary heart disease. *Cardiol. Rev.* **12**, 107–119 (2004).
34. Yamazaki, K., Kuromitsu, J. & Tanaka, I. Microarray analysis of gene expression changes in mouse liver induced by peroxisome proliferator- activated receptor alpha agonists. *Biochem. Biophys. Res. Commun.* **290**, 1114–1122 (2002).
35. Blangero, J. *et al.* Quantitative trait nucleotide analysis using Bayesian model selection. *Hum. Biol.* **77**, 541–559 (2005).
36. Kent, J.W. Jr., Dyer, T.D., Göring, H.H.H. & Blangero, J. Type I error rates in association versus joint linkage/association tests in related individuals. *Genet. Epidemiol.* **31**, 173–177 (2007).
37. Chekmenev, D.S., Haid, C. & Kel, A.E. P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.* **33**, W432–W437 (2005).
38. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
39. Kadonaga, J.T., Carner, K.R., Masiarz, F.R. & Tjian, R. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. *Cell* **51**, 1079–1090 (1987).
40. Turner, J. & Crossley, M. Mammalian Kruppel-like transcription factors: more than just a pretty finger. *Trends Biochem. Sci.* **24**, 236–240 (1999).
41. Kaczynski, J., Cook, T. & Urrutia, R. Sp1- and Kruppel-like transcription factors. *Genome Biol.* **4**, 206 (2003).
42. Perez-Enciso, M. In silico study of transcriptome genetic variation in outbred populations. *Genetics* **166**, 547–554 (2004).
43. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
44. McPeek, M.S. & Sun, L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**, 1076–1094 (2000).
45. Sobel, E. & Lange, K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.* **58**, 1323–1337 (1996).
46. Sobel, E., Papp, J.C. & Lange, K. Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* **70**, 496–508 (2002).
47. Boehnke, M. Allele frequency estimation from data on relatives. *Am. J. Hum. Genet.* **48**, 22–25 (1991).
48. Boerwinkle, E., Chakraborty, R. & Sing, C.F. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann. Hum. Genet.* **50**, 181–194 (1986).
49. Warnick, G.R., Benderson, J. & Albers, J.J. Dextran sulfate-Mg2+ precipitation procedure for quantitation of high-density-lipoprotein cholesterol. *Clin. Chem.* **28**, 1379–1388 (1982).
50. Li, Y.C., Ross, J., Scheppler, J.A. & Franza, B.R. Jr. An *in vitro* transcription analysis of early responses of the human immunodeficiency virus type 1 long terminal repeat to different transcriptional activators. *Mol. Cell. Biol.* **11**, 1883–1893 (1991).