Human Molecular Genetics, 2005, Vol. 14, No. 23 ***3741–3749***
doi:10.1093/hmg/ddi404
Advance Access published on October 26, 2005

Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes

Samuel Deutsch[1,†], Robert Lyle[1,††], Emmanouil T. Dermitzakis[2,†], Homa Attar[1], Lakshman Subrahmanyan[5], Corinne Gehrig[1], Leila Parand[1], Maryline Gagnebin[1], Jacques Rougemont[3], C. Victor Jongeneel[3,4] and Stylianos E. Antonarakis[1,*]

[1]Department of Genetic Medicine and Development, Geneva University Medical School, 1 Rue Michel Servet, CH-1211 Geneva, Switzerland, [2]Population and comparative genomics group, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK, [3]Vital-IT Centre, Swiss Institute of Bioinformatics, Lausanne, Switzerland, [4]Ludwig Institute for Cancer Research, Office of Information Technology, Ch. des Boveresses 155 Epalinges, Switzerland and [5]School of Medicine, University of Massachusetts, 55 Lake Avenue North, Worcester, MA, USA

Received August 9, 2005; Revised and Accepted October 19, 2005

Inter-individual differences in gene expression are likely to account for an important fraction of phenotypic differences, including susceptibility to common disorders. Recent studies have shown extensive variation in gene expression levels in humans and other organisms, and that a fraction of this variation is under genetic control. We investigated the patterns of gene expression variation in a 25 Mb region of human chromosome 21, which has been associated with many Down syndrome (DS) phenotypes. Taqman real-time PCR was used to measure expression variation of 41 genes in lymphoblastoid cells of 40 unrelated individuals. For 25 genes found to be differentially expressed, additional analysis was performed in 10 CEPH families to determine heritabilities and map loci harboring regulatory variation. Seventy-six percent of the differentially expressed genes had significant heritabilities, and genomewide linkage analysis led to the identification of significant eQTLs for nine genes. Most eQTLs were in *trans*, with the best result ($P = 7.46 \times 10^{-8}$) obtained for *TMEM1* on chromosome 12q24.33. A *cis*-eQTL identified for *CCT8* was validated by performing an association study in 60 individuals from the HapMap project. SNP rs965951 located within *CCT8* was found to be significantly associated with its expression levels ($P = 2.5 \times 10^{-5}$) confirming *cis*-regulatory variation. The results of our study provide a representative view of expression variation of chromosome 21 genes, identify loci involved in their regulation and suggest that genes, for which expression differences are significantly larger than 1.5-fold in control samples, are unlikely to be involved in DS-phenotypes present in all affected individuals.

INTRODUCTION

Understanding the relationship between sequence variation present in human populations (1) and phenotypic diversity is one of the main challenges of functional genomics. Only a fraction of polymorphic variation is likely to be of functional significance, and traditionally, studies have focused on polymorphisms that alter the primary sequence of proteins as prime candidates for functional variation (qualitative changes). Early studies in model organisms as well as in humans have shown that functional protein sequence variants or hypomorphs are indeed a valid paradigm to explain trait

*To whom correspondence should be addressed. Tel: +41 223795708; Fax: +41 223795706; Email: stylianos.antonarakis@medecine.unige.ch
[†]These authors equally contributed to this study.
[††]Present address: EPAM, Norwegian Institute of Public Health, Oslo, Norway.

differences within populations (2–4). Interestingly, however, numerous recent association studies implicating genetic variation of a gene to a particular phenotype have failed to identify coding SNPs as candidates for the etiological effect, suggesting that other mechanisms are likely to underlie the molecular pathology in these disorders (5–7).

An alternative way in which sequence variation can have a functional impact is by affecting the steady-state level of mRNA molecules of a particular gene in a given cell (quantitative changes). Since the 1970s, it has been suggested that quantitative differences in gene expression might provide a significant source of variation in natural populations, representing an important substrate for evolution and accounting for a considerable fraction of phenotypic diversity (8).

Changes in the dosage of a gene or group of genes have previously been shown to be associated with human disorders such as trisomy 21 and other contiguous gene syndromes (9–12). It is thus likely that differences in gene expression can also explain the population variance of many traits.

High-throughput expression profiling in a number of organisms has revealed that variation in gene expression levels within and among populations is abundant, with a large proportion of genes (20–40% in most studies) showing significant patterns of inter-individual variation (13–16). Although this variation could be of stochastic, environmental or genetic origin, analyses of expression differences in segregating populations (from lower eukaryotes to mammals) have demonstrated that a significant proportion of the variance has a genetic component (16–19). As such, gene expression levels can be considered as a quantitative trait (eQTL) and thus be dissected using standard genetic approaches such as quantitative linkage analysis.

Understanding the pattern of gene expression variation in humans is important, as it is likely to underlie a significant proportion of the risk for many common disorders. In the context of human chromosome 21 (Hsa21), this question is particularly relevant because it could help to understand the relationship between trisomy of Hsa21 and Down syndrome (DS) phenotypes and, in particular, give new insights into the molecular basis of the extensive phenotypic heterogeneity observed in trisomy 21 patients (20).

In our study, we set out to investigate the patterns of gene expression variation in a 25 Mb region of Hsa21, using a high-precision approach combining TaqMan real-time PCR and multiple replicates per sample.

We screened for inter-individual differences in expression in 41 genes using a panel of lymphoblastoid cell lines derived from 40 unrelated individuals from the CEPH collection (21). Genes found to be differentially expressed were further analyzed in 10 three-generation families in order to determine: (i) What fraction of the variation has a genetic component? (ii) Which loci in the genome are associated with these expression traits?

RESULTS

Gene expression variation

Out of 71 Taqman assays initially designed, a total of 41 met our efficiency criteria of $0.95 < E < 1.05$ (10) in RNAs from lymphoblastoid cell lines. These 41 genes were screened for differential gene expression in cDNAs obtained from 40 unrelated individuals, namely the grandparents of 10 CEPH families (Fig. 1). For each individual, we calculated the normalized relative expression for each of the 41 genes as described in Methods section. The median 95% CIs for relative expression measurements was ± 0.18, indicating that inter-individual differences as small as $\sim$1.4-fold could be readily detected for the large majority of assays.

To exclude Epstein–Barr virus (EBV) transformation effects and cell culture conditions as major sources of gene expression variation, we transformed lymphocytes from the same individual six independent times over a 2-week period. Expression analysis of 25 HSA21 genes and seven normalization genes in these cell lines showed high-expression correlations for all genes, and inter- and intra-cell line variation were not significantly different (R. Lyle and A. Reymond, unpublished data).

To determine which genes were differentially expressed between individuals, we used the variance ratio (15), which gives the ratio of the inter-individual variance over the average intra-individual error. Twenty-five genes were selected for further study based on having a variance ratio >1 (Table 1).

The median fold change in expression (ratio of individual with highest expression over individual with lowest expression) for the set of differentially expressed genes was 3.2, but values ranged from 2 to 47.43 (Table 1). However, for some genes (e.g. *APP* and *CBS*) in which a number of individuals display very low expression levels, the fold change might be overestimated due to limitations in the sensitivity of detection.

Heritability

To determine what fraction of the variance of the differentially expressed genes could be attributed to genetic components, we measured expression levels in 10 three-generation families from the CEPH collection (21). In total, we assayed cDNAs from 135 individuals, including the initial 40 unrelated individuals who are the grandparents of these pedigrees.

Seventeen out of the 25 differentially expressed genes had significant heritability values ($P < 0.05$), with a maximum of 0.84 observed for *SLC37A1* ($P = 8.07 \times 10^{-11}$) and a median of 0.42 for all genes (Fig. 2).

We also calculated the significance of the heritability estimates for each trait by multiple permutation analysis. For this purpose, we permuted the phenotype (normalized relative expression level) for each gene 500 times and calculated the heritability using the same family structures. Results from this calculation showed that all 17 genes previously found to have significant heritabilities remained significant, and in addition, two of the genes that had been considered marginally not significant (*INFAR2* and *PRDM15* both at $P = 0.06$) were significant at the 95% level by multiple permutation and were thus considered for quantitative linkage analysis.

eQTL mapping

To identify candidate loci involved in the transcriptional control of chromosome 21 genes, we performed genomewide quantitative linkage analysis on the 10 study families. We used

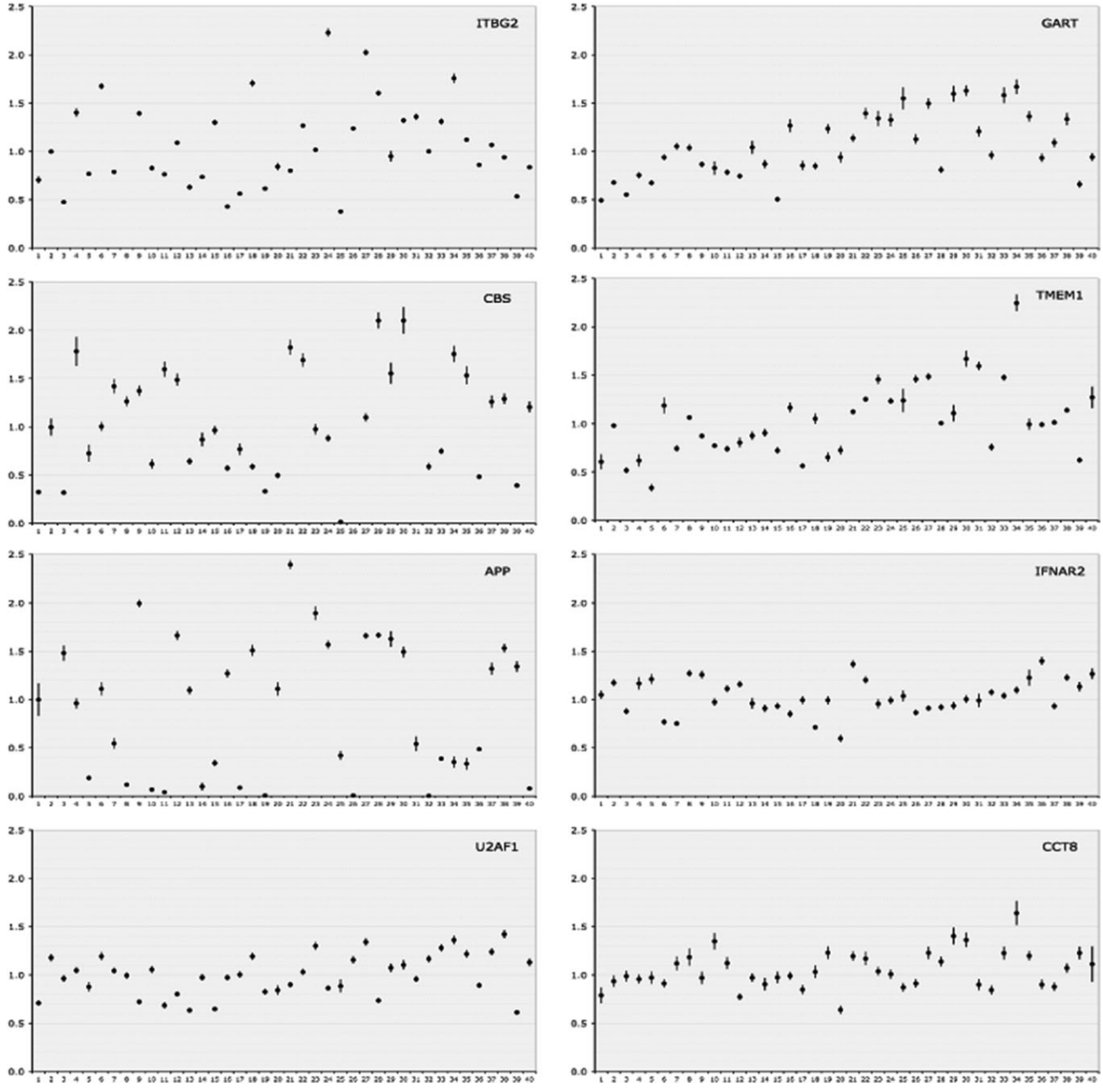

Figure 1. Normalized relative expression values for eight genes. Circles denote mean values for each of the 40 unrelated individuals, and bars indicate standard error of the mean.

SNP genotyping information available in public databases. All genotyping errors as assessed by Mendelian incompatibility patterns were removed (22).

For our pedigrees, we estimated 60% power to detect a suggestive ($P < 0.001$) eQTL given the median heritability of 0.42. This power increases to >90% for heritability values over 0.6.

Multipoint linkage analysis resulted in the identification of at least one eQTL for four genes (first four genes) (Table 2) using a threshold of $P < 1.6 \times 10^{-4}$ (which corresponds to the theoretical genomewide significance of 0.05) (23). The most significant result was obtained for *TMEM1* with $P = 7.46 \times 10^{-8}$ (LOD score of 6). In one case, the eQTL for a gene (*CCT8*) mapped to within 5 Mb of its physical location (*cis*-eQTL), whereas the other eQTLs were located elsewhere in the genome (*trans*). In addition, for *TMEM1*, multiple *trans*-eQTLs were identified (Fig. 3).

If the criteria are relaxed to a threshold of $P < 0.001$ (suggestive linkage), eQTLs for eight additional genes are identified (last eight genes) (Table 2).

Table 1. Inter-individual gene expression variation

Gene	Variance ratio	Fold expression difference
ITGB2	34.22	6.96
CBS	28.7	13.8
APP	24.32	47.43
PFKL	22.76	2.43
U2AF1	17.31	2.30
PRDM15	13.01	14.65
LSS	11.64	3.61
PDXK	9.81	2.77
SLC19A1	9.52	4.60
SLC37A1	8.96	3.90
PWP2H	7.34	2.0
MCM3AP	6.65	2.17
GART	6.28	2.56
CBR1	6.11	4.82
TMEM1	5.5	5.12
BTG3	5.4	3.23
DSCR1	5.23	3.36
ETS2	4.26	9.36
IFNAR2	3.97	2.34
ANKRD3	3.56	16.53
WRB	3.04	2.77
GABPA	2.14	2.54
SON	2.14	2.51
IFNAR1	1.99	2.51
CCT8	1.36	2.37

Simulations

Variance component analysis is one of the most commonly used model-free methods to perform quantitative linkage analysis (24,25); however, it assumes that the traits analyzed are normally distributed. Violations of normality by the presence of outliers or excessive kurtosis can result in an increased level of type 1 errors. (24,26,27) Many quantitative traits are not normally distributed, for example, 84% of our expression traits significantly deviate from normality. Ignoring such deviations can lead to important misinterpretation of the results (24).

One way to deal with the problem of non-normally distributed traits is to determine the empirical significance of the linkage results (obtained through variance component or regression methods) by performing simulation studies. To this end, we computed 1000 simulations by randomizing the genotypes but keeping the expression phenotypes constant. Genotype randomization has the advantage of keeping heritability patterns and other aspects of the data structure, such as marker density, allele frequencies and missing data unchanged.

We analyzed the results of the simulations by extracting the highest linkage score per genome-scan per trait, in order to build a significance distribution (the simulated maxium LOD scores for each trait follow chi-square distributions). This analysis showed that in nine out of 12 cases, the genes with significant or suggestive eQTLs remained significant at the 95% level (genes in bold, Table 2). Hence, a high proportion of the eQTLs with suggestive *P*-values can be considered as significant on a genomewide basis according to the simulations.

For some traits, the rates of type 1 errors were considerably higher than expected (Fig. 3). For example, in the case of *TMEM1* where several *trans*-eQTLs were initially identified, only one locus remained significant after adjustment.

As performing large-scale simulations is a computationally intensive process, we assessed an alternative method to address the problem of non-normality. We used Box–Cox and bivariate-normal copula transformations in order to approximate the traits distributions to normal. These transformations have a minimal effect on the internal correlations of the data. One thousand simulations were performed for each transformed trait to determine the effect this would have on the power and the rates of type 1 errors. Results of these simulations showed that while the power was not greatly affected by either type of data transformation (data not shown), the effects on the levels of type 1 error were variable and produced, at best, only small improvements (Supplementary Material, Fig. S1). These results suggest that data transformations are not efficient at solving the increased rates of type 1 errors observed for some traits, and thus,

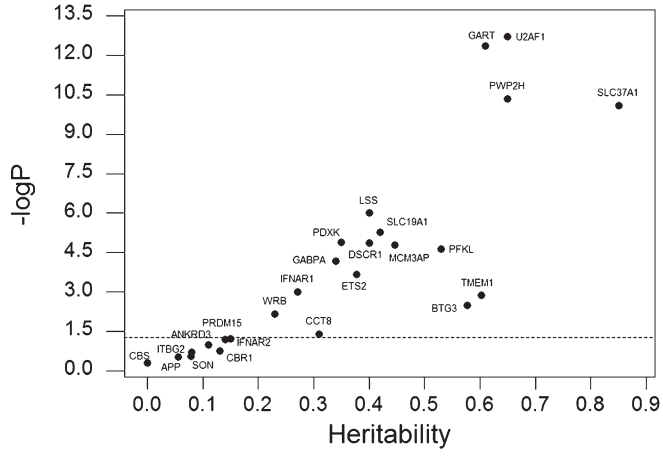

Figure 2. Heritabilities and associated *P*-values (expressed as $-\log 10P$) for the 25 differentially expressed genes. Dotted line indicates 95% significance threshold.

Table 2. Results of genomewide scan for gene expression traits

Gene	eQTL(s) position	SNP multipoint *P*-values (Merlin)
TMEM1	**12q24.33,** 11q24.2, 4q26.2	**7.46×10^{-8},** 4.24×10^{-6}, 2.47×10^{-5}
GART	**6p25.1**	**1.66×10^{-6}**
U2AF1	**6p25.1**	**2.36×10^{-5}**
CCT8	**21q22.11**	**1.3×10^{-4}**
PDXK	**2q21.1**	**2.72×10^{-4}**
SLC19A1	**3q26.2**	**3.94×10^{-4}**
GABPA	**11p15.4**	**7.15×10^{-4}**
PRDM15	**15q13.2**	**7.52×10^{-4}**
SLC37A1	**19q13.42**	**9.42×10^{-4}**
BTG3	12q21.32	4.08×10^{-4}
ETS2	12	4.46×10^{-4}
MCM3AP	4q26.1	8.74×10^{-4}

Values in bold are significant on a genomewide basis as determined by simulation studies.

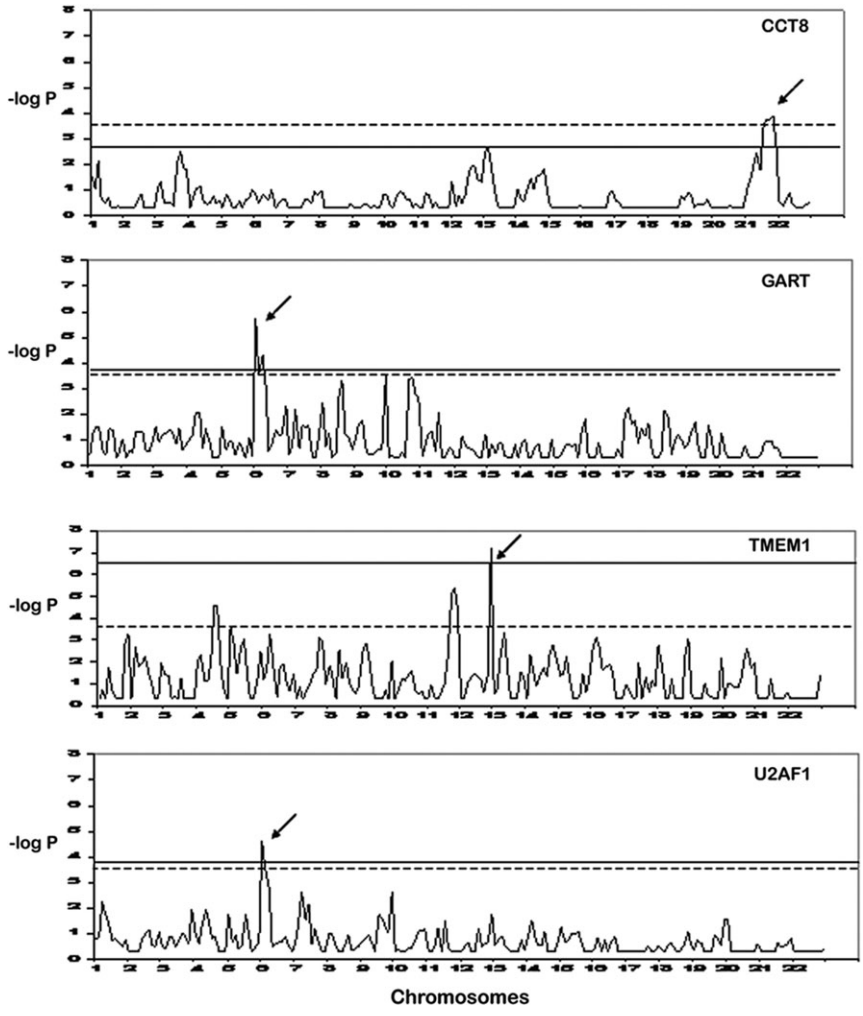

Figure 3. Genomewide multipoint quantitative linkage plots. Dotted line indicates theoretical 95% significance threshold. Full line indicates empirical 95% threshold as determined through simulations. Arrowheads indicate significant peaks.

simulations are highly advisable to properly assess the significance of the results for each trait studied.

Association study

To further validate the *cis*-eQTL identified for the *CCT8* gene, we performed an association study using 60 additional unrelated individuals from the CEPH collection for which high-density genotyping data are available as part of the international HapMap project (28).

Gene expression variation was measured as before, and genotypes for 41 SNPs from a 100 kb region surrounding the *CCT8* gene were obtained. We performed association analysis for each SNP (predictor) to the normalized relative *CCT8* expression values (response) using analysis of variance (ANOVA), and corrected for multiple tests using a step-down approach (29). To guard against the effect of non-normal trait distribution, we transformed the raw phenotypes to their logarithmic values to yield a normally distributed trait (Shapiro test for normality, $P = 0.56$). Four SNPs were significantly associated with *CCT8* expression levels ($P = 2.5 \times 10^{-5}$) even after correction for multiple testing ($P = 9 \times 10^{-4}$). Out of these SNPs, rs965951 is located in intron 14 of the gene, and the three other SNPs (rs2832159, rs8133819, rs2832160) are located 12 kb upstream. All four SNPs are in complete linkage disequilibrium ($r^2 = 1.0$) (Fig. 4).

DISCUSSION

We set out to dissect the patterns of gene expression variation on a 25 Mb segment of Hsa21, which has traditionally been associated with many of the defining features of DS (mental retardation, characteristic facies, muscle hypotonia) (20,30), as well as with some of its variable phenotypes including congenital heart defects (31). In addition, this region of Hsa21 has also been linked to some common complex disorders such as bipolar affective disorder of unknown genetic etiology (32,33).

Our approach consisted in using real-time Taqman PCR and multiple replicates per sample, as this has been shown to allow the detection of small inter-individual differences in gene expression (10) that might nonetheless be physiologically important (34,35). Our data show that differences as small

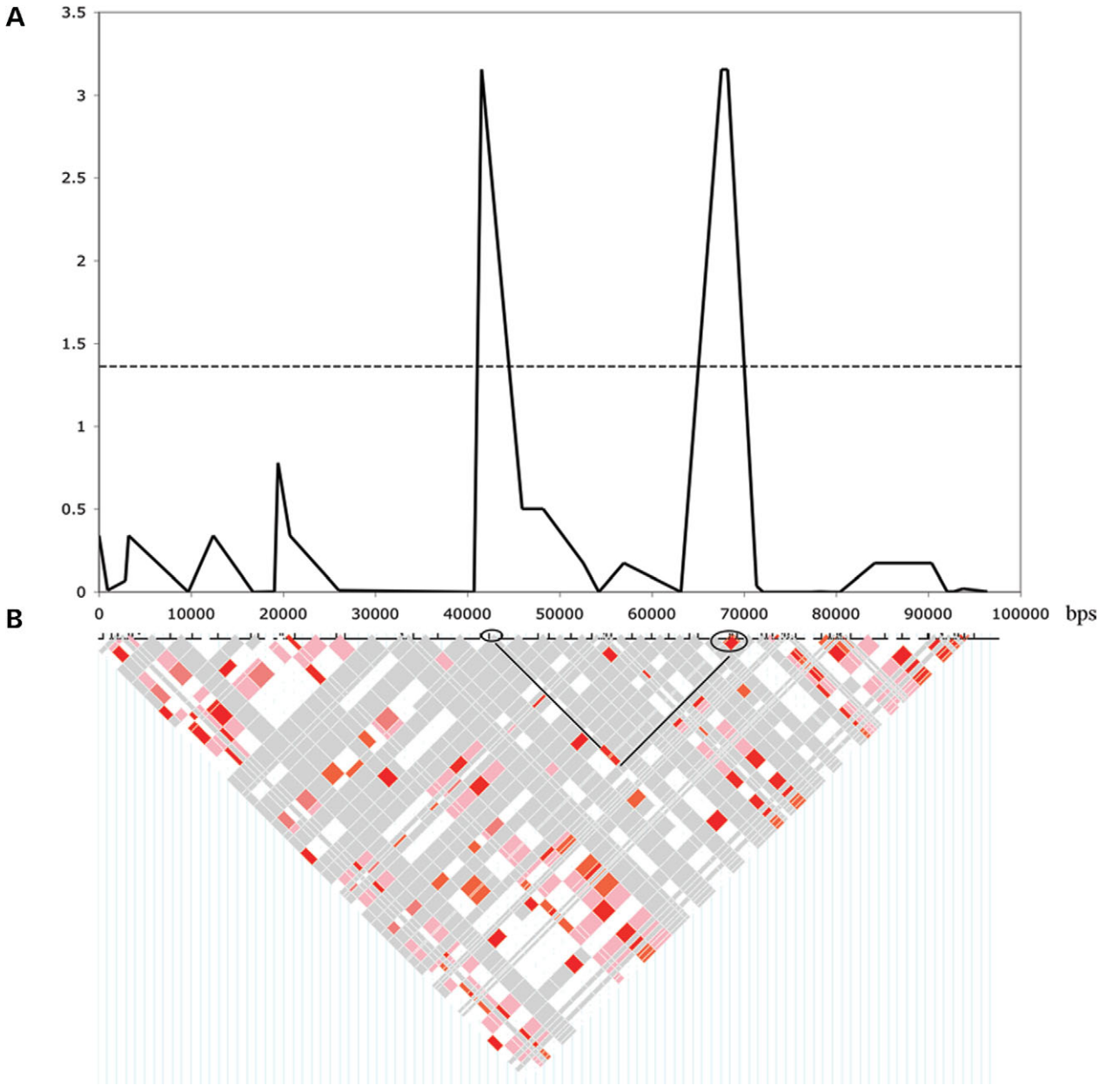

Figure 4. Association of *CCT8* expression levels to surrounding nucleotide variation. (**A**) $-\log(P)$ versus SNP position for 100 kb surrounding *CCT8* (41 SNPs from HapMap data set). *P*-values shown were corrected for multiple hypothesis testing as described in Methods. Dotted line indicates nominally significant ($P < 0.05$) results. Two regions of the locus comprising four SNPs are significant after correction. (**B**) Linkage disequilibrium between SNPs in selected region as viewed with LD plot plugin (http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap). Bright red regions are in complete LD ($r^2 = 1.0$). The two groups of SNPs (four SNPs total) significantly correlated with CCT8 expression are circled. Dark lines point to squares representing complete LD ($r^2 = 1.0$) between these sites.

as 1.4-fold could be reliably measured, which compares favorably to microarray procedures in which few replicates per sample are performed, and for which typically only changes >2-fold can be confidently detected (36).

Our results revealed a substantial amount of inter-individual gene expression variation, as 25 out of the 41 genes analyzed (61%) showed differential expression (based on our conservative criteria). This finding supports the notion that gene expression variation is a highly important but poorly characterized source of molecular diversity that is likely to explain a considerable fraction of the phenotypic variance in the population (37). The median fold change in expression level between the lowest and highest values per gene among individuals was 3.2, with the largest inter-individual expression differences observed for *APP* (Fig. 1). This is interesting, as *APP* is crucially involved in the pathology of Alzheimer's disease, and different levels of *APP* could influence the pathogenesis of this disorder.

DS is a disease caused by alterations in gene dosage such that affected individuals are expected to have an average

1.5-fold overexpression of Hsa21 genes (20). Given the extensive variation in gene expression observed in our samples among normal individuals (at least 2-fold), we would predict that for many Hsa21 genes, there is a considerable overlap in total expression levels between normal and trisomy 21 individuals due to allelic variation. Thus, expression variation at the levels we report here may help to explain two related aspects of DS phenotypes: penetrance and variability. Overexpressed genes which have low levels of expression variation would be predicted to lead to the more penetrant phenotypes (e.g. *CCT8* and *U2AF1*). In contrast, genes with high variation in expression (e.g. *ITGB2* and *CBS*) would contribute to incompletely penetrant/variable DS-related phenotypes (model shown in Supplementary Material, Fig. S2).

However, because the gene expression variation data reported here are based on lymphoblastoid cell lines, confirmation of these results in additional cell types/tissues would be of great interest.

Analysis of the segregation patterns of these 'expression traits' in 10 CEPH pedigrees revealed that for 76% (19/25) of genes, a significant fraction of the variation is explained by genetic factors, with a median heritability of 0.42. Genome-wide quantitative linkage analysis using previously generated genotyping data (38) led to the identification of significant eQTLs for nine out of the 19 genes, after correction for biases in the data structure and for multiple testing. The large majority of the eQTLs identified were in *trans*, suggesting that quantitative or qualitative differences in certain genes (transcription factors, proteins involved in signal transduction and others) (18) have an impact on larger transcriptional networks. However, given the results of other studies showing abundant allele-specific expression differences (39,40), we expect *cis*-regulatory variation to be common and most likely under-represented in our results.

For the *CCT8* gene, which was the only one with a *cis*-eQTL, we performed an association study to validate the linkage findings. Results of the association analysis revealed that a group of four SNPs within or close to CCT8 were significantly associated with its expression levels, independently confirming the presence of *cis*-regulatory variation.

An interesting finding is that for some genes with high expression heritabilities (e.g. *PWP2H* and *PFKL*), no significant eQTL was identified even though power calculations on our sample size suggest ~90% power to detect such loci. We thus hypothesize that in many cases, expression traits are regulated by multiple loci, each of which contributes only modestly to the trait. This higher complexity in the genetic architecture (41,42) underlying expression variation indicates that only in cases in which few loci account for a large proportion of the expression variability, eQTLs will be identified effectively. This may explain the under-representation of *cis*-eQTLs, and suggests that larger sample sizes are required to dissect more complex genetic regulation.

An important observation from our study is the increased type 1 error for many of the expression traits when performing variance components calculations for eQTL mapping. This highlights the need to take appropriate measures, such as simulations, to enable correct interpretation of the significance of the results.

In two previous studies, QTL mapping of gene expression traits was performed in humans using microarrays (17,43). These studies have a higher throughput, with thousands of genes being analyzed in parallel, which gives a more global view of the genetic control of gene expression. However, for specific regions involved in the etiology of human disorders, such as the one presented here, a more focused study design with higher number of replicates is possible, leading to a more accurate estimation of individual expression levels.

In this study, we provide a detailed view of gene expression variation of Hsa21 genes and a screen for eQTLs involved in their regulation. The extensive expression variation observed has important implications concerning the molecular pathogenesis and phenotypic variability in DS patients. Large-scale association studies with appropriate samples sizes will constitute the next step for the identification of regulatory variation.

MATERIALS AND METHODS

Families and lymphoblastoid cell culture

We obtained EBV-transformed lymphoblastoid cell lines of 135 individuals belonging to 10 CEPH (21) families (1333, 1334, 1340, 1341, 1345, 1346, 1347, 1362, 1408, 13292) from Coriell cell repositories. All cell lines were grown in RPMI 1640 with Glutamax I medium (Invitrogen Corporation) supplemented with 10% fetal calf serum and 1% penicillin and streptomycin mix (Invitrogen Corporation). Cells lines were harvested at a density of $0.6-1 \times 10^6$ cells/ml and at least 80% viability. Cultures were spun for 5 min at 1000 *g*, and the resulting pellets were washed once in PBS and lysed by adding 2 ml of micro-glass beads (Sigma) and vortexing in 1 ml lysis solution containing β-mercaptoethanol (Qiagen, RNeasy kit). Cell lysates were stored at -80°C.

RNAs were extracted using RNeasy mini kits with on-column DNAse I digestion (Qiagen). RNA samples were quantified by spectrophotometry, and the quality was assessed using an Agilent 2100 BioAnalyzer with the RNA 6000 Nano LabChip. All RNAs had a 260/280 nm ratio between 1.8–2, and a 28s/18s rRNA ratio above 2.

cDNAs were synthesized from total RNA using SuperScriptII reverse transcriptase (Invitrogen Corporation) and a poly d(T) primer. For each cell line, 5 μg of total RNA in a total volume of 20 μl were used, and the resulting cDNA was diluted 1:14 prior to PCR.

Gene selection and assay design

We chose to study a 25 Mb region of Hsa21 encompassing both the so-called DS critical (30) region and the minimal region thought to be involved in the development of the heart defect phenotypes (atrial or ventricular septal defects or complete atrioventricular canal) present in ~40% of DS patients (31). This region contains around 109 well-characterized genes.

On the basis of experimental data indicating expression in lymphoblastoid cell lines (www.ensembl.org/Multi/martview?species=Homo_sapiens), we selected 71 genes for which we designed Taqman assays. Assay designs are listed in Supplementary Material, Table S1.

Real-time quantitative PCR

Real-time quantitative PCR was carried out essentially as described in Lyle *et.al* (10). Briefly, intron-spanning Taqman assays were designed using PrimerExpress (Applied Biosystems) with default parameters. Assay efficiencies were calculated using a cDNA dilution series (44). All PCRs were performed using a qPCR mastermix (RT-QP2X-03, Eurogentec).

Reactions were set up using a Biomek 2000 robot (Beckman), in a 10 μl volume in 384-well plates. Six replicates per gene per sample were performed. PCRs were run in an ABI 7900 Sequence Detection System (Applied Biosystems) with the following conditions: 50°C for 2 min, 95°C for 10 min and 50 cycles of 95°C 15 s/60°C for 1 min.

To maximize the reproducibility of our results, we premixed the required amount of cDNA, with a large volume of PCR mastermix, to ensure uniformity in the starting concentration of cDNA in all assays and replicates. In addition, the experiment was designed so that all assays for a particular individual were run at the same time on a single plate, as this renders the normalizations more accurate.

We checked that all C_T-values were within the ranges tested in assay efficiency tests. In addition, no systematic biases were detected in the results as determined by assessing correlation between the following variables: PCR efficiency (E), threshold cycle (C_T), relative expression and error (10).

qPCR data analysis

Raw cycle threshold (C_T) values were obtained using SDS 2.0 software (Applied Biosystems). A threshold value of 0.2 was used for all genes, and background was changed manually for individual genes as recommended by Applied Biosystems. From the six replicates per gene, outliers were detected using Grubb's test at the 95% significance level. For all calculations, C_T-values were converted to quantity (q) with the formula $q = 2^{-C_T}$. Analyses were carried out in Excel, SPSS and Minitab (MINITAB Inc.).

We normalized expression values by performing a median normalization across all genes, excluding the 75th percentile to minimize the effect of outliers. Each gene was then median normalized across individuals. Normalized relative expression values thus have a median of 1.

To determine which genes show significant levels of expression variation in our cohort, we measured expression levels in a group of 40 unrelated individuals. We evaluated the levels of gene expression variation by means of the variance ratio obtained by dividing the inter-individual variance of the means of each gene by the mean intra-individual error. We selected 25 genes which had a variance ratio above 1.3.

Heritability, linkage, simulations and association

Heritability calculations were performed using the 'polygenic-screen' command from the SOLAR software (45). To calculate the empirical significance of the heritability values, we performed a multiple permutation test in which the phenotypic values for all traits were randomly permuted 500 times and heritabilities calculated as before. These results were used to determine the 95% significance thresholds.

SNP genotyping data, consisting of 2688 autosomal SNPs with an effective resolution of 3.9 cM, were downloaded from the SNP Consortium database (http://snp.cshl.org/linkage_maps/) (38). Multipoint linkage with the SNP map was performed using Merlin (46) with the –VC option, after Mendelian inconsistencies (PEDCHECK) (22) and unlikely genotypes (PEDWIPE) (46) were removed.

To calculate the empirical significance of the linkage results, we performed 1000 simulations for each quantitative trait using the –simulate command from Merlin with different seed numbers. We extracted the highest result from each simulation to build significance distributions. The distribution of the maximum scores of permutations followed a chi-squared distribution.

To evaluate the effect of non-normality on the power and levels of type 1 errors in our data, we transformed all traits by: (i) Box–Cox transformation using Minitab, (ii) the bivariate normal copula as described by Basrak *et al*. (27). We performed 1000 simulations for each transformed trait and built significance distributions as described above. All simulations were performed using a cluster of 32 HP/Intel Itanium 2 based servers at the Vital-IT Center.

Association of CCT8 expression and *cis*-nucleotide variation was done using anovasnp (version 0.7) (D. Posada, manuscript in preparation) with 100 000 permutations to correct for multiple hypotheses using a step-down procedure (29). Genotypes were downloaded from the HapMap project URL (http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap), HapMap public release no. 16c.1.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Alexandre Reymond, Bernard Conrad and Jacques Beckmann for useful discussions and critical reading of the manuscript. Denis Cohen for bioinformatics help, and Dorret Boomsma and Bojan Basrak for providing the scripts for copula transformation. This work was funded by the Foundation Jérôme Lejeune, the ChildCare Foundation, The Swiss National Science Foundation (3100-057149.99/1), the EU/OFES (QLG1-CT-2002-00816) and the NCCR Frontiers in Genetics.

Conflict of Interest statement. Authors have not declared any conflict of interest.

REFERENCES

1. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
2. Marinkovic, D. and Ayala, F.J. (1975) Fitness of allozyme variants in *Drosophila* pseudoobscura. I. Selection at the PGM-1 and Me-2 loci. *Genetics*, **79**, 85–95.
3. McDonald, J.F., Anderson, S.M. and Santos, M. (1980) Biochemical differences between products of the Adh locus in *Drosophila*. *Genetics*, **95**, 1013–1022.

4. Kasvosve, I., Delanghe, J.R., Gomo, Z.A., Gangaidzo, I.T., Khumalo, H., Wuyts, B., Mvundura, E., Saungweme, T., Moyo, V.M., Boelaert, J.R. *et al.* (2000) Transferrin polymorphism influences iron status in blacks. *Clin. Chem.*, **46**, 1535–1539.
5. Zhang, Y., Leaves, N.I., Anderson, G.G., Ponting, C.P., Broxholme, J., Holt, R., Edser, P., Bhattacharyya, S., Dunham, A., Adcock, I.M. *et al.* (2003) Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nat. Genet.*, **34**, 181–186.
6. Chumakov, I., Blumenfeld, M., Guerassimenko, O., Cavarec, L., Palicio, M., Abderrahim, H., Bougueleret, L., Barry, C., Tanaka, H., La Rosa, P. *et al.* (2002) Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc. Natl Acad. Sci. USA*, **99**, 13675–13680.
7. Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T.T. *et al.* (2002) Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.*, **71**, 877–892.
8. King, M.C. and Wilson, A.C. (1975) Evolution at two levels in humans and chimpanzees. *Science*, **188**, 107–116.
9. FitzPatrick, D.R., Ramsay, J., McGill, N.I., Shade, M., Carothers, A.D. and Hastie, N.D. (2002) Transcriptome analysis of human autosomal trisomy. *Hum. Mol. Genet.*, **11**, 3249–3256.
10. Lyle, R., Gehrig, C., Neergaard-Henrichsen, C., Deutsch, S. and Antonarakis, S.E. (2004) Gene expression from the aneuploid chromosome in a trisomy mouse model of down syndrome. *Genome Res.*, **14**, 1268–1274.
11. Kahlem, P., Sultan, M., Herwig, R., Steinfath, M., Balzereit, D., Eppens, B., Saran, N.G., Pletcher, M.T., South, S.T., Stetten, G. *et al.* (2004) Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of down syndrome. *Genome Res.*, **14**, 1258–1267.
12. Mao, R., Zielke, C.L., Zielke, H.R. and Pevsner, J. (2003) Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain. *Genomics*, **81**, 457–467.
13. Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. (2002) Variation in gene expression within and among natural populations. *Nat. Genet.*, **32**, 261–266.
14. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
15. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.
16. Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
17. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
18. Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.*, **35**, 57–64.
19. Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A. and Schadt, E.E. (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, **75**, 1094–1105.
20. Antonarakis, S.E., Lyle, R., Dermitzakis, E.T., Reymond, A. and Deutsch, S. (2004) Chromosome 21 and down syndrome: from genomics to pathophysiology. *Nat. Rev. Genet.*, **5**, 725–738.
21. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M. and White, R. (1990) Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, **6**, 575–577.
22. O'Connell, J.R. and Weeks, D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **63**, 259–266.
23. Lander, E. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, **11**, 241–247.
24. Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I. and Blangero, J. (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.*, **65**, 531–544.
25. Cherny, S.S., Sham, P. and Cardon, L.R. (2004) Introduction to the special issue on variance components methods for mapping quantitative trait loci. *Behav. Genet.*, **34**, 125–126.
26. Shete, S., Beasley, T.M., Etzel, C.J., Fernandez, J.R., Chen, J., Allison, D.B. and Amos, C.I. (2004) Effect of winsorization on power and type 1 error of variance components and related methods of QTL detection. *Behav. Genet.*, **34**, 153–159.
27. Basrak, B., Klaassen, C.A., Beekman, M., Martin, N.G. and Boomsma, D.I. (2004) Copulas in QTL mapping. *Behav. Genet.*, **34**, 161–171.
28. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
29. Westfall, P.H. and S.S. Young (1993) *Resampling-Based Multiple Testing*. John Wiley & Sons, New York.
30. Delabar, J.M., Theophile, D., Rahmani, Z., Chettouh, Z., Blouin, J.L., Prieur, M., Noel, B. and Sinet, P.M. (1993) Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet.*, **1**, 114–124.
31. Barlow, G.M., Chen, X.N., Shi, Z.Y., Lyons, G.E., Kurnit, D.M., Celle, L., Spinner, N.B., Zackai, E., Pettenati, M.J., Van Riper, A.J. *et al.* (2001) Down syndrome congenital heart disease: a narrowed region and a candidate gene. *Genet. Med.*, **3**, 91–101.
32. Straub, R.E., Lehner, T., Luo, Y., Loth, J.E., Shao, W., Sharpe, L., Alexander, J.R., Das, K., Simon, R., Fieve, R.R. *et al.* (1994) A possible vulnerability locus for bipolar affective disorder on chromosome 21q22.3. *Nat. Genet.*, **8**, 291–296.
33. Liu, J., Juo, S.H., Terwilliger, J.D., Grunn, A., Tong, X., Brito, M., Loth, J.E., Kanyas, K., Lerer, B., Endicott, J. *et al.* (2001) A follow-up linkage study supports evidence for a bipolar affective disorder locus on chromosome 21q22. *Am. J. Med. Genet.*, **105**, 189–194.
34. Gouya, L., Puy, H., Robreau, A.M., Bourgeois, M., Lamoril, J., Da Silva, V., Grandchamp, B. and Deybach, J.C. (2002) The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype FECH. *Nat. Genet.*, **30**, 27–28.
35. Yan, H., Dobbie, Z., Gruber, S.B., Markowitz, S., Romans, K., Giardiello, F.M., Kinzler, K.W. and Vogelstein, B. (2002) Small changes in expression affect predisposition to tumorigenesis. *Nat. Genet.*, **30**, 25–26.
36. Mutch, D.M., Berger, A., Mansourian, R., Rytz, A. and Roberts, M.A. (2002) The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, **3**, 17.
37. Stamatoyannopoulos, J.A. (2004) The genomics of gene expression. *Genomics*, **84**, 449–457.
38. Matise, T.C., Sachidanandam, R., Clark, A.G., Kruglyak, L., Wijsman, E., Kakol, J., Buyske, S., Chui, B., Cohen, P., de Toma, C. *et al.* (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.*, **73**, 271–284.
39. 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.
40. 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.
41. Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
42. Wang, W.Y., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.
43. Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A. and Schadt, E.E. (2004) Genetic Inheritance of Gene Expression in Human Cell Lines. *Am. J. Hum. Genet.*, **75**.
44. Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*, **25**, 402–408.
45. Almasy, L. and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, **62**, 1198–1211.
46. Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.