OPEN ACCESS Freely available online

PLoS GENETICS

The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study

Alexandra C. Nica[1,2], Leopold Parts[1], Daniel Glass[3], James Nisbet[1], Amy Barrett[4], Magdalena Sekowska[1], Mary Travers[4], Simon Potter[1], Elin Grundberg[1,3], Kerrin Small[1,3], Åsa K. Hedman[4], Veronique Bataille[3], Jordana Tzenova Bell[3,4], Gabriela Surdulescu[3], Antigone S. Dimas[2,4], Catherine Ingle[1], Frank O. Nestle[5], Paola di Meglio[5], Josine L. Min[4], Alicja Wilk[1], Christopher J. Hammond[3], Neelam Hassanali[4], Tsun-Po Yang[1], Stephen B. Montgomery[2], Steve O'Rahilly[6], Cecilia M. Lindgren[4], Krina T. Zondervan[4], Nicole Soranzo[1,3], Inês Barroso[1,6], Richard Durbin[1], Kourosh Ahmadi[3], Panos Deloukas[1]*, Mark I. McCarthy[4,7,8]*, Emmanouil T. Dermitzakis[2]*, Timothy D. Spector[3]*, The MuTHER Consortium

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, **2** Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland, **3** Department of Twin Research, King's College London, London, United Kingdom, **4** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **5** St. John's Institute of Dermatology, King's College London, London, United Kingdom, **6** University of Cambridge Metabolic Research Labs, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, United Kingdom, **7** Oxford Centre for Diabetes, Endocrinology, and Metabolism, University of Oxford, Churchill Hospital, Oxford, United Kingdom, **8** Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom

Abstract

While there have been studies exploring regulatory variation in one or more tissues, the complexity of tissue-specificity in multiple primary tissues is not yet well understood. We explore in depth the role of *cis*-regulatory variation in three human tissues: lymphoblastoid cell lines (LCL), skin, and fat. The samples (156 LCL, 160 skin, 166 fat) were derived simultaneously from a subset of well-phenotyped healthy female twins of the MuTHER resource. We discover an abundance of *cis*-eQTLs in each tissue similar to previous estimates (858 or 4.7% of genes). In addition, we apply factor analysis (FA) to remove effects of latent variables, thus more than doubling the number of our discoveries (1,822 eQTL genes). The unique study design (Matched Co-Twin Analysis—MCTA) permits immediate replication of eQTLs using co-twins (93%–98%) and validation of the considerable gain in eQTL discovery after FA correction. We highlight the challenges of comparing eQTLs between tissues. After verifying previous significance threshold-based estimates of tissue-specificity, we show their limitations given their dependency on statistical power. We propose that continuous estimates of the proportion of tissue-shared signals and direct comparison of the magnitude of effect on the fold change in expression are essential properties that jointly provide a biologically realistic view of tissue-specificity. Under this framework we demonstrate that 30% of eQTLs are shared among the three tissues studied, while another 29% appear exclusively tissue-specific. However, even among the shared eQTLs, a substantial proportion (10%–20%) have significant differences in the magnitude of fold change between genotypic classes across tissues. Our results underline the need to account for the complexity of eQTL tissue-specificity in an effort to assess consequences of such variants for complex traits.

Citation: Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS Genet 7(2): e1002003. doi:10.1371/journal.pgen.1002003

Editor: Gregory Barsh, Stanford University School of Medicine, United States

Received October 18, 2010; **Accepted** December 15, 2010; **Published** February 3, 2011

Funding: Wellcome Trust funded the study and also supported ACN. Additional support was provided by the Louis-Jeantet Foundation to ETD and ACN, the Wellcome Trust grant 077016/Z/05/Z, and the United Kingdom NIHR Cambridge Biomedical Research Centre to IB. JTB is a Wellcome Trust Henry Wellcome Fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: emmanouil.dermitzakis@unige.ch (ETD); tim.spector@kcl.ac.uk (TDS); mark.mccarthy@drl.ox.ac.uk (MIM); panos@sanger.ac.uk (PD)

Introduction

Gene expression is an essential cellular function whose regulation determines a significant proportion of the phenotypic variance. Using microarrays and recently second generation sequencing (RNA-seq) [1,2], major progress has been made in understanding the genetics of human gene expression and identifying loci that drive differential expression across individuals [3,4], populations [5–7] and tissues [7–11]. This development is especially valuable for the biological analysis of genome-wide association (GWAS) signals [12], which often map to non-genic regions and are thus hard to interpret in the absence of additional information [13].

Transcript abundance is a very proximal endophenotype affected by genetic variation and has already facilitated the identification of candidate susceptibility genes for metabolic disease traits [14], asthma [15] or Crohn's disease [16]. This has been mostly possible when the tissue of expression was relevant to the interrogated complex trait, as disease phenotypes manifest themselves only in certain tissues. eQTLs discovered in LCLs have primarily helped explain GWAS associations with immunity-related disorders [17,18] while associations with obesity-related traits were mostly observed when gene expression was quantified in adipose tissue [9]. Nevertheless, our guess of tissue relevance is yet far from satisfactory [19], reinforcing thus the incontestable

Author Summary

Regulation of gene expression is a fundamental cellular process determining a large proportion of the phenotypic variance. Previous studies have identified genetic loci influencing gene expression levels (eQTLs), but the complexity of their tissue-specific properties has not yet been well-characterized. In this study, we perform *cis*-eQTL analysis in a unique matched co-twin design for three human tissues derived simultaneously from the same set of individuals. The study design allows validation of the substantial discoveries we make in each tissue. We explore in depth the tissue-dependent features of regulatory variants and estimate the proportions of shared and specific effects. We use continuous measures of eQTL sharing to circumvent the statistical power limitations of comparing direct overlap of eQTLs in multiple tissues. In this framework, we demonstrate that 30% of eQTLs are shared among tissues, while 29% are exclusively tissue-specific. Furthermore, we show that the fold change in expression between eQTL genotypic classes differs between tissues. Even among shared eQTLs, we report a substantial proportion (10%–20%) of significant tissue differences in magnitude of these effects. The complexities we highlight here are essential for understanding the impact of regulatory variants on complex traits.

value of measuring expression in multiple cell-types (including primary tissues reflecting in vivo patterns).

Transcriptional regulatory networks are expected to dictate tissue-specificity of regulatory effects [20], but the extent of this is still under debate. Depending on the cell-types compared and the eQTL discovery methods used, current estimates for tissue-specificity of eQTLs range from ~30% (liver, adipose tissue) [21] to 70–80% (LCL, fibroblasts, T cells) [7].

In this study we investigated various aspects of tissue-specificity and we emphasize the importance of accounting not only for statistical significance but also for continuous biological properties of regulatory variants, such as fold change in expression. We explored the complexity of the human *cis*-regulatory variation landscape in three tissues (LCL, skin and fat) derived from a subset of female Caucasian twins aged between 40 and 87 years old (mean 62 years) from the UK Adult Twin registry [22]. The present study represents the pilot phase of the MuTHER project (Multiple Tissue Human Expression Resource—http://www.muther.ac.uk/), a major resource initiated to enhance our knowledge about common trait susceptibility by providing genome-wide expression, methylation and eventually transcriptome sequencing information for 855 extensively phenotyped twins (clinical, anthropometric, life-style information as well as a wide range of biological measurements are available).

Results

Gene expression was quantified in LCL, skin and fat using Illumina's whole genome expression array (HumanHT-12 version 3) containing 48,803 probes in three technical replicates [E-MTAB-522]. Log_2 - transformed expression signals were normalized separately per tissue by quantile normalization across replicates followed by quantile normalization across individuals. 27,499 probes mapping uniquely to 18,170 Ensembl genes were retained for further analysis. The same individuals had also been genotyped with Illumina's 1M-Duo and 1.2M-Duo chips; 865,544 SNPs with MAF>1% passed quality check (QC). The overlapping set of successfully genotyped samples with available expression data amounted to 156 individuals for LCL (30 MZ pairs, 37 DZ pairs, 22 singletons), 160 for skin (31 MZ pairs, 37 DZ pairs, 24 singletons) and 166 for fat (31 MZ pairs, 40 DZ pairs, 24 singletons). This final dataset was used for eQTL analysis.

We tested for SNP-gene expression associations (eQTLs) separately in each tissue. We considered only unrelated individuals at a time by separating twins from the same pair and thus performing two independent eQTL analyses per tissue. This study design, hereafter named Matched Co-Twin Analysis (MCTA), permits immediate replication and validation of eQTL discoveries. We used Spearman Rank Correlation (SRC) to detect associations and restricted our search to *cis* effects located within 1Mb on either side of a gene's transcription start site (TSS). Statistical significance was assessed at different thresholds using permutations (10,000 per gene) [5]. We detected an abundance of *cis* eQTLs (Table S1A) per tissue at a comparable rate to other studies of similar sample size [5,7]. The reported eQTLs appear robust as they replicate well between individuals of the two co-twin groups per tissue. We measured the eQTL overlap in a continuous fashion by taking the significant SNP-gene associations from one co-twin set and estimating the proportion of true associations (π_1 statistic [23], see Materials and Methods) on the distribution of corresponding p-values in the reciprocal co-twin validation set. High levels of eQTL replication were observed across co-twins, with a mean π_1 of 0.93 in skin and 0.98 in LCL and fat (Table 1). We also measured the estimated proportion of true positives among the subset of genes that did not replicate in the co-twin at the same threshold. This too is high ($\pi_1 = 0.84$ for skin and 0.94 for LCL and fat), suggesting that exact overlap of genes at a given permutation threshold (PT) is an underestimate of eQTL replication due to winner's curse. In other words, we detected eQTLs in the co-twin that clearly replicated the initial findings, but at p-values that marginally missed the initial discovery threshold. To further confirm the robustness of our discoveries, we overlapped the MuTHER LCL results with available eQTL data from two recent independent studies. 40% of the genes for which we detect LCL eQTLs overlap

Table 1. *Cis* eQTL discoveries (number of genes) per tissue at 10^{-3} PT.

Number of significant genes at 10^{-3} PT								
	SRC analysis				SRC-FA analysis			
	Twin 1	Twin 2	Shared	Replication (Mean π_1)	Twin 1	Twin 2	Shared	Replication (Mean π_1)
LCL	509	556	363	0.98	1064	1220	781	0.97
SKIN	238	231	132	0.93	532	542	338	0.95
FAT	462	488	304	0.98	1052	1070	735	0.97

Results from both the Spearman Rank Correlation (SRC) and Factor Analysis (SRC-FA) presented. Proportion of replicating signals calculated as the mean co-twin π_1 estimates from the p-value distribution of same SNP-gene associations in the reciprocal twin set.

doi:10.1371/journal.pgen.1002003.t001

with eQTLs detected in HapMap 3 samples of European ancestry (CEU) (Stranger et al. submitted). Likewise, 36% of the *cis* associations detected by Gibson et. al. in leukocytes derived from 194 southern Moroccan individuals [24] overlap with genes reported in our study. Given the differences in gender distribution, sample preparation or even cell-type tested (LCL versus leukocytes) across these studies, the gene overlap observed is reassuring.

The observed variation in gene expression is not entirely due to genetic effects. Experimental noise and environmental conditions also affect transcript levels in a global manner. Therefore, it is desirable to remove the effects of such random variables and thus increase the power to detect eQTLs. For this purpose, we employed factor analysis (FA) on each tissue separately and corrected for global latent effects on all individuals in each tissue [25]. We fitted various parameters such as number of learned factors and proportion of variance explained, in order to maximize for replication of eQTLs per tissue between twin sets. After performing standard SRC eQTL analysis on the factor-corrected expression data (SRC-FA), we obtained a substantial improvement in eQTL discovery at each of the standard permutation thresholds used (Table S1B). The improvement (twice as many eQTLs at 10^{-3} PT) is consistent in all tissues. The high eQTL replication between twin sets persists after FA, with an additional improvement of true positives detection in skin: $\pi_1 = 0.95$ (Table 1). As expected, FA correction recovers the majority of the eQTLs discovered with the initial analysis (90% of LCL and fat and 80% of skin) ensuring that proximal genetic effects have not been corrected out. The FA correction enabled the discovery of additional signals (Table S2) likely representing real effects that could not be detected initially due to low power. This is supported by the significant overrepresentation of low association p-values ($\pi_1 = 0.99$, Figure 1) estimated in the uncorrected data for eQTLs detected only after FA correction.

Direct tissue overlap of significant eQTLs supports an extensive level of tissue-specificity for the three tissues, with very similar proportions in both the SRC and SRC-FA analyses (Figure 2). In the first co-twin set we discovered 858 eQTL genes (non-redundant union) at 10^{-3} PT in all three tissues (Table 2). Of these, 106 genes (12.35%) are shared across all tissues, 139 (16.2%) are shared in at least two tissues and 613 genes (71.44%) are detected in only one tissue. In skin we detect proportionally fewer tissue-specific effects (10.02% of skin eQTLs are specific to skin at 10^{-3} PT), an observation likely due to tissue heterogeneity and larger variety of present cell-types. SRC-FA results confirm the estimated ~30% of eQTLs to be shared in at least two tissues based on threshold eQTL discovery (Table S3).

Tissue-specific effects are largely not due to tissue-specific expression of the underlying transcripts. We detected regulatory variants active only in one tissue for genes that are expressed at high levels in the other two tissues (Figure S1). The strength of tissue-specificity was investigated further by performing a joint repeated-measures ANOVA analysis with the tissue modelled as a categorical predictor variable (i.e. tissue type comprised the repeated measure). We assessed the relationship to the genotype by inspecting the SNP×tissue interaction p-value term. As expected, we detected a large enrichment of significant SNP×tissue interaction p-values for all associations ($\pi_1 = 0.56$) with tissue-specific effects having higher enrichment ($\pi_1 = 0.6$) than shared ones ($\pi_1 = 0.41$) (Figure S2). The enrichment in the shared category suggests additional attributes of tissue-specificity beyond statistical significance, as presented in the succeeding fold change analysis.

The direction of allelic effects for shared eQTLs (10^{-3} and 10^{-2} PT) is consistent across the three given tissues (Figure S3). As expected, for eQTLs significant in one tissue only the SRC correlation coefficient rho (reflecting direction and magnitude of effects) explains a substantially higher fraction of gene expression variation in the tissue of discovery compared to the other two tissues (identical SNP-gene associations - Figure S4). On the other hand, the amount of expression variance explained by shared eQTLs (10^{-3} PT) is comparable across tissues.

To refine regulatory signals and describe independently acting variants, we mapped eQTLs to recombination hotspot intervals and filtered markers in high LD (Materials and Methods). We found that ~7% of the genes tested are regulated by more than one independent *cis* eQTL, with similar estimates obtained from the standard and factor eQTL analysis (Figure S5). For finer comparison of eQTL effects, we conducted an analysis where sharing was required for both the gene and the genomic interval harboring the eQTL. This analysis yielded similar counts of tissue-shared and specific effects (Tables S4, S5), suggesting that the vast majority of shared genes also share regulatory variants across tissues. Furthermore, as shown previously [7], we observed that eQTLs cluster symmetrically around the TSS, with shared effects being distributed tightly around the TSS and tissue-specific effects spanning a greater range of distances (Figures S6, S7).

The results described so far are based on thresholds, which are driven by statistical significance. Overlaps at these levels are heavily dependent on power and affected by winner's curse. In addition, eQTLs sharing statistical significance may still have notable effect differences on gene expression levels across tissues, with potentially different biological consequences. Given these caveats, we examined tissue-specificity in a continuous manner by quantifying the proportion of true positives estimated from the enrichment of low p-values (π_1). Specifically, the p-value distribution of significant SNP-probe pairs (10^{-3} PT) from a reference tissue was investigated in the other two tissues. The p-value distribution in the other tissues indicates a high degree of tissue sharing (53 to 80%) both with the SRC and SRC-FA, varying slightly depending on the reference tissue in the comparison (Table S6). This suggests that there are effect size differences (both fold change and amount of variance explained) among tissues for the same regulatory variants, which is the basis for the previously described higher eQTL tissue-specificity estimates [7]. Overall, 29% of eQTLs (1-mean π_1) are estimated with the continuous approach to be tissue-specific, when comparing the three tissues studied.

As described above, tissue overlap of eQTLs should encompass not only sharing of a statistically significant regulatory effect, but also a similar effect size (fold change in expression) of that variant across tissues. In this respect, we report the fold change as the difference between the gene expression means of the heterozygous and major homozygous genotypic classes. Within the same tissue, the two co-twin sets are only slightly different in their fold change estimates. These minor differences reflect most probably the winner's curse effect (0.96 Pearson's correlation of fold change between Twin 1 and Twin 2 in LCL, 0.93 in skin and 0.93 in fat - Figure 3, Figures S8, S9). The difference in estimated effect size is much more apparent however across tissues (e.g. LCL eQTLs have a 0.69 and 0.77 fold change correlation with skin and fat eQTLs respectively, skin eQTLs have a 0.69 fold change correlation with fat eQTLs). This is largely a consequence of eQTL tissue-specificity, but a small effect of winner's curse is also expected (as observed in the comparison of co-twin sets). Furthermore, additional possible hidden tissue-specific effects are implied by the fact that shared eQTLs (at the same threshold of significance) don't always share the same effect size across tissues (LCL fold change correlation of 0.78 in skin and 0.84 in fat for shared eQTLs i.e. up to 20% difference in fold change magnitude between tissues compared to within-tissue difference). This suggests that even statistically tissue-shared eQTLs have additional

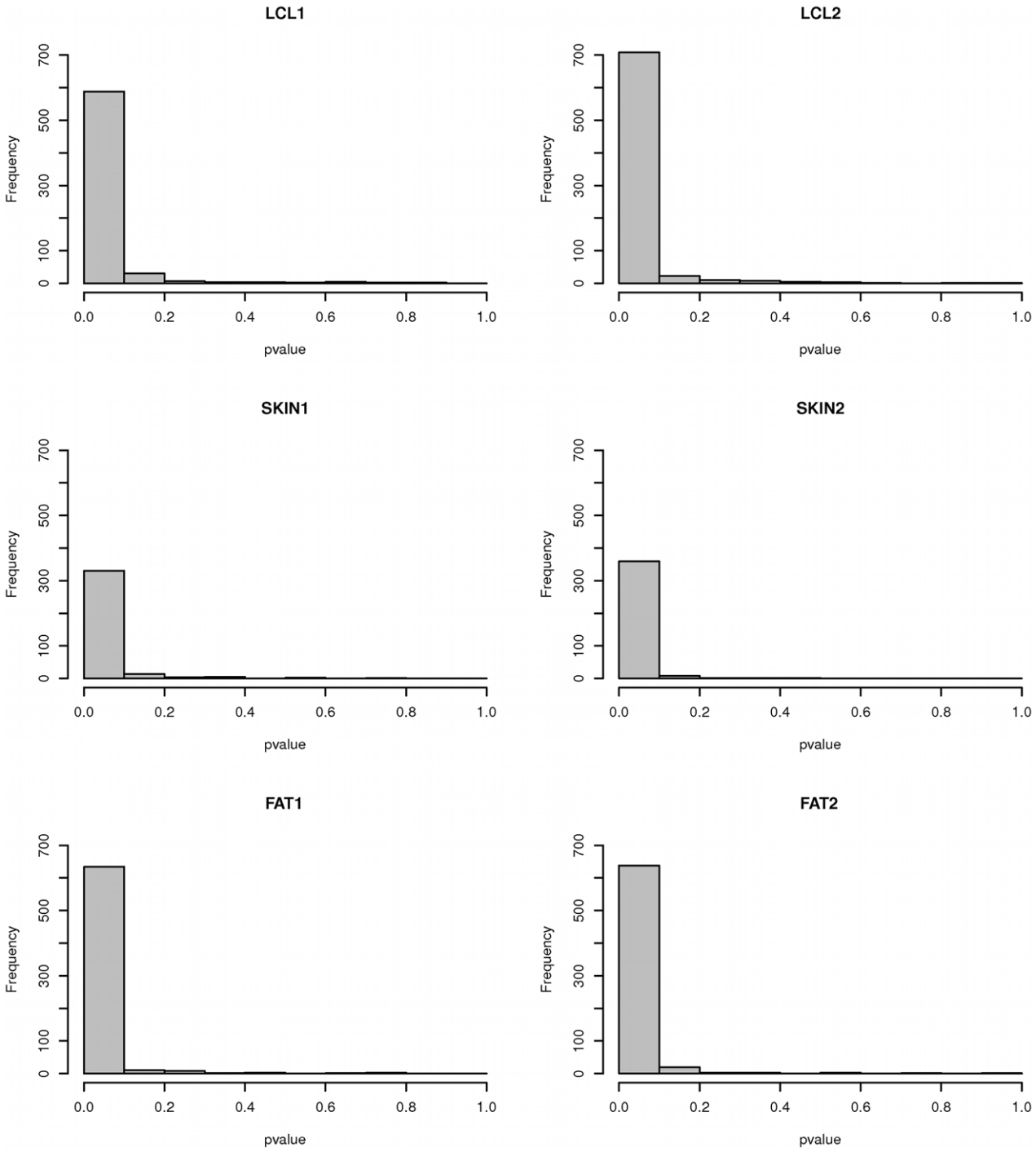

Figure 1. P-value distribution of *cis* eQTLs (10^{-3} PT) gained with FA correction in the uncorrected data. The significant overrepresentation of low p-values for the new eQTLs ($\pi_1 = 0.99$) shows that the signal existed in the uncorrected data but wasn't called significant due to low power. In each tissue, the exact SNP-gene combinations (eQTLs) tested are presented for both co-twin sets (Twin 1—first column, Twin 2—second column).
doi:10.1371/journal.pgen.1002003.g001

dimensions of tissue-specificity and their mere discovery in multiple tissues does not guarantee similar magnitude of consequences.

Discussion

We have performed eQTL analysis in one cell-line (LCL) and two primary tissues of clinical importance (skin – previously uncharacterized and fat). For each tissue we report robust eQTLs replicating in independent samples with identical (MZ) or on average 50% similar (DZ) genetic background using a matched co-twin design (MCTA). To further increase our power to detect eQTLs and uncover smaller genetic effects, we applied factor analysis accounting for global variance components in the data. We refined our signals to detect independently acting *cis* eQTLs and for

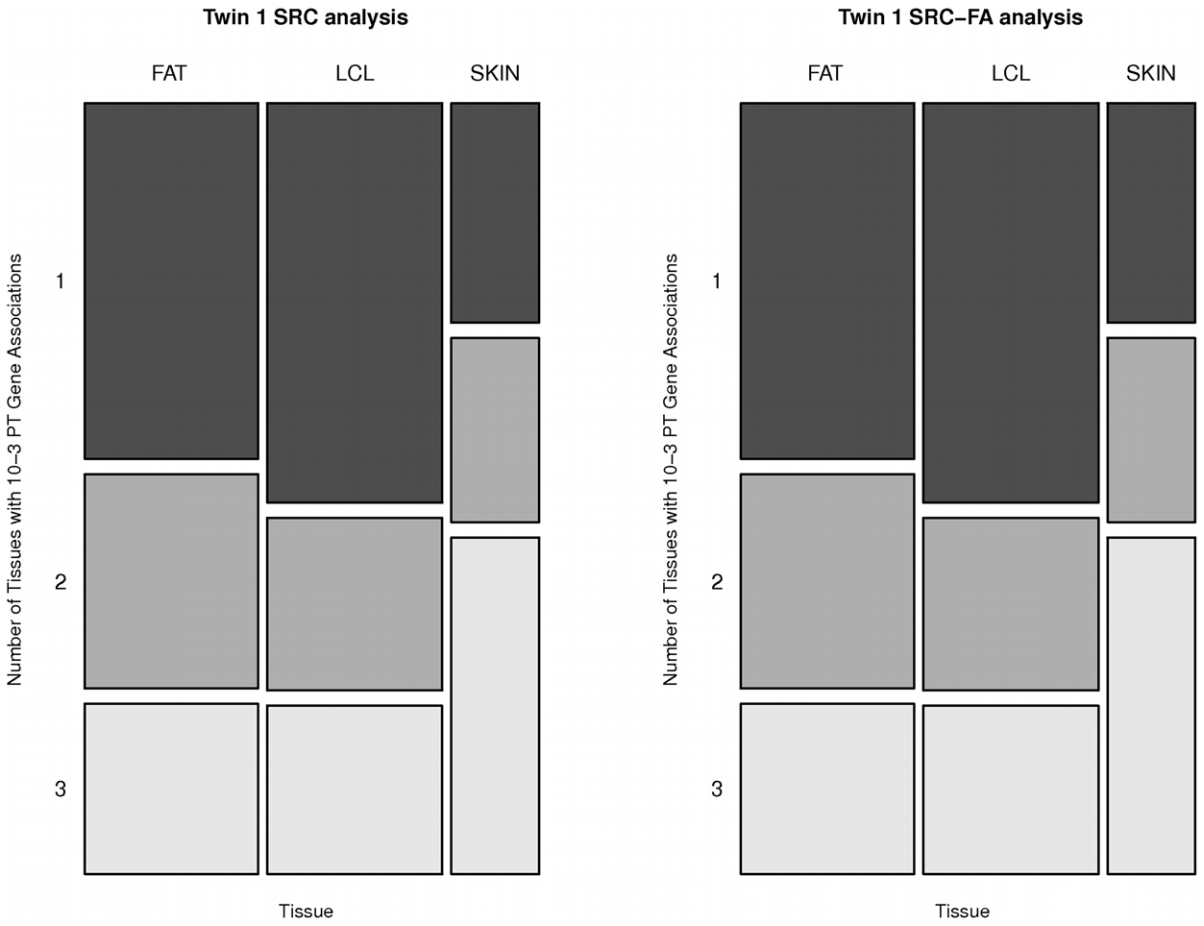

Figure 2. Proportion of tissue shared and tissue-specific eQTLs (10^{-3} PT) from the SRC analysis and SRC-FA respectively. Both methods reveal similarly high extents of tissue-specificity. Skin specific eQTLs of smaller effects are harder to detect due to low power.
doi:10.1371/journal.pgen.1002003.g002

most genes we found single associated regulatory variants. When these variants are shared across tissues, they also share the same direction of allelic effects and map to the same recombination hotspot interval. Using threshold-based criteria, tissue overlap of eQTLs supports a large degree of tissue-specificity for the three tissues studied. However, this estimate is dependent on power and we therefore put forth a continuous measure of tissue-specificity that provides a refined view of the decay of statistical significance as well as fold change effect on gene expression. Using this approach we observed a significant overrepresentation of low p-values in all

Table 2. Tissue-shared and tissue-specific gene associations (10^{-3} PT), SRC analysis.

		Twin 1			Twin 2	
		10^{-3} PT	% total	Overlap	10^{-3} PT	% total
3 tissues	LCL-SKIN-FAT	106	12.35	78	102	11.02
2 tissues only	LCL-SKIN	19	2.21	4	12	1.29
	LCL-FAT	93	10.84	52	107	11.56
	SKIN-FAT	27	3.15	11	26	2.81
1 tissue only	LCL	291	33.92	150	335	36.18
	SKIN	86	10.02	17	91	9.82
	FAT	236	27.50	103	253	27.32
Total significant	LCL	509		363	556	
	SKIN	238		132	231	
	FAT	462		304	488	
Union of total significant		858	100	563	926	100

doi:10.1371/journal.pgen.1002003.t002

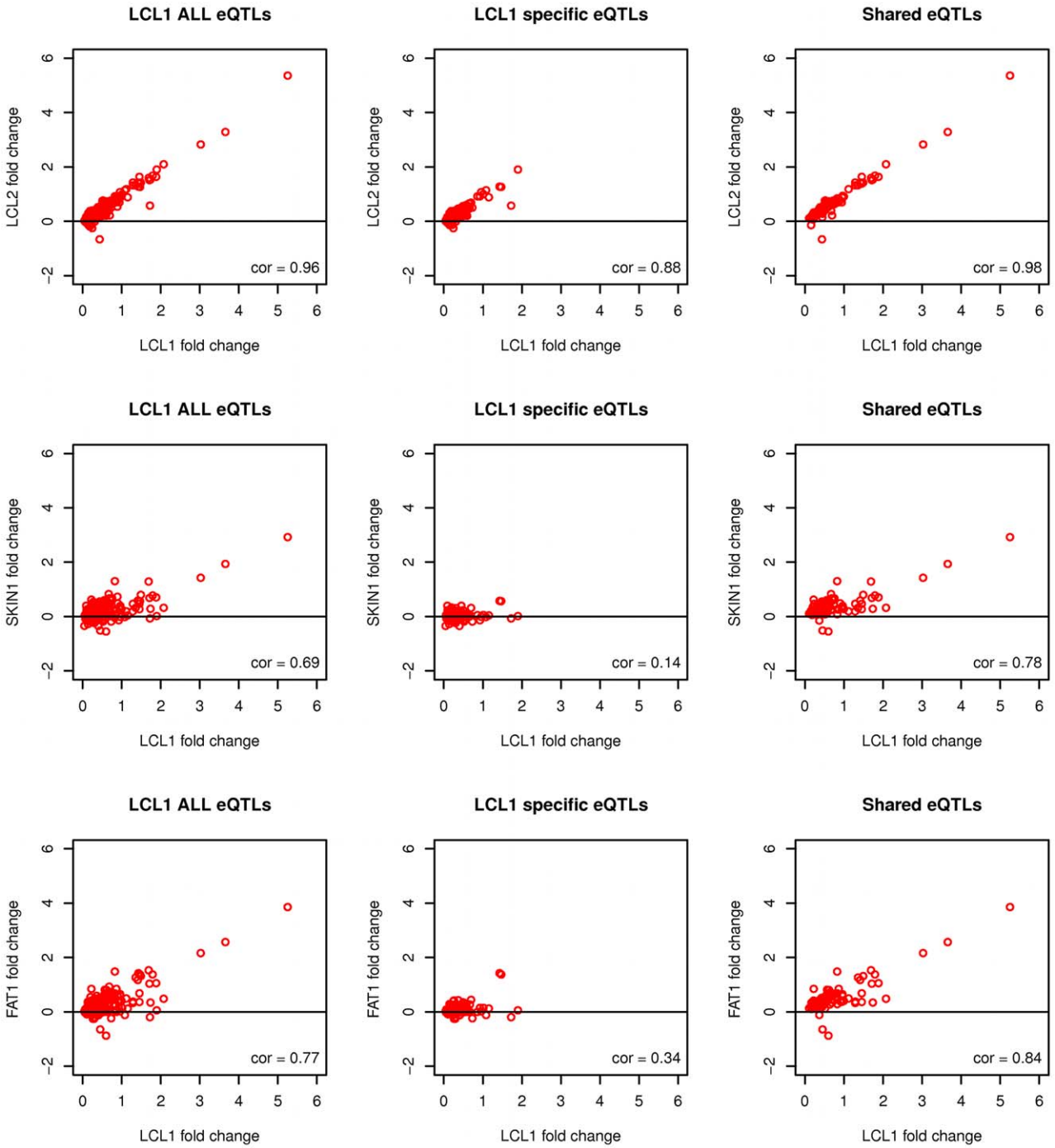

Figure 3. Fold change within twins and across tissues for LCL eQTLs (10^{-3} PT, SRC) discovered in Twin 1. The plotted fold change on the X-and Y-axes was calculated as the difference in mean expression of the heterozygous and major homozygous genotypic classes. For each pairwise tissue comparison, the Pearson's correlation coefficient between fold changes is shown.
doi:10.1371/journal.pgen.1002003.g003

pairwise tissue comparisons, indicating larger proportions of shared statistically significant regulatory effects, some yet to be discovered with bigger sample sizes. However, we also observed significant eQTLs at the same threshold exhibiting differential fold changes in expression between genotypes across tissues. These cases represent tissue-specific effects as well, since differential fold change in expression is likely to have different biological consequences. Overall biological interpretation of regulatory effects - much like in the case of complex traits – is tissue-dependent, highlighting the value of multiple tissue expression datasets. Understanding such complexities and context-dependent effects in the genetic architecture of gene expression and other cellular phenotypes is essential for the interpretation of the biological properties of disease causing variants.

Materials and Methods

All samples and information were collected with written and signed informed consent. The project has been approved by the local ethics committees of all institutions involved.

Sample collection

All individuals recruited in this study were Caucasian female twins aged between 40 and 87 years old (mean age 62). Skin punch biopsies (N = 196) were taken from a relatively photo-protected area adjacent and inferior to the umbilicus. The fat sample was then carefully dissected from the same skin biopsy incision. A peripheral blood sample to generate lymphoblastoid cell lines (LCL) was taken contemporaneously. For a full description of the biopsy technique see Text S1.

Gene expression measurements and genotyping

RNA levels were measured in LCL, skin and fat using Illumina's whole-genome expression array HumanHT-12 version 3 as previously described [5]. Each sample had three technical replicates. Illumina's v3 probes were mapped to unique Ensembl gene IDs by combining and cross-checking two methods. The first approach used Illumina's probe annotation to RefSeq IDs. These were further queried with BioMart (Ensembl 54) for corresponding Ensembl genes. RefSeq IDs mapping to multiple EnsGenes were excluded. The second approach used BLAT to map the 50-mer probe sequences to Ensembl transcripts and to extract genomic locations matching for all 50 bases of the probe sequence. Probes with unique perfect match to the genome and corresponding transcripts matching to the same genes were kept. The union of the two mappings after excluding 196 conflictingly matching probes resulted in 27,499 probes corresponding to 18,170 autosomal genes available for association analysis.

Genotyping has been performed in parallel using Illumina's 1M-Duo and 1.2M-Duo custom chips on different subsets of individuals. Before further filtering, there were 106 samples with call rate (CR)$\geq$0.90 on the 1.2M and 88 samples with CR$\geq$0.90 on the 1M chip. Combined intensity files were created for Illuminus [26] by retaining on a per-chromosome basis only SNPs common to both chips. Additionally, any SNPs that moved position between the two chips were removed. Following further quality checks (Hardy-Weinberg $p>10^{-4}$, MAF>1%), 865,544 SNPs were kept for analysis.

The overlapping set of successfully genotyped samples with available expression data amounted to 156 (LCL), 160 (skin) and 166 (fat) individuals.

Post-experimental normalization of gene expression data

Log_2 - transformed expression signals were normalized separately per tissue as follows: quantile normalization was performed across the 3 replicates of each individual followed by quantile normalization across all individuals.

Genotype-gene expression associations and multiple testing correction

The eQTL analysis was done separately for each tissue. Within each tissue, twins from the same pair were separated by id in two samples analyzed independently. This separation resulted in the following sample size for LCL, skin and fat respectively: Twin 1 (74, 76, 79) and Twin 2 (82, 84, 87). Associations between SNP genotypes and normalized expression values were conducted using Spearman Rank Correlation (SRC). We considered only SNPs in *cis*, i.e. within a 1MB window from the TSS. We assess the statistical significance of the nominal associations using permutations as previously described [5]. We call an eQTL significant at 10^{-3} permutation threshold (PT) if the nominal association P-value is greater than the 0.001 tail of the minimal P-value distribution resulting from the SNP's associations with 10,000 permuted sets of expression values for each gene.

Factor analysis

We applied a Bayesian factor analysis model [25] to the expression data in each tissue. This approach uses an unsupervised linear model to account for global variance components in the data, and yields a residual expression dataset that can be used in further analysis.

We tested a wide range of parameter settings for the model, controlling the amount of variance explained by it. This was achieved by setting the parameters of the prior distributions for gene expression precision (inverse variance) and factor weight precision. These random variables are modelled using Gamma distributions, thus we varied their natural exponential family parameters - the prior mean and number of prior observations. We varied the prior mean from 10^{-6} to 10^{-2}, and number of prior observations from $N*10^{-3}$ to N, where N is the number of observations from data, and learned 120 latent factors. In the subsequent analysis, we used for each tissue the residual dataset that gave the best eQTL overlap between the two twin samples. The prior values used for each dataset are given in Table S7. The eQTL analysis on the corrected expression data was performed identically to the standard analysis: SRC followed by permutation testing.

Proportion of true positives from p-value distribution

For quantifying eQTL replication and tissue sharing in a continuous way, we used Storey's QVALUE software [23] (implemented in the R package qvalue 1.20.0, default recommended settings). The program takes a list of p-values and computes their estimated π_0 - the proportion of features that are truly null - based on their distribution (the assumption used is that p-values of truly alternative cases tend to be close to zero, while p-values of null features will be uniformly distributed among [0,1]). The quantity $\pi_1 = 1-\pi_0$ estimates the lower bound of the proportion of truly alternative features, i.e. the proportion of true positives (TP). Replication and sharing between two samples is reported as the proportion of TP (π_1) estimated from the p-value distribution of independent eQTLs discovered in sample 1 in the second sample (exact SNP-probe combinations are tested).

Recombination hotspot interval mapping and LD filtering

We refined the eQTL signals in order to characterize likely independent effects per gene. For this purpose, we mapped all common autosomal SNPs to recombination hotspot intervals as defined by McVean et.al [27]. We map significant eQTLs to recombination hotspot intervals and save the most significant SNP per gene. For each gene, SNPs resulting from this mapping are in addition filtered for LD in a pairwise manner (for each pair with $D'>0.5$ the least significant SNP is ignored). This filtering ensures that true shared effects (interval-gene combinations) are compared and not just genes.

Supporting Information

Figure S1 Median expression values of tissue-specific genes in the tissue of discovery and the other two tissues. Tissue-specific effects are not restricted to genes expressed in a tissue-specific manner.
10.1371/journal.pgen.1002003.s001(TIFF)

Figure S2 SNP×tissue interaction p-value from repeated measures ANOVA for all, shared and tissue-specific eQTLs respectively. Greater enrichment of significant SNP×tissue p-values is observed for tissue-restricted effects.
10.1371/journal.pgen.1002003.s002(TIFF)

Figure S3 eQTLs (10^{-2} PT, SRC) shared in all three tissues tested have the same direction of allelic effect (SRC rho) across tissues.
10.1371/journal.pgen.1002003.s003(TIFF)

Figure S4 Cumulative SRC rho distribution across tissues for tissue-specific and shared eQTLs (10^{-3} PT, Twin1). eQTLs discovered in one tissue only have distinctively higher variance in the tissue of discovery compared to shared effects.
10.1371/journal.pgen.1002003.s004(TIFF)

Figure S5 Most regulatory signals come from single independent eQTLs (SRC, 10^{-2} PT).
10.1371/journal.pgen.1002003.s005(TIFF)

Figure S6 Distribution of independent *cis* eQTLs (10^{-3} PT, SRC) around TSS, Twin 1.
10.1371/journal.pgen.1002003.s006(TIFF)

Figure S7 Distribution of independent *cis* eQTLs gained with FA correction (10^{-3} PT) around TSS, Twin 1.
10.1371/journal.pgen.1002003.s007(TIFF)

Figure S8 Fold change within twins and across tissues for SKIN eQTLs (10^{-3} PT, SRC) discovered in Twin 1. Fold change was calculated as the difference in mean expression of the heterozygous and major homozygous genotypic classes. For each pairwise tissue comparison, the Pearson's correlation coefficient between fold changes is shown.
10.1371/journal.pgen.1002003.s008(TIFF)

Figure S9 Fold change within twins and across tissues for FAT eQTLs (10^{-3} PT, SRC) discovered in Twin 1. Fold change was calculated as the difference in mean expression of the heterozygous and major homozygous genotypic classes. For each pairwise tissue comparison, the Pearson's correlation coefficient between fold changes is shown.
10.1371/journal.pgen.1002003.s009(TIFF)

Table S1 *Cis* eQTL associations with SRC and SRC-FA.
10.1371/journal.pgen.1002003.s010(DOC)

Table S2 eQTL recovery with FA. FA correction recovers the majority of eQTLs from the SRC analysis and adds twice as many discoveries.
10.1371/journal.pgen.1002003.s011(DOC)

Table S3 Tissue-shared and tissue-specific gene associations (10^{-3} PT), SRC-FA.
10.1371/journal.pgen.1002003.s012(DOCX)

Table S4 Tissue-shared and tissue-specific interval-gene associations (10^{-3} PT), SRC analysis.
10.1371/journal.pgen.1002003.s013(DOC)

Table S5 Tissue-shared and tissue-specific interval-gene associations (10^{-3} PT), SRC-FA.
10.1371/journal.pgen.1002003.s014(DOC)

Table S6 Continuous estimates of tissue sharing by enrichment of low p-values (π_1) of reference eQTLs (SNP-probes 10^{-3} PT) in the other two secondary tissues.
10.1371/journal.pgen.1002003.s015(DOC)

Table S7 FA weight and noise prior values used for each tissue.
10.1371/journal.pgen.1002003.s016(DOC)

Text S1 Biopsy technique protocol.
10.1371/journal.pgen.1002003.s017(DOC)

Author Contributions

Conceived and designed the experiments: SO KTZ RD KA PD MIM ETD TDS. Performed the experiments: DG JN AB MS MT CI. Analyzed the data: ACN LP. Contributed reagents/materials/analysis tools: EG KS ÅKH ASD TPY JLM SBM SP VB JTB GS FON PdM AW CJH NH CML NS IB. Wrote the paper: ACN ETD.

References

1. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777.
2. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772.
3. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.
4. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78. doi:10.1371/journal.pgen.0010078.
5. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224.
6. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.
7. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246–1250.
8. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nat Genet 39: 1494–1499.
9. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423–428.
10. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107. doi:10.1371/journal.pbio.0060107.
11. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208–1216.
12. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
13. Dermitzakis ET (2008) From gene expression to disease risk. Nat Genet 40: 492–493.
14. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, et al. (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452: 429–435.
15. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.
16. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 40: 1107–1112.
17. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet 6: e1000895. doi:10.1371/journal.pgen.1000895.
18. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6: e1000888. doi:10.1371/journal.pgen.1000888.

19. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 40: 768–775.
20. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140: 744–752.
21. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE (2010) Integrating Pathway Analysis and Genetics of Gene Expression for Genome-wide Association Studies. Am J Hum Genet.
22. Spector TD, Williams FM (2006) The UK Adult Twin Registry (TwinsUK). Twin Res Hum Genet 9: 899–906.
23. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.
24. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, et al. (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet 42: 62–67.
25. Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol 6: e1000770. doi:10.1371/journal.pcbi.1000770.
26. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, et al. (2007) A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 23: 2741–2746.
27. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304: 581–584.