Survey of allelic expression using EST mining

Bing Ge, Scott Gurd, Tiffany Gaudin, Carole Dore, Pierre Lepage, Eef Harmsen, Thomas J. Hudson, and Tomi Pastinen[1]

McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A, Canada

Cis-acting allelic variation in gene regulation is a source of phenotypic variation. Consequently, recent studies have experimentally screened human genes in an attempt to initiate a catalog of genes possessing _cis_-acting variants. In this study, we use human EST data in dbEST as the source of allelic expression data, and the HapMap database to provide expected allele frequencies in human populations. We demonstrate a greater concordance of allele frequencies estimated from human ESTs in dbEST with those derived from the CEPH HapMap sample representing Caucasians from northern and western Europe, than population samples obtained in Asia and Africa. Deviations between allele frequencies observed in EST databases and the ones obtained from the CEPH HapMap samples may result from common heritable _cis_-acting variants altering the relative allele distribution in RNA. We provide in silico as well as experimental evidence that this strategy does allow significant enrichment of genes harboring common heritable _cis_-acting polymorphisms in linkage disequilibrium with expressed alleles.

[Supplemental material is available online at www.genome.org.]

Progress in human disease gene discovery and characterization of global expression patterns in unrelated individuals suggest that genetic variation affecting gene expression has significant effects on phenotypic variability, including risk to disease. *Cis*-acting variation that regulates expression leads to preferential expression of an allelic transcript. This can be detected directly by in vivo comparisons of the relative abundancies of allelic transcripts using intragenic polymorphisms (Pastinen and Hudson 2004). Such surveys of allelic expression by us and others (Yan et al. 2002; Bray et al. 2003; Lo et al. 2003; Pastinen et al. 2004) have established that unequal representation of marker alleles in transcripts from heterozygous individuals is a common phenomenon, affecting 5%–20% of heterozygous individuals for 15%–50% of genes. These studies did not systematically address the possible tissue specificity of allelic expression that was observed in mice (Cowles et al. 2002). Allelic expression analysis may also detect epigenetic regulation including classical imprinting and random monoallelic expression (Pastinen and Hudson 2004).

Increased attention to regulatory polymorphisms underlying human traits (for review, see Knight 2005) calls for cataloging of genes demonstrating differences in allelic expression. Data regarding common *cis*-acting regulatory variants and information on tissue specificity and the regulatory mechanism(s) that modify allelic imbalance would optimally be included in such a resource.

The largest public source of sequence variation in human transcripts that exists is in dbEST and contains 4,636,789 sequence traces (build #173) in UniGene clusters. Along with EST-based gene identification (Adams et al. 1992; for review, see Mathe et al. 2002), this database has found uses in identifying single nucleotide polymorphisms (SNPs) (Buetow et al. 1999; Marth et al. 1999; Irizarry et al. 2000). Recently, a method to look for evidence of imprinting using EST data was suggested (Yang et al. 2003). Here we propose to use the dbEST data along with the recently released Human Haplotype Map Project data (The International HapMap Consortium 2003) for large-scale discovery of genes harboring common regulatory variants. In addition to the analysis of ~11,000 SNPs for "in silico allelic imbalance," we present corroborating bioinformatics and experimental evidence as well as new tools for experimental validation.

Results and Discussion

If a common *cis*-acting regulatory variant (or regulatory SNP, rSNP) is in linkage disequilibrium (LD) with a marker SNP found in redundant ESTs, the marker allele corresponding to the underexpressed allele should be underrepresented in EST allele counts. In order to detect such occurrences, the expected EST allele distribution needs to be estimated. Figure 1 illustrates an in silico approach to detecting deviations between the observed versus expected allele counts in EST data sets.

The major caveats in this approach relate to the quality of EST sequencing, estimating allele frequencies in EST data sets when the ethnic origin of the donor is unknown, and biased sampling. To handle these issues, we made several assumptions. The impact of errors in EST sequencing is alleviated by focusing on SNPs that have been validated independently: We consider the demonstration of Mendelian transmission as an ultimate proof that an SNP is valid, and only SNPs showing evidence of transmission in Caucasian HapMap trios were further studied. Because each cDNA library can maximally contribute two alleles in the EST allele counts, we randomly selected a maximum of two sequences from each library. This reduces the sampling bias attributed to unequal sequencing density of individual libraries. The most difficult issue is the lack of information about the origin of the donor. Based on the location of the largest EST sequencing projects, we assume that Caucasian alleles are overrepresented in the EST libraries contained in dbEST. This hypothesis was tested by comparing the observed EST allele counts versus expected allele frequencies (population allele frequencies) in four populations across >2500 loci (Fig. 2). The Caucasian allele frequencies fit relatively well with the allele frequencies observed in ESTs, whereas the allele frequencies in other populations generally deviate considerably. We therefore used only Caucasian al-

[1] **Corresponding author.**
E-mail tomi.pastinen1@staff.mcgill.ca; fax (514) 398-2622.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.4023805. Freely available online through the *Genome Research* Immediate Open Access option.

15:1584–1591 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

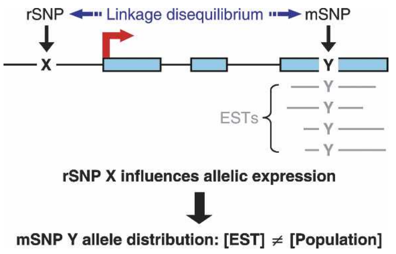
rSNP
Linkage disequilibrium
mSNP
X
Y
ESTs
Y
Y
Y
Y
rSNP X influences allelic expression
mSNP Y allele distribution: [EST] ≠ [Population]

Figure 1. Principle of EST–genotype comparisons. If a marker SNP (mSNP) in observed ESTs is in linkage disequilibrium with an unknown, *cis*-acting, regulatory SNP (rSNP) affecting the allelic transcription of the corresponding transcript, the allelic distribution in ESTs deviates from the allele frequency observed in the population (genomic DNA).

lele frequencies to estimate the expected allele frequencies in further steps. We cannot correct for the possibility that some genes may be preferentially detected in EST libraries derived from few non-Caucasian donors, nor can we adjust for multiple libraries having been derived from the same individual. The normalization (Soares et al. 1994) and subtraction (Bonaldo et al. 1996) procedures in EST library construction to improve coverage of EST sequencing should not have a major impact on the allelic representation of transcripts, since these methods generally use long probes that are usually not affected by minor sequence differences such as SNPs. In contrast, the power to detect allele-specific rare transcript variants in our method may actually be enhanced by the normalization/subtraction used in constructing a subset of human EST libraries. Finally, we note that the power to detect deviations from the expected frequencies differs from gene to gene, since the representation of the transcripts varies in different cDNA libraries.

Table 1 lists the top 117 SNPs (nominal $p < 0.001$) showing the greatest deviation between CEPH allele frequencies derived from HapMap (http://www.HapMap.org release#13; The International HapMap Consortium 2003), and allele frequencies derived from polymorphic ESTs mapping to human RefSeq genes. This list of in silico allelic imbalances pinpoints genes that have an increased likelihood for harboring regulatory SNPs. Most candidate SNPs are derived from a wide range of tissues, arguing for tissue independency of detected effects. Detailed description of SNP–EST comparison for the SNPs included in Table 1 as well as the complete data set can be accessed at http://genomequebec.mcgill.ca/EST-HapMap.

In order to detect *cis*-acting differences using the EST data, the marker SNP has to be in LD with the unknown regulatory SNP. Most of the known or suggested variants affecting human gene regulation map to the 5′-end of transcripts (Rockman and Wray 2002). In addition, the highest density of validated transcriptional control sequences map to 5′-ends of mammalian genes (Xie et al. 2005). Given that LD is highly dependent on physical distance (Reich et al. 2001), we hypothesized that, on average, the EST–SNPs showing significant deviation would show a tendency to map closer to 5′-ends of genes. To test this, we compared the significant hits to a matched set of SNPs in ESTs (to ensure similar sample size and power) showing no evidence of deviation in distribution between genomic (CEPH) and EST-derived allele frequencies. The results of these comparisons are shown in Figure 3. On average, our best hits ($n = 117$ at $P < 0.001$) mapped 22 kb closer (29 vs. 51 kb) to the 5′-ends of the respective genes as compared to nonsignificant, matched data points (2-tailed *t*-test, $P = 0.005$ for difference between the groups). Extending the comparison to hits at $P < 0.01$ ($n = 358$) still shows significant difference (*t*-test, $P = 0.006$) in the distance from transcript start between the matched groups (37 vs. 50 kb), whereas the effect is no longer statistically significant with SNPs showing a weaker association (Fig. 3).

For experimental validation of the predicted differences between expressed alleles, we chose to use direct sequencing, which allows normalization of polymorphic bases relative to the surrounding invariant bases and other quality measures that are not possible with most genotyping methods. To facilitate the analysis of allelic expression from these sequence traces, we developed the software PeakPicker, as illustrated in Figure 4. Genomic DNA and RNA (cDNA) samples are sequenced in parallel, with heterozygous genomic DNA samples serving to establish the expected range of 50:50 allelic ratios. We evaluated allelic imbalance (AI) using a sample-specific test (called AI_{95}) and a locus-specific test (called $\mathrm{AI}_{\mathrm{LS}}$). For the AI_{95} test, we estimated a 95% confidence interval (CI) for equal expression, based on the relative peak heights observed in heterozygous genomic DNA samples. The 95% CI threshold corresponds, on average, to a <1.2-fold difference between transcript abundance (or a 45:55 allelic ratio), as determined by dilution experiments (see Supplemental Fig. 1). Unequal expression of allelic transcripts in a heterozygous sample (i.e., AI) is called when we observe consistent deviations beyond the 95% CI in independent RNA preparations derived from the same LCL. If only one sample falls outside the 95% CI or if the replicate samples give conflicting deviations (suggesting unequal expression of opposite alleles), the allelic expression status of the sample is categorized as undefined. The determination

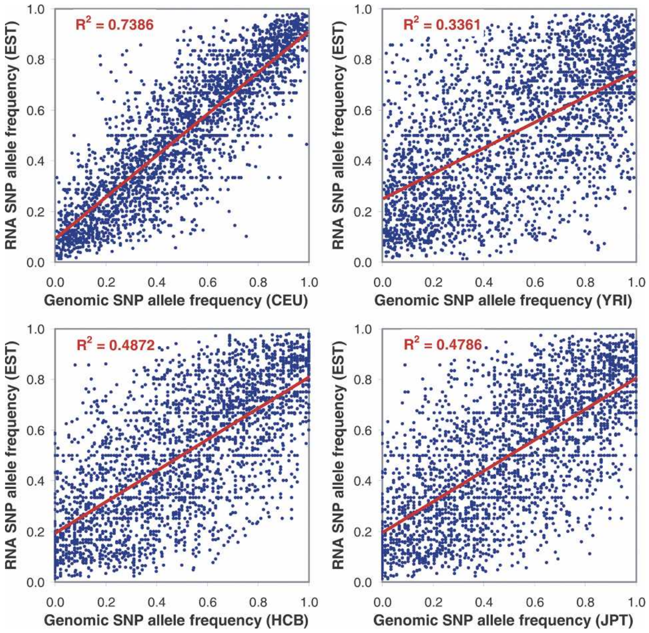

Figure 2. EST allele frequencies versus allele frequencies in different ethnic groups. Allele frequencies for 2678 SNPs genotyped in all HapMap populations (*x*-axis) were compared to EST allele-counting derived allele frequencies (*y*-axis). The Caucasian allele (CEU, *top left* graph) frequencies show relatively good concordance with those observed in the ESTs. The African allele frequencies (YRI, *top right*) deviate most from the EST-derived frequencies, whereas the allele frequencies in Asian populations (HCB and JPT, *bottom left* and *right*, respectively) show similar—but significantly lower—concordance with allele frequencies observed in ESTs as compared to the comparisons of ESTs versus Caucasian allele frequencies.

Table 1. Top 117 SNPs in genotype (HapMap/CEU Nov. 2004)—EST (UniGene) comparison

Gene	SNP	CHR	*P*-value[a]	Gene	SNP	CHR	*P*-value	Gene	SNP	CHR	*P*-value
STEAP1	rs2888782	7	5.62E-19	*MAPK7*	rs2233072	17	6.68E-05	*HTATIP2*	rs3824886	11	4.48E-04
STEAP1	rs4015375	7	1.88E-17	*IFI27*	rs2799	14	6.80E-05	*ALS4*	rs2296869	9	4.61E-04
FCGR3A	rs396716	1	7.92E-12	*WASPIP*	rs7739	2	6.82E-05	*TXNDC13*	rs1058007	20	4.82E-04
ANKRD15	rs3739586	9	3.24E-10	*FLJ10287*	rs687513	1	7.65E-05	*RNASE3*	rs2073342	14	4.94E-04
WBSCR16	rs7375	7	1.07E-09	*ADPGK*	rs9460	15	9.38E-05	*CXCL16*	rs1051007	17	4.98E-04
DPYSL2	rs11863	8	1.51E-09	*TF*	rs1799852	3	9.46E-05	*HRG*	rs1042445	3	5.10E-04
PSMB4	rs7172	1	1.69E-09	*TM4SF4*	rs9793	3	1.07E-04	*PBK*	rs1052874	8	5.10E-04
AUP1	rs2231250	2	2.90E-09	*METAP1*	rs1238741	4	1.11E-04	*PAX8*	rs1478	2	5.30E-04
CORO1C	rs2111211	12	2.93E-08	*MRC1*	rs941	10	1.13E-04	*ACHE*	rs7636	7	5.30E-04
RBM33	rs6962201	7	1.40E-07	*DHX34*	rs1064202	19	1.13E-04	*RPAP1*	rs3743031	15	5.50E-04
TACC3	rs8389	4	5.38E-07	*SRP14*	rs7535	15	1.20E-04	*SLC35C2*	rs1044369	20	5.60E-04
TXNIP	rs7211	1	6.98E-07	*PURB*	rs9701	7	1.23E-04	*DOCK2*	rs3763048	5	6.20E-04
ACRBP	rs1045553	12	7.73E-07	*LIPC*	rs690	15	1.27E-04	*LOC51321*	rs3213690	17	6.40E-04
CYP3A5	rs15524	7	1.05E-06	*ANKRD13*	rs1044994	12	1.32E-04	*FVT1*	rs6810	18	6.40E-04
RUVBL2	rs1062708	19	2.27E-06	*SLC7A9*	rs2287884	19	1.46E-04	*IL17R*	rs2229151	22	6.40E-04
PSCA	rs1045605	8	3.29E-06	*GRWD1*	rs1643487	19	1.62E-04	*ADAT1*	rs3743598	16	6.50E-04
PLK1	rs27770	16	4.06E-06	*WBSCR18*	rs8891	7	1.73E-04	*SLC20A2*	rs6841	8	6.80E-04
EMID1	rs743920	22	4.99E-06	*COASY*	rs615942	17	1.75E-04	*HADHB*	rs1056471	2	6.80E-04
SPATA5L1	rs1365610	15	7.28E-06	*SNX6*	rs9264	14	1.81E-04	*PRL*	rs6239	6	6.90E-04
GCS1	rs1063588	2	8.51E-06	*KBTBD8*	rs7623808	3	1.99E-04	*GABBR1*	rs2267633	6	6.90E-04
PSCA	rs1045574	8	8.69E-06	*C22orf5*	rs1059804	22	2.00E-04	*PTPN12*	rs3750050	7	7.00E-04
UBXD2	rs1050115	2	9.11E-06	*CXCL5*	rs425535	4	2.04E-04	*FLJ12788*	rs2301984	2	7.00E-04
VPS39	rs7086	15	1.04E-05	*DMTF1*	rs3747807	7	2.04E-04	*LOC133957*	rs13474	5	7.00E-04
C20orf111	rs9346	20	1.78E-05	*BBOX1*	rs2305095	11	2.31E-04	*QDPR*	rs3733570	4	7.30E-04
GATM	rs1049508	15	1.83E-05	*GLT8D2*	rs3817602	12	2.37E-04	*NOB1P*	rs3811348	16	7.30E-04
GPX3	rs11548	5	1.86E-05	*ITGA3*	rs3744538	17	2.44E-04	*HCRTR1*	rs2271933	1	7.80E-04
SLC25A28	rs880568	10	1.88E-05	*PPIL5*	rs2281836	14	2.48E-04	*FCGBP*	rs741143	19	7.90E-04
SURF2	rs12763	9	2.30E-05	*ISG20L2*	rs3795737	1	2.72E-04	*OAS1*	rs2660	12	8.00E-04
BTBD1	rs1045742	15	2.61E-05	*FLJ12681*	rs3751667	16	2.82E-04	*GRN*	rs5848	17	8.10E-04
MOBK1B	rs2272239	2	3.25E-05	*NMNAT2*	rs549191	1	2.94E-04	*KLP1*	rs7478	19	8.30E-04
PAH	rs1042503	12	3.31E-05	*CD200*	rs1050572	3	3.43E-04	*LOC133957*	rs1054174	5	8.30E-04
KIAA0020	rs12171	9	4.06E-05	*GEMIN6*	rs1056104	2	3.58E-04	*DEXI*	rs3087519	16	8.70E-04
CYB5R3	rs137124	22	4.07E-05	*RPUSD4*	rs2276312	11	3.63E-04	*TRAF3*	rs1131877	14	9.00E-04
TRIM4	rs2572010	7	4.39E-05	*ARTS-1*	rs27434	5	3.81E-04	*SEC61A1*	rs1042907	3	9.10E-04
FBLN1	rs9682	22	4.69E-05	*CYP3A7*	rs10211	7	3.94E-04	*PRSS35*	rs3812141	6	9.20E-04
ARNT2	rs7484	15	5.38E-05	*CD1E*	rs1065457	1	4.02E-04	*TRUB2*	rs2231645	9	9.60E-04
PSCA	rs2976396	8	5.50E-05	*RUNX3*	rs13157	1	4.08E-04	*NUMA1*	rs3750912	11	9.70E-04
CCDC3	rs2280076	10	6.07E-05	*SERPINA3*	rs6116	14	4.09E-04	*ERAL1*	rs2242345	17	9.90E-04
MYBL2	rs285162	20	6.10E-05	*DLAT*	rs9371	11	4.12E-04	*BACE1*	rs638405	11	1.00E-03

[a] *P*-value for difference in comparison of EST- and CEPH-derived allele frequencies.

of allelic expression status in individual samples using the AI_{95} test is useful if further mapping studies are pursued. Locus-specific allelic imbalance is defined using the AI_{LS} test, which is a two-tailed *t*-test comparing the average peak ratios observed in genomic DNA and RNA (cDNA). The advantages of the AI_{LS} test is that all data points (including discordant results that were discarded by the AI_{95} test) are used; the test appears to offer a small improvement in sensitivity for allelic expression differences, but it does not specify the AI status of individual samples.

From the set of 976 genes having in silico allelic imbalance ($P < 0.05$) (http://genomequebec.mcgill.ca/EST-HapMap), we chose 40 genes that are expressed in lymphoblastoid cell lines (LCLs), based on earlier DNA-microarray studies (Pastinen and Hudson 2004). The selection is not random, as we picked genes over many months when the HapMap data were being produced; the final set is biased toward genes that we observe to have more significant differences between CEPH and EST allele frequencies. The LCL panel for this validation test included the unrelated CEPH parents used in the International HapMap project (The International HapMap Consortium 2003). In a limited number of cases in which both parents were uninformative, allelic imbalance was measured in LCLs from their offspring. The results of DNA/RNA sequencing are shown in Table 2. One assay could not be confidently interpreted (*SNX6*) (see Table 2, footnote j) and was excluded. Using the AI_{95} method to evaluate each heterozygous sample, we observed a statistically significant ($\chi^2 = 9.0$, 1df, empirical $P < 0.005$) enrichment of genes with LCLs having allelic imbalance, as predicted by the EST-genotype comparison: 14 of 39 genes (36%) predominantly overexpress the predicted allele (with >80% of heterozygous samples called showing the predicted allele being overexpressed), and only two (5%) showed predominant deviation toward the opposite allele. Similar results were observed by the locus-specific AI_{LS} method to determine allelic imbalance (see Table 2, two rightmost columns and footnotes h and i), with 15 genes showing statistically significant overexpression of the predicted allele, whereas two genes showed significant overexpression of the opposite allele; again, the distribution of the results deviates significantly from chance ($\chi^2 = 9.9$, 1df, empirical $P < 0.005$). The same genes were generally identified with both the AI_{95} and AI_{LS} methods of analysis. Overall, using the AI_{95} test, the predicted allele was overexpressed in 242 informative cases, the alleles were equally expressed or labeled undefined in 348 cases, and the opposite allele showed evidence of overexpression in 105 cases. This distribution deviates highly significantly ($\chi^2 = 54.1$, 1df, $P < 0.0001$) from chance. We note that the thresholds to call allelic expres-

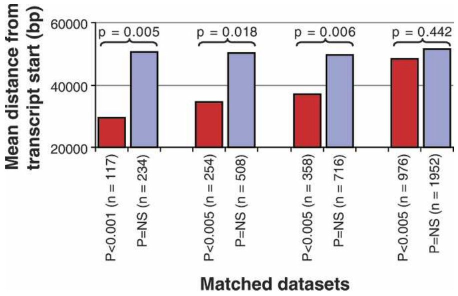

Figure 3. Comparison of SNP-transcript start distances based on significance in EST–genomic allele frequency comparisons. Four pairs of matched (i.e., equal allele frequencies and EST counts) groups of SNPs were compared for distance from transcript start site in the genomic sequence as a proxy for linkage disequilibrium from a putative "regulatory SNP." The *leftmost* pair of bars shows mean transcript start site distances for the test set of 117 most significantly ($P < 0.001$) deviating sets of SNPs (*leftmost* red bar; also see Table 1) as compared to the control set of 234 matched SNPs showing nonsignificant deviation in EST–genomic allele frequency comparison (*leftmost* blue bar). The mean distance from the transcript start for the test set is ~29 kb, which is significantly shorter (*t*-test, $P = 0.005$) than the average distance (~51 kb) observed in the control set. The pairs of bars to the *right* of the first pair show progressively larger, and less significantly associated, test sets as compared to the two matched sets of controls. The difference in distance between test and control sets diminishes in a stepwise manner: For the test set at $P < 0.005$, the distance is 34 kb versus 50 kb observed in controls; at $P < 0.01$, the distance is 37 kb versus 50 kb; and at $P < 0.05$, the distance is 48 kb versus 52 kb, which is no longer significant.

sion using the AI_{95} test can influence the total number of allelic imbalances called. For genes that are expressed at widely different expression levels (and show more variability in allele ratios in cDNA samples), this may inflate the total number of RNA samples demonstrating putative allelic expression. The potential for false-positive assignments of allelic imbalance due to too nonstringent threshold definition would be expected to work as a random confounder and should not bias the assessment of the direction of relative expression differences between allelic transcripts. Similarly, allelic expression is quantitatively deviated toward the predicted allele across the tested 39 loci: The average predicted high allele versus predicted low allele ratio ($H_{\mathrm{DNA}}/L_{\mathrm{DNA}}$ ratio) was 0.96 (95% CI: 0.91–1.02) in control heterozygous genomic DNA samples, while in RNA the average $H_{\mathrm{RNA}}/L_{\mathrm{RNA}}$ ratio was 1.21 (95% CI: 1.12–1.30). The difference between distribution of allele ratios in genomic DNA and RNA (cDNA) is highly significant ($P = 0.0000011$, two-tailed *t*-test).

By combining the dbEST and International HapMap data, we generated a list enriched for candidate genes exhibiting allelic expression differences. These candidates differ from those derived from experimental surveys (Yan et al. 2002; Bray et al. 2003; Lo et al. 2003; Pastinen et al. 2004) in two important aspects: (1) the expression data is derived from multiple tissues; and (2) the significantly deviating marker SNPs should be in LD with common regulatory variants in order to be detected in this screening approach. The rate of validation of genes (36% and 38% using the AI_{95} and AI_{LS} methods, respectively) with expected allelic imbalance is not different from the 15%–50% of genes reported to demonstrate preferential allelic expression in unselected genes. However, we note that in comparison to previous experiments using genes not predicted by EST analyses, the genes validated here have a distinguishing characteristic: The allelic expression is common and shows consistent bias in regard to which allele is overexpressed (as opposed to showing variability among samples from different individuals). In our earlier experimental survey, <5% of randomly selected genes demonstrated such characteristics (Pastinen et al. 2004). Finally, we acknowledge that our validation strategy is only moderately powered to detect weak or rare regulatory variants and tissue-specific effects. The candidate SNPs identified here are therefore optimal for investigations of regulatory variants in case-control studies. In fact, some of the genes validated in this study have already been associated with complex disease traits in earlier studies: *EPHX2*, *OAS1*, and *ARTS-1* have been suggested to influence lipid phenotypes, susceptibility to diabetes, or development of hypertension, respectively (Yamamoto et al. 2002; Sato et al. 2004; Field et al. 2005).

The candidate gene list provided here offers a relatively straightforward approach for discovering regulatory variants in other tissues as well. We also note that the data used in this study are based on the HapMap release of November 2004, which may represent only 15%–20% of the final data to be included in the International HapMap project (The International HapMap Consortium 2003). In addition to data presented here, we have established a Web resource containing updated lists of candidate genes based on new HapMap data releases (http://genomequebec.mcgill.ca/EST-HapMap). High-sensitivity allelic expression analysis is facilitated by the normalized sequencing approach and the PeakPicker software. Demonstration of preferential expression of predicted allele indicates that the regulatory variants are in LD with the tested alleles and thus allow delineation of the "search space" for regulatory variants to a region delimited by ancestral recombination events, or to "haplotype blocks" (Daly et al. 2001).

Methods

Samples and RNA preparation for allelic expression assays

A description of the lymphoblastoid cell lines used in this study is listed in Supplemental Table 1. The RNA from cultured cells used in allelic expression assays consisted of 60 unrelated lymphoblastoid cell lines (LCLs) of Caucasian origin, corresponding to the parents of the CEPH trios used in the International HapMap project (The International HapMap Consortium 2003). Four children from these 30 trios were also included in the analysis of chromosomal association of the allelic expression in cases in which the parents were uninformative or failed. The lymphoblastoid cell lines were obtained from the Coriell Cell Repositories (Coriell Institute for Medical Research, Camden, NJ). The LCLs were cultured in RPMI 1640 medium (Invitrogen) supplemented with penicillin/streptomycin, 2 mM L-glutamine, and 15% heat-inactivated fetal bovine serum (Sigma-Aldrich). The cells were grown at 37°C and 5% CO_2. The cell growth was monitored using a hemocytometer, and cultures were harvested at a density of $0.8–1.1 \times 10^6$ cells/mL. The cells were lysed by resuspension in TRIzol Reagent (Invitrogen).

cDNA synthesis

RNA was isolated from LCLs using Trizol reagent according to the manufacturer's instructions (Invitrogen). The RNA quality was verified by observing samples on the Agilent 2100 Bioanalyzer (Agilent). In the reaction, 50-µg aliquots of total RNA were treated with 8 U of DNase I for 40 min at 37°C (Ambion), extracted with phenol/chloroform (Invitrogen), and reprecipitated. The resulting RNA was annealed to 1000 ng of ran-

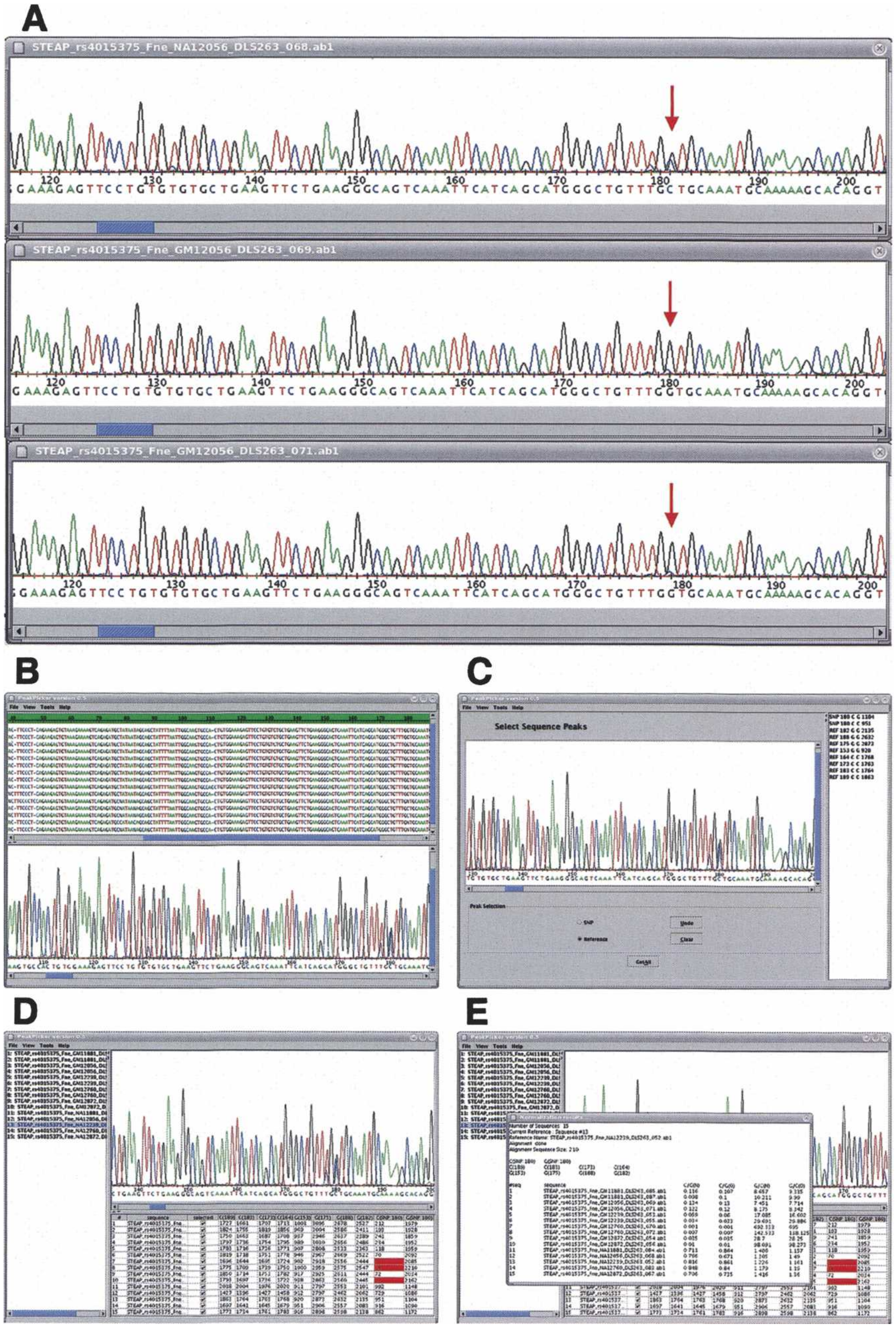
A
STEAP_rs4015375_Fne_NA12056_DLS263_068.ab1
STEAP_rs4015375_Fne_GM12056_DLS263_069.ab1
STEAP_rs4015375_Fne_GM12056_DLS263_071.ab1
120
130
140
150
160
170
180
190
200
B
C
Select Sequence Peaks
Peak Selection
SNP
Reference
Undo
Clear
D
E

Figure 4. PeakPicker software developed for quantitative allele ratio analysis. (*A*) PeakPicker allows parallel viewing of sequence traces. The three traces shown are illustrating analysis of rs4015375 (*STEAP*); the *top* trace represents genomic DNA with a clear heterozygous (C/G) genotype (red arrow); the two traces *below* show the independent replicates of the corresponding RNA samples, in which predominant expression of one allele (G) is evident. (*B*) Multiple alignment is carried out and samples passing the user-defined alignment score (% match) are analyzed further. (*C*) A reference sequence (typically one of the genomic DNA controls) is used to select SNP and control peaks. (*D*) PeakPicker identifies the SNP and control peaks in all sequences analyzed in parallel and generates a table flagging suspicious values to allow manual review of peak selection. (*E*) A text file with SNP allele ratios normalized based on the control peaks is generated as an output.

dom hexamers (Invitrogen), and first-strand cDNA synthesis (RT) was performed using Superscript II reverse transcriptase according to the manufacturer's instructions (Invitrogen). RT-reactions were carried out in duplicate from each RNA sample.

Allelic expression analysis by high-sensitivity sequencing

PCR and sequencing-primer designs avoided known SNPs underlying primer sequences, and applied same designs for RNA and DNA samples unless an exonic SNP was located close to the exon–intron boundary. All primer designs are available at http://genomequebec.mcgill.ca/EST-HapMap. All RNA samples were amplified in duplicate from independent cDNA preparations, and a subset (4–10/SNP) of informative gDNA heterozygotes were amplified in identical conditions to establish the expected 50:50 heterozygote profiles. Sequencing was carried out using 0.5 µL of BigDye Terminator (BDT) v3.1 Cycle Sequencing Kit (Applied Biosystems) with 2 µL of a PCR product, 1.75 µL of 5× sequencing buffer (0.4 M Tris-HCl at pH 9.0, 10 mM $MgCl_2$), and 10 pmol of the sequencing primer in a 10-µL reaction in conditions suggested by the manufacturer. Reaction products were separated on an Applied Biosystems 3730XL DNA analyzer. The sequence traces were analyzed using in-house software (see below). The normalized heterozygote ratios of genomic DNA samples were used to establish a 95% confidence interval (95% CI) for each SNP. If both heterozygote ratios in independent RNA samples showed convergent deviation beyond 95% CI derived from genomic DNA data, the sample was called to have allelic imbalance. If one of the two RNA replicates was within the 95% CI or if the replicates deviated to opposite directions, the sample was defined as "unknown."

Databases and resources

EST sequences were obtained fromUniGene (Build #173 of *Homo sapiens* with 4,636,789 total sequences in clusters) and public SNP genotypes from the International HapMap Project at http://www.HapMap.org (release #13, Nov. 2004). The human genome reference sequence was downloaded from UCSC at http://genome.ucsc.edu (release hg16), which is based on NCBI Build 34. RefSeq gene structures were also from UCSC, which were based on hg16 and updated in November 2004. The EST sequences were mapped on the Human genome sequence by BLAT (Kent 2002).

Genomic SNP allele frequencies in four populations were calculated from HapMap data. Unrelated individuals were used for frequency calculation. We used 60 independent individual data from the CEPH Europe population (CEU), 45 from Han Chinese in Beijing (HCB), 44 from Japanese in Tokyo (JPT), and 60 from

Table 2. Experimental validation by sequencing RNA samples of informative heterozygotes

			In silico allelic imbalance data								AI_{95} analysis per LCL Validation by sequencing in LCL RNA				Locus-specific analysis (AI_{LS})	
			Allele[b]		Allele counts HapMap		Allele counts dbEST									
Gene[a]	SNP	CHR	*H*	*L*	*H*	*L*	*H*	*L*	No. of ESTs[c]	*P*-value	*H* > *L*[d]	*L* > *H*[e]	*L* = *H*[f]	Unknown[g]	Average allele ratio RNA/DNA[h]	Allele ratio *P*-value[i]
STEAP1	rs4015375	7	C	G	8	112	19	0	23	1.88E-17	**5**	**0**	**0**	**0**	**7.70**	**1.84E-04**
GCS1	rs1063588	2	T	C	11	109	17	23	48	8.51E-06	2	0	1	7	**1.07**	**4.04E-02**
WRB	rs2837005	21	T	C	25	95	10	2	12	2.57E-05	**16**	**0**	**0**	**1**	**2.34**	**4.49E-04**
METAP1	rs1238741	4	C	T	11	109	12	17	34	1.11E-04	**9**	**0**	**0**	**2**	1.07	1.27E-01
GEMIN6	rs1056104	2	A	G	6	114	12	35	56	3.58E-04	**3**	**0**	**0**	**3**	**1.19**	**2.28E-02**
PBK	rs1052874	8	G	C	6	114	11	32	53	5.06E-04	**4**	**0**	**0**	**3**	**1.40**	**1.74E-03**
PAX8	rs1478	2	G	T	17	103	21	32	65	5.35E-04	6	1	2	8	**1.57**	**3.42E-02**
HADHB	rs1056471	2	G	C	10	110	38	122	211	6.85E-04	**6**	**0**	**0**	**2**	**1.20**	**2.30E-05**
OAS1	rs2660	12	G	A	43	77	37	22	67	7.97E-04	**29**	**0**	**0**	**0**	**2.17**	**1.11E-10**
LGMN	rs2236264	14	T	C	13	107	25	65	199	2.00E-03	**8**	**0**	**0**	**2**	**1.28**	**1.34E-04**
EPHX2	rs1042064	8	C	T	31	89	34	36	84	2.44E-03	**20**	**2**	**0**	**4**	**2.87**	**2.77E-04**
MTHFD2	rs12196	2	A	G	68	52	72	22	119	2.45E-03	**25**	**0**	**0**	**6**	**1.18**	**2.81E-08**
RAB7L1	rs823137	1	G	A	59	59	15	2	19	3.37E-03	**7**	**1**	**0**	**17**	**1.07**	**3.18E-02**
PISD	rs8461	22	T	C	28	92	24	27	68	3.37E-03	**10**	**0**	**0**	**15**	**1.41**	**1.91E-02**
ARTS-1	rs26653	5	C	G	33	85	7	2	17	4.33E-03	**17**	**0**	**3**	**3**	**1.43**	**6.06E-03**
ELL3	rs2788	15	G	A	8	112	7	25	43	1.78E-02	**6**	**0**	**0**	**2**	**1.33**	**3.94E-04**
CORO1C	rs2111211	12	C	T	54	64	55	8	67	2.93E-08	6	0	9	7	1.50	1.37E-01
VPS39	rs7086	15	C	G	12	108	34	64	119	1.04E-05	0	4	4	3	0.89	1.92E-01
SNX6[j]	rs9264	14	C	T	65	53	62	14	97	1.81E-04	NA	NA	NA	NA	NA	NA
CD200	rs1050572	3	A	G	6	114	11	30	72	3.43E-04	1	3	0	2	0.93	6.31E-02
CXCL16	rs1051007	17	G	A	7	111	12	32	60	4.98E-04	0	3	0	3	0.84	1.05E-02
FVT1	rs6810	18	A	G	57	63	56	21	86	6.42E-04	1	11	5	13	0.93	1.25E-01
FLJ12788	rs2301984	2	G	A	12	108	11	18	39	6.98E-04	0	3	1	6	1.01	8.48E-01
PTPN12	rs3750050	7	G	A	11	109	12	23	39	7.05E-04	2	4	0	2	0.91	8.98E-02
GRN	rs5848	17	T	C	20	100	103	212	518	8.14E-04	9	3	1	4	1.64	5.62E-02
TRAF3	rs1131877	14	C	T	26	92	9	4	22	9.05E-04	1	4	9	4	0.95	5.55E-01
SEC61A1	rs1042907	3	G	C	89	31	103	10	155	9.11E-04	1	4	1	13	0.98	2.98E-01
MGC5576	rs6823	12	C	G	55	63	78	37	144	1.44E-03	0	4	12	15	0.91	2.89E-01
STK33	rs2289921	11	G	C	49	71	14	3	24	1.53E-03	9	7	1	10	1.15	1.50E-01
LMAN1	rs1127220	18	C	T	30	90	23	23	56	2.85E-03	0	0	15	0	0.99	9.30E-01
GATM	rs1049518	15	A	G	39	77	38	30	96	3.49E-03	9	11	1	7	1.09	2.48E-01
HPS4	rs3747134	22	G	A	10	110	8	17	38	3.59E-03	0	0	7	2	0.85	2.53E-01
ARPC5	rs11755	1	A	G	51	69	54	31	100	4.44E-03	0	5	15	9	0.93	1.99E-01
FXYD2	rs11999	11	C	A	36	80	20	14	69	4.58E-03	10	7	0	6	4.91	9.38E-02
MCM2	rs893293	3	C	T	24	96	28	46	165	7.78E-03	3	0	3	14	1.19	1.66E-01
PIK3R1	rs3756668	5	A	G	56	64	14	3	23	8.21E-03	6	0	13	13	0.99	9.01E-01
ZNF350	rs2278414	19	A	G	16	104	6	7	17	8.30E-03	2	4	0	11	0.88	7.52E-02
ACSL5	rs8624	10	C	T	30	90	22	25	56	8.99E-03	6	0	4	13	1.12	9.60E-02
CDK2	rs2069398	12	G	A	108	12	59	0	128	9.39E-03	3	4	2	2	0.97	6.69E-01
PPID	rs2070629	4	C	T	78	42	19	2	26	2.13E-02	0	20	0	5	0.80	6.79E-03

[a] Genes on top (from *STEAP1* to *ELL3*) were validated by either qualitative or quantitative analysis of allelic expression data. The data points in bold correspond to the data fulfilling the validation criteria mentioned in the text. Genes from *CORO1C* to *PPID* did not fulfill the criteria for validation.
[b] Alleles are ordered based on the "expected high (*H*) expressor" and "expected low (*L*) expressor" as predicted by the EST-genotype comparison.
[c] Total number of EST sequence traces in UniGene, a maximum of two per library were included in the EST allele counts.
[d] Number of heterozygous individuals showing overexpression of the predicted high allele as determined by consistent deviation in independent cDNA samples beyond the 95% confidence interval. If >80% of samples fulfilled the prediction, the data points fulfill the validation criteria and are shown in bold.
[e] Number of heterozygous individuals showing overexpression of the expected "low" allele as determined by consistent deviation in independent cDNA samples beyond the 95% confidence interval.
[f] Number of heterozygous individuals showing equal expression of alleles as determined by both RNA samples falling to 95% confidence interval observed for genomic DNA controls.
[g] Number of informative samples that did not fall into the preceding three categories in the allele ratio analysis and remained "unclassified."
[h] The average ratio of predicted high-allele versus the predicted low allele in RNA is divided by the value of predicted high-allele versus the predicted low-allele in control heterozygous DNA samples (i.e., H_{RNA}/L_{RNA}:H_{DNA}/L_{DNA}). If this ratio >1 and the distribution of values in RNA versus DNA is statistically significant (*t*-test) the candidate SNP is considered validated and is shown in bold.
[i] *P*-value (*t*-test, two-tailed) for difference between the *H*/*L* ratios in genomic DNA versus RNA.
[j] Both RNA and DNA samples showed (concordant) variation of allele ratios in *SNX6,* thus unequal expression could be caused by DNA-copy number variation or unidentified SNPs underlying the sequencing primers. The data was omitted from further analysis.

Yoruba in Ibadan (YRI). The criteria for SNP selection for the comparison were: A SNP should be transmitted at least once in 30 trios of CEPH families and should be located inside the RefSeq gene exon region, based on the UCSC RefSeq gene annotation database. For four different population frequency comparisons, we used a subset of SNPs that were common to all four populations.

We used EST sequence data to estimate SNP allele frequencies in RNA samples. EST sequences from UniGene were first mapped to the reference genome sequence by using BLAT. To avoid mapping errors, the EST sequences were required to be at least 70% identical to genomic sequence in the matched region and the EST were expected to map to the same UniGene cluster region. Approximately 3,856,000 sequences satisfied these two requirements and then were used in this study. Based on EST sequence alignment, which involved base-to-base matching to reference genome sequence, and SNP location on the genome, we counted both SNP alleles on the EST sequence. To limit the bias introduced by unequal representation of EST libraries in EST sequence data, a maximum of two sequences were counted from one library. Two-sided Fisher's exact tests (implemented in R) were performed on 11822 SNPs to compare the allele frequencies between CEPH-derived genomic and EST-derived RNA samples. The same SNP was included under two UniGene entries for ~5% of data; in our candidate gene list we eliminated the SNPs with double entries, as these might reflect UniGene errors or true overlapping transcripts (the complete data set is also available on the Web site). To compare the distances from the SNP to the 5′-end of the respective gene for SNPs with (test group) or without significant deviation (control group) in EST–SNP comparisons, we selected the annotated isoform with the shortest distance to the SNP (UCSC hg 16). The control SNPs were selected by matching two SNPs for each test SNP. The randomly chosen control SNPs fulfilled the following criteria: (1) the *P*-value of the Fisher's test was >0.9 in the original EST allele frequency comparison; and (2) the averages of EST counts per site were matched; as well as (3) the averages of allele frequencies were matched between test and control groups.

Software for normalized peak height analysis

To improve the consistency of heterozygote ratio measurements, we applied normalization of SNP allele height with reference peaks selected from the flanking sequence. To facilitate this process, we developed the "PeakPicker" software. PeakPicker is developed for quantitative allele ratio analysis and can be used to determine differential allelic expression in cells heterozygous for a marker SNP expressed in mRNA by measuring and calculating the peak height ratios of the marker SNP.

The input files for PeakPicker are the raw sequence files from ABI sequencers (*.AB1). Multiple alignments of the selected sequences are carried out with a user-defined sequence range using Needleman-Wunsch for global alignment. After identifying the marker SNP in a sequence and identifying bases to which its height should be compared (reference peaks), PeakPicker identifies and analyzes the SNP and reference peaks in all sequences. Because peak heights vary depending on sample, base type, and their position within the chromatogram, a normalization step is performed. A text file with SNP allele ratios normalized on the reference peaks is generated as an output. Suspicious values are flagged to allow manual review of the peak selection.

PeakPicker is written in Java language and can be used in many platforms such as Windows and Linux systems. "PeakPicker" is available at http://genomequebec.mcgill.ca/EST-HapMap. Sequence traces for the assays described in Table 2 are available upon request.

Statistical analysis of validation data

For the data analysis using the AI_{95} test defined above, each allele was assigned a state: (1) concordant direction of expression, (2) discordant direction of expression, (3) no preferential expression of alleles, or (4) unknown. The statistical evaluation of the concordant (i.e., observed direction matching the predicted) direction of overexpression was compared to the null hypothesis (i.e., no preferential bias in allelic expression), which was calculated using χ^2 statistics. Empirical *P*-values were obtained by performing 10,000 simulations. Both steps were performed using an on-line Java-applet for calculations of χ^2 values and performing permutations for the generation of empirical *P*-values (available at http://davidmlane.com/hyperstat/chi_square.html). Similarly, the difference in the number of genes showing >80% of overexpressed alleles concordant with predictions was compared to the null hypothesis (i.e., no preferential bias in allelic expression). The comparison of allele ratios derived from the $\mathrm{AI}_{\mathrm{LS}}$ method required *H*/*L* ratios to be generated using the predicted high-expressing allele (*H*) divided by the predicted low-expressing (*L*) allele in genomic DNA and RNA (cDNA). When the *H*/*L* ratio is higher in RNA than in DNA, the overexpression is concordant with the predicted direction. The statistical significance of this comparison was obtained using a two-tailed *t*-test statistic within a locus (see Table 2, rightmost column) or across loci.

Acknowledgments

T.J.H. is the recipient of a Clinician-Scientist Award in Translational Research by the Burroughs Wellcome Fund and an Investigator Award from CIHR. This work is supported by Genome Canada and Genome Quebec.

References

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., and Venter, J.C. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

Bray, N.J., Buckland, P.R., Owen, M.J., and O'Donovan, M.C. 2003. *Cis*-acting variation in the expression of a high proportion of genes in human brain. *Hum. Genet.* **113:** 149–153.

Buetow, K.H., Edmonson, M.N., and Cassidy, A.B. 1999. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21:** 323–325.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D., and Lander, E.S. 2002. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32:** 432–437.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229–232.

Field, L.L., Bonnevie-Nielsen, V., Pociot, F., Lu, S., Nielsen, T.B., and Beck-Nielsen, H. 2005. OAS1 splice site polymorphism controlling antiviral enzyme activity influences susceptibility to type 1 diabetes. *Diabetes* **54:** 1588–1591.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W., and Lee, C.J. 2000. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* **26:** 233–236.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Knight, J.C. 2005. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **83:** 97–109.

Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H., and Lee, M.P. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res.* **13:** 1855–1862.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23:** 452–456.

Mathe, C., Sagot, M.F., Schiex, T., and Rouze, P. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30:** 4103–4117.

Pastinen, T. and Hudson, T.J. 2004. *Cis*-acting regulatory variation in the human genome. *Science* **306:** 647–650.

Pastinen, T., Sladek, R., Gurd, S., Sammak, A., Ge, B., Lepage, P., Lavergne, K., Villeneuve, A., Gaudin, T., Brandstrom, H., et al. 2004.

A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics* **16:** 184–193.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199–204.

Rockman, M.V. and Wray, G.A. 2002. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19:** 1991–2004.

Sato, K., Emi, M., Ezura, Y., Fujita, Y., Takada, D., Ishigami, T., Umemura, S., Xin, Y., Wu, L.L., Larrinaga-Shum, S., et al. 2004. Soluble epoxide hydrolase variant (Glu287Arg) modifies plasma total cholesterol and triglyceride phenotype in familial hypercholesterolemia: Intrafamilial association study in an eight-generation hyperlipidemic kindred. *J. Hum. Genet.* **49:** 29–34.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.

Yamamoto, N., Nakayama, J., Yamakawa-Kobayashi, K., Hamaguchi, H., Miyazaki, R., and Arinami, T. 2002. Identification of 33 polymorphisms in the adipocyte-derived leucine aminopeptidase (ALAP) gene and possible association with hypertension. *Hum. Mutat.* **19:** 251–257.

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B., and Kinzler, K.W. 2002. Allelic variation in human gene expression. *Science* **297:** 1143.

Yang, H.H., Hu, Y., Edmonson, M., Buetow, K., and Lee, M.P. 2003. Computation method to identify differential allelic gene expression and novel imprinted genes. *Bioinformatics* **19:** 952–955.

Web site references

http://davidmlane.com/hyperstat/chi_square.html; Web-based statistical tools for determination of χ^2 values and calculation of empirical *P*-values.

http://genomequebec.mcgill.ca/EST-HapMap; Additional information relating to current manuscript.

http://genome.ucsc.edu; Human Genome Browser.

http://www.HapMap.org; Web site for description of and data from the International Haplotype Map Consortium.

Received April 9, 2005; accepted in revised form July 12, 2005.