Mapping cis- and trans-regulatory effects across multiple tissues in twins

NICA, Alexandra, *et al*.

Abstract

Sequence-based variation in gene expression is a key driver of disease risk. Common variants regulating expression in cis have been mapped in many expression quantitative trait locus (eQTL) studies, typically in single tissues from unrelated individuals. Here, we present a comprehensive analysis of gene expression across multiple tissues conducted in a large set of mono- and dizygotic twins that allows systematic dissection of genetic (cis and trans) and non-genetic effects on gene expression. Using identity-by-descent estimates, we show that at least 40% of the total heritable cis effect on expression cannot be accounted for by common cis variants, a finding that reveals the contribution of low-frequency and rare regulatory variants with respect to both transcriptional regulation and complex trait susceptibility. We show that a substantial proportion of gene expression heritability is trans to the structural gene, and we identify several replicating trans variants that act predominantly in a tissue-restricted manner and may regulate the transcription of many genes.

Reference

NICA, Alexandra, *et al*. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 2012, vol. 44, no. 10, p. 1084-9

PMID : 22941192
DOI : 10.1038/ng.2394

nature genetics

Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins

Elin Grundberg[1,2,20], Kerrin S Small[1,2,20], Åsa K Hedman[3,20], Alexandra C Nica[4,5,20], Alfonso Buil[4,5,20], Sarah Keildson[3], Jordana T Bell[2,3], Tsun-Po Yang[1], Eshwar Meduri[1,2], Amy Barrett[6], James Nisbett[1], Magdalena Sekowska[1], Alicja Wilk[1], So-Youn Shin[1], Daniel Glass[2], Mary Travers[6], Josine L Min[3], Sue Ring[7], Karen Ho[7], Gudmar Thorleifsson[8], Augustine Kong[8], Unnur Thorsteindottir[8,9], Chrysanthi Ainali[10], Antigone S Dimas[4,5], Neelam Hassanali[6], Catherine Ingle[1], David Knowles[11], Maria Krestyaninova[12], Christopher E Lowe[13,14], Paola Di Meglio[15], Stephen B Montgomery[4,5,19], Leopold Parts[1], Simon Potter[1], Gabriela Surdulescu[2], Loukia Tsaprouni[1], Sophia Tsoka[10], Veronique Bataille[2], Richard Durbin[1], Frank O Nestle[15], Stephen O'Rahilly[13,14], Nicole Soranzo[1], Cecilia M Lindgren[3], Krina T Zondervan[3], Kourosh R Ahmadi[2], Eric E Schadt[16], Kari Stefansson[8,9], George Davey Smith[7], Mark I McCarthy[3,6,17], Panos Deloukas[1], Emmanouil T Dermitzakis[4,5] & Tim D Spector[2], The Multiple Tissue Human Expression Resource (MuTHER) Consortium[18]

Sequence-based variation in gene expression is a key driver of disease risk. Common variants regulating expression in *cis* have been mapped in many expression quantitative trait locus (eQTL) studies, typically in single tissues from unrelated individuals. Here, we present a comprehensive analysis of gene expression across multiple tissues conducted in a large set of mono- and dizygotic twins that allows systematic dissection of genetic (*cis* and *trans*) and non-genetic effects on gene expression. Using identity-by-descent estimates, we show that at least 40% of the total heritable *cis* effect on expression cannot be accounted for by common *cis* variants, a finding that reveals the contribution of low-frequency and rare regulatory variants with respect to both transcriptional regulation and complex trait susceptibility. We show that a substantial proportion of gene expression heritability is *trans* to the structural gene, and we identify several replicating *trans* variants that act predominantly in a tissue-restricted manner and may regulate the transcription of many genes.

The risk of developing common complex diseases, such as type 2 diabetes and obesity, involves multiple genetic and environmental factors. Genome-wide association studies (GWAS) have been successful in identifying common genetic variants associated with these complex human diseases. A suggested approach for finding additional genetic components is to focus on low-frequency and rare variants[1]. In parallel, approaches to disentangle the underlying molecular mechanisms for identified disease loci are also needed. As the majority of the common genetic variants that are associated with complex traits map to non-coding regions and may thus alter gene regulation, the use of gene expression data integrated with sequence variation—eQTL studies—is a commonly applied approach using various cell[2–5] and tissue samples[6–8]. More recently, we and others have been able to collect multiple cells and/or tissues from the same individuals[9–12], showing the degree of tissue dependency of *cis*-regulatory effects[11,13]. Tissue dependency seems to be an important feature of disease susceptibility variants that regulate gene expression[11,14], promoting the use of multiple disease-targeted cell types in future large-scale eQTL studies. However, despite success in mapping common *cis*-regulatory variants in these studies[8,15,16], *trans* variants have been more difficult to

[1]Wellcome Trust Sanger Institute, Hinxton, UK. [2]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [5]Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland. [6]Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [7]Medical Research Council (MRC) Centre for Causal Analyses in Translational Epidemiology (CAiTE), School of Social and Community Medicine, University of Bristol, Bristol, UK. [8]deCODE genetics, Reykjavik, Iceland. [9]Faculty of Medicine, University of Iceland, Reykjavík, Iceland. [10]Department of Informatics, School of Natural and Mathematical Sciences, King's College London, London, UK. [11]Department of Engineering, University of Cambridge, Cambridge, UK. [12]European Bioinformatics Institute, Hinxton, UK. [13]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK. [14]Cambridge National Institute for Health Research (NIHR) Biomedical Research Centre, Addenbrooke's Hospital, Cambridge, UK. [15]St. John's Institute of Dermatology, King's College London, London, UK. [16]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA. [17]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, UK. [18]A full list of members and affiliations is provided in the **Supplementary Note**. [19]Present address: Department of Pathology, Stanford University School of Medicine, Stanford, California, USA, and Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [20]These authors contributed equally to this work. Correspondence should be addressed to M.I.M. (mark.mccarthy@drl.ox.ac.uk), P.D. (panos@sanger.ac.uk), E.T.D. (emmanouil.dermitzakis@unige.ch) or T.D.S. (tim.spector@kcl.ac.uk).

Received 4 January; accepted 8 August; published online 2 September 2012; doi:10.1038/ng.2394

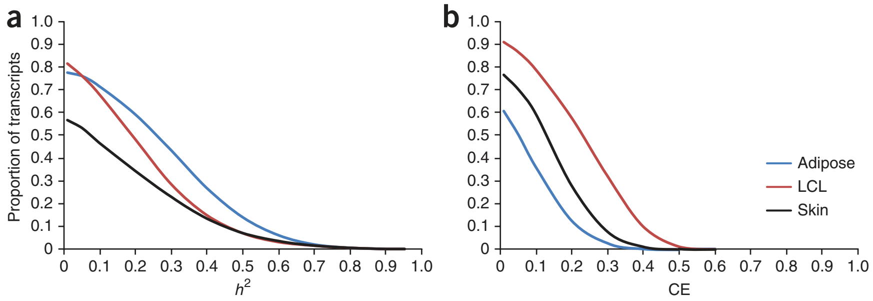

Figure 1 Genetic and non-genetic effect of gene expression across multiple tissues. (**a**,**b**) Estimation of the proportion of variation in expressed transcripts in adipose (N = 11,394), LCLs (N = 10,631) and skin (N = 11,932) that is attributable to genetic (narrow-sense heritability, h^2) (**a**) and familial non-genetic factors (shared common environment, CE) (**b**). The y axis shows the proportion of transcripts at the h^2 or CE cutoff indicated on the x axis.

map, mainly due to small effect sizes, emphasizing the need for well-powered studies and thorough replication efforts in multiple tissues.

To this end, we designed the MuTHER project (Multiple Tissue Human Expression Resource) to develop a major resource of detailed genomic and transcriptome data from three disease-relevant tissues (adipose, lymphoblastoid cell lines (LCLs) and skin) originating from a cohort of 856 deeply phenotyped twins (one-third monozygotic and two-thirds dizygotic from the TwinsUK adult registry). The increased sample size compared to our pilot study[11] allows us to use the classical twin design for systematic dissection of genetic (*cis* and *trans*) effects on gene expression, providing for the first time (i) estimates of additional heritable *cis* effects unexplained by common SNPs (with minor allele frequency (MAF) of >5%) identified in standard *cis*-eQTL analysis and (ii) in-depth characterization of the architecture of *trans* regulation of gene expression highlighted by thorough replication efforts in multiple independent data sets. In addition, the large-scale multitissue design of our study also allows us to provide the most precise estimates to date of, not only gene expression heritability, but also the degree of tissue dependency of eQTL function. The relevance of tissue-dependent eQTLs in complex trait susceptibility is further highlighted, as we identify for hundreds of known GWAS SNPs the candidate causal eQTLs, with good correlation of the phenotype to the candidate tissue.

RESULTS

Data structure

In total, 856 female twins (154 monozygotic twin pairs, 232 dizygotic twin pairs and 84 singletons) aged 38.7–84.6 years were recruited from the TwinsUK resource[17], and adipose (subcutaneous fat) and skin tissue biopsies, as well as peripheral blood samples (for generation of LCLs), were collected for subsequent genome-wide expression profiling. The TwinsUK cohort has previously been shown to be comparable to population singletons in terms of disease-related and lifestyle characteristics[18]. Cohort characteristics are presented in **Supplementary Table 1**.

Genetic and non-genetic effects on gene expression

We estimated narrow-sense heritability, h^2, for each transcript across the three tissues in all available twin pairs using a variance component model, adjusting for known technical cofactors[19]. The average h^2 estimates of expressed transcripts corresponded to $h^2_{\text{adipose}} = 0.26$, $h^2_{\text{LCL}} = 0.21$ and $h^2_{\text{skin}} = 0.16$ (**Fig. 1** and **Supplementary Table 2**). The cell type heterogeneity expected in skin probably explains the lower estimates in that tissue. We tested how heritability of transcripts compares across tissues and found that approximately 50% of the top 5,000 heritable transcripts in each tissue (corresponding to $h^2_{\text{adipose}} > 0.33$, $h^2_{\text{LCL}} > 0.27$ and $h^2_{\text{skin}} > 0.22$) are in fact heritable across 2 or more tissues (**Supplementary Fig. 1a**), with similar results obtained when we restricted analysis to transcripts expressed in all 3 tissues (**Supplementary Fig. 1b**).

Twin studies also allow calculation of the proportion of phenotypic variation attributable to familial non-genetic factors, meaning the shared common environment. Notably, we found that as much as 32% of expressed LCL transcripts had a common environmental component that explained over 30% of the total variance, compared to 2% and 8% in adipose and skin tissue, respectively (**Fig. 1b**). This larger shared environmental effect in LCLs most likely reflects the impact of additional correlated sample handling steps not applicable for tissue biopsies, such as blood sampling, cell isolation, Epstein-Barr virus (EBV)-mediated transformation and cell culture procedures, as the study subjects visited the clinic in pairs.

Large-scale *cis* eQTL mapping

To map the underlying common, genetic effect of transcript levels, we performed global *cis* eQTL mapping, associating the 23,596 expression traits with imputed HapMap 2 genotypes in a linear mixed (polygenic) model and then performed a score test taking relatedness into account (**Supplementary Fig. 2**). *cis* eQTLs were called with a per-tissue false discovery rate (FDR) of 1%, which corresponds to $P < 5.0 \times 10^{-5}$ in adipose, $P < 7.8 \times 10^{-5}$ in LCLs and $P < 3.8 \times 10^{-5}$ in skin. Across all transcripts, we detected an abundance of *cis* eQTLs per tissue (N_{adipose} = 3,529, N_{LCL} = 4,625 and N_{skin} = 2,796; **Supplementary Table 3**), with 14%, 17% and 10% of transcripts with a *cis* eQTL in adipose, LCL and skin tissue, respectively, having more than 1 independent *cis* eQTL. For these transcripts associated with at least one *cis* variant, the average h^2 estimates were 0.31, 0.25 and 0.21 in adipose, LCLs and skin, respectively. The probability of detecting *cis* eQTLs of large effect size across tissues increased with heritability, as shown by average h^2 estimates of transcripts associated with a *cis* variant at $P < 5 \times 10^{-8}$, whereas the maximum average h^2 seen in adipose tissue was 0.38.

We validated identified *cis*-regulatory effects by performing replication studies in independent expression data sets (**Supplementary Table 4**). Using the list of replication P values from the different data sets, we first estimated π_0, which is the overall proportion of true null hypotheses among all tests performed. We could then quantify the proportion of significant replicated *cis* results in each study, π_1, corresponding to $\pi_1 \equiv 1 - \pi_0$ (ref. 20) and noted a high replication rate of *cis* eQTLs across studies of similar size (π_1 = 0.70–0.76) (**Supplementary Fig. 3**).

Most previous efforts to examine tissue dependency of *cis*-eQTL effects have only used a P-value threshold, but this has obvious limitations. Here, we employed several complementary approaches in addition to the threshold-based approach to address this question. First, we assessed tissue dependency by studying shared effects at 1% FDR and found substantial tissue independence of *cis* eQTLs (**Table 1**). For instance, 47% of *cis* eQTLs identified at 1% FDR in adipose tissue were identified in at least one other tissue, and as many as 22% were seen across all three tissues at a similar FDR threshold. This degree of overlap was further confirmed by estimating the proportion of

Table 1 Estimated degree of tissue overlap of *cis* effects (1% FDR)

		Approach 1	Approach 2	
Reference tissue	Secondary tissue	*N* (%)	Twin 1 π_1	Twin 2 π_1
Adipose	LCL	1,221 (34.6)	0.64	0.56
	Skin	1,207 (34.3)	0.79	0.77
	LCL and skin	767 (21.8)	–	–
	LCL or skin	1,661 (47.1)	–	–
LCL	Adipose	1,118 (24.2)	0.65	0.63
	Skin	978 (21.1)	0.65	0.61
	Adipose and skin	728 (15.7)	–	–
	Adipose or skin	1,368 (29.6)	–	–
Skin	Adipose	1,265 (45.2)	0.77	0.83
	LCL	1,104 (39.5)	0.64	0.64
	Adipose and LCL	790 (28.3)	–	–
	Adipose or LCL	1,579 (56.5)	–	–

Two approaches were used, including a threshold-based approach (approach 1) and a matched co-twin design comparing *P*-value distributions across tissues, where π_1 represents the proportion of true positives (approach 2).

significant results across tissues (π_1 = 0.5–0.7; **Supplementary Fig. 4**). As with previous smaller studies[9–11,13], tissue-independent *cis* eQTLs had larger effect sizes and were over-represented close to transcription start sites (TSSs) compared to tissue-dependent effects (**Supplementary Fig. 5**). In general, *cis* effects that were located less than 200 kb from TSSs explained a larger proportion of the variance in expression levels ($r^2_{\text{average adipose}}$ = 0.07, $r^2_{\text{average LCL}}$ = 0.08 and $r^2_{\text{average skin}}$ = 0.08) than long-range effects (located >200 kb from TSSs) ($r^2_{\text{average adipose}}$ = 0.04, $r^2_{\text{average LCL}}$ = 0.04 and $r^2_{\text{average skin}}$ = 0.04) (**Supplementary Fig. 6**).

We then characterized tissue dependency of regulatory effects in more detail using a matched co-twin design, as previously described[11], comparing the *P*-value distribution of significant SNP–expression probe pairs within and across tissues (**Supplementary Fig. 7**). We found that 56–83% of *cis* effects were shared across tissues, with adipose and skin sharing more with each other than with LCLs (**Table 1**). eQTLs with statistical significance in multiple tissues might still have tissue-dependent biological consequences if they have different effect sizes (fold change in expression) across tissues. Thus, we also evaluated tissue dependency by contrasting fold changes in expression between tissues and estimating the predictive value (r^2) of each tissue for the other two (**Supplementary Fig. 8**). After accounting for winner's curse (subtracted unexplained intratissue variance), we estimate that 41–62% of *cis* eQTLs are not only tissue independent but also have a similar magnitude of effect in multiple tissues. Considering together data generated using several complementary approaches, we find evidence that >60% of *cis* eQTLs have statistically significant effects in multiple tissues.

Dissection of the contribution of *cis* effects to heritability of gene expression

To estimate the proportion of the heritability of each transcript that is driven in *cis* by alleles of high frequency (common SNPs, defined here as SNPs with MAF of > 5%), we combined the results from our heritability and *cis*-eQTL analyses. As the current sample size was not sufficient to obtain reliable h^2 estimates of less than 0.1, we focused on transcripts with h^2 of >0.1, which corresponds to 10,027 (43%), 10,219 (44%) and 7,511 (32%) transcripts in adipose, LCLs and skin, respectively (**Fig. 1** and **Supplementary Table 2**).

Overall, when taking all transcripts into account, we found that common *cis* SNPs (MAF > 5%) explained on average only 9% (adipose), 12% (LCLs) and 10% (skin) of the total genetic variance at each locus (**Fig. 2**). Less than a third (27% (adipose), 33% (LCLs) and 21% (skin)) of the transcripts were in fact associated with a *cis* variant at 1% FDR; therefore, when focusing on these transcripts only and taking independent *cis* effects into account, the *cis* component accounted for a greater proportion of the genetic variance, namely 25% (adipose), 31% (LCLs) and 32% (skin) on average. Notably, the effect of common *cis* variants increased as heritability increased. If we filtered for transcripts that were highly heritable in all tissues (h^2 > 0.6 across all tissues, N = 24), ~95% of these transcripts were found to be associated with a *cis* variant, and, at 18 of the 24, a single *cis* SNP explained over 50% of the genetic variance (**Supplementary Table 5**).

These results from multiple tissues indicate that a large proportion of the heritability of gene expression remains unexplained by the common SNPs (MAF > 5%) analyzed in standard *cis*-eQTL analyses. Thus, we asked whether there are other genetic *cis* effects that account for the additional genetic variance in gene expression.

We therefore performed quantitative linkage analysis in the *cis* regions of transcripts with h^2 of >0.1 that were associated with a common *cis* SNP, using a global regression approach that analyzes the data of all expression traits (N_{adipose} = 2,537, N_{LCL} = 3,157 and N_{skin} = 1,493) in a single linear regression. We phased all SNPs in each *cis* region (~2 Mb) and counted the haplotypes with shared identity by descent (IBD) for all dizygotic twin pairs. We then estimated the average heritability at each *cis* region, using the global regression approach based on the Haseman-Elston algorithm, but taking all selected transcripts into account. We noted that, on average, 30% (adipose), 35% (LCL) and 36% (skin) of the total genetic variance was explained by variants in *cis*, which was in fact 40% more than if only common *cis* SNPs identified from *cis*-eQTL analysis were included (**Table 2**). This added genetic component is likely due to low-frequency and/or rare *cis* variants. However, the estimates of its magnitude should be considered as a lower bound, as our sample size limited our ability to conduct the analyses on a subset of the heritable transcripts.

Integration of *cis*-eQTL data with disease loci

A major application of eQTL data has been the functional annotation of loci identified in GWAS. We investigated the regulatory impact of GWAS variants by integrating *cis* eQTLs (1% FDR; **Supplementary Table 6**) and disease-associated SNPs (National Human Genome Research Institute (NHGRI) database, accessed 21 December 2010), using regulatory trait concordance (RTC) methodology as previously

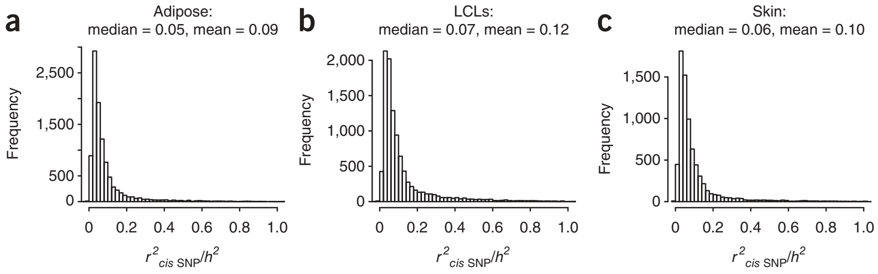

Figure 2 Contribution of heritable *cis* components to gene expression variation. (**a**–**c**) The contribution of individual *cis* SNPs (MAF > 5%) to the genetic variance of gene expression (h^2 > 0.1) in adipose (**a**), LCLs (**b**) and skin (**c**). The *x* axis shows the proportion of the heritability of each transcript that is explained by independent *cis* SNPs.

Table 2 Proportion of the heritability explained by *cis* effects

Tissue	*N* transcripts	Average h^2	h^2_{cis}	h^2_{cis}/h^2	h^2_{SNP}	h^2_{SNP}/h^2	$h^2_{\mathrm{SNP}}/h^2_{cis}$
Adipose	2,537	0.40	0.12	0.30	0.072	0.18	0.60
LCL	3,157	0.34	0.12	0.35	0.095	0.28	0.79
Skin	1,497	0.36	0.13	0.36	0.094	0.26	0.72

h^2_{cis} corresponds to the average heritability estimate at the *cis* regions, and h^2_{SNP} corresponds to the average heritability estimate at the *cis* regions that is due to common SNPs.

described[14]. RTC scores of ≥0.9 indicate that overlapping eQTL and GWAS signals likely tag the same functional variant. In all three tissues, we observed an over-representation of high-RTC-scoring candidates, which suggests that disease effects are mediated through changes to gene expression (**Supplementary Fig. 9**). Of the total number of interval-disease combinations tested in adipose ($N = 765$), LCLs ($N = 887$) and skin ($N = 639$), we detected 181 (23.7%), 225 (25.4%) and 145 (22.7%) signals with RTC of ≥0.9 in each tissue, respectively, more than twice the number expected by chance ($P_{\mathrm{adipose}} = 0.0009$, $P_{\mathrm{LCL}} = 0.008$ and $P_{\mathrm{skin}} = 0.0009$). Disease-associated eQTL candidates are largely tissue dependent (~60%), in line with the estimated proportion of tissue-dependent *cis* effects (52 of the total nonredundant 358 RTC signals were discovered across all 3 tissues; **Supplementary Table 6**).

Our efforts in integrating eQTL data with disease associations suggest that the ability to interpret the functionality of GWAS loci is highly dependent on the tissue where gene expression is interrogated and the tissue's relevance to the trait of interest. Indeed, we observed a significant enrichment of immunity-related GWAS signals among high-RTC-scoring *cis* eQTLs in LCLs ($P = 6.6 \times 10^{-5}$, Fisher's exact test), much more so than in the other two tissues ($P_{\mathrm{adipose}} = 0.003$ and $P_{\mathrm{skin}} = 0.013$). Likewise, disease-associated eQTLs detected in adipose and skin samples explained associations with biologically relevant traits (**Supplementary Table 7**). For example, in adipose, we discovered regulatory effects that potentially explained associations with triglyceride concentrations (rs2304130 (ref. 21) on *ATP13A1* and rs439401 (ref. 21) on *APOE*) or birth weight (rs900400 (ref. 22) on *TIPARP*), whereas, in skin, associations with melanoma (rs910873 (ref. 23) on *ASIP*) and skin sensitivity to sun (rs1805007 (ref. 24) on *DBNDD1*) stood out.

Trans regulation of gene expression across tissues

Although low-frequency and/or rare *cis* variants seem to contribute significantly to the total heritable *cis* effect, a large proportion of the total heritability (>60%) of gene expression is still unexplained, indicating that the effects of *trans*-regulatory variants on gene expression are likely to be critical to interindividual differences in gene expression. We thus proceeded to explore the *trans*-regulatory landscape across tissues. Given the large number of tests performed and the relatively small effect sizes of *trans* eQTLs, we chose the GWAS threshold of $P < 5 \times 10^{-8}$ (corresponding to an FDR of less than 10%) to select possible candidates for further investigation, including replication analysis in independent samples. At $P < 5 \times 10^{-8}$, we found 639, 557 and 609 *trans* eQTLs in adipose, LCLs and skin, respectively (**Supplementary Table 8**). The relative proportion of *trans* eQTLs per tissue is the inverse of that seen in the *cis*-eQTL analysis, perhaps reflecting the different external environments present for complex tissues versus cultured cells. In contrast to the *cis* results, nearly all *trans* eQTLs seemed to be tissue dependent, had relatively small effect sizes and were associated with transcripts with lower average h^2 values ($h^2_{\mathrm{adipose}} = 0.19$, $h^2_{\mathrm{LCL}} = 0.18$ and $h^2_{\mathrm{skin}} = 0.13$).

Notably, many *trans* SNPs at $P < 5 \times 10^{-8}$ were associated with multiple transcripts, suggesting that they are multigene regulators. In adipose tissue, 48 SNPs accounted for 169 (32%) of the *trans* eQTLs, and, in LCLs and skin tissue, 48 SNPs accounted for 121 (21%) and 44 SNPs for 164 (27%) of the *trans* eQTLs, respectively. These multigene regulators (defined here as *trans* SNPs associated with at least two distinct transcripts at $P < 5 \times 10^{-8}$) consistently showed enrichment for additional *trans* associations with low *P* values beneath the 5×10^{-8} threshold (**Fig. 3a**), indicating that *trans* SNPs may regulate additional genes below our *P*-value threshold. In contrast, the *P*-value distribution across all measured transcripts in the other two tissues approximated the distribution expected under a null hypothesis of no enrichment in *trans* associations (**Fig. 3**). To quantify the genome-wide effect of these *trans* SNPs, we again used π_1 for estimation of the proportion of true positives in

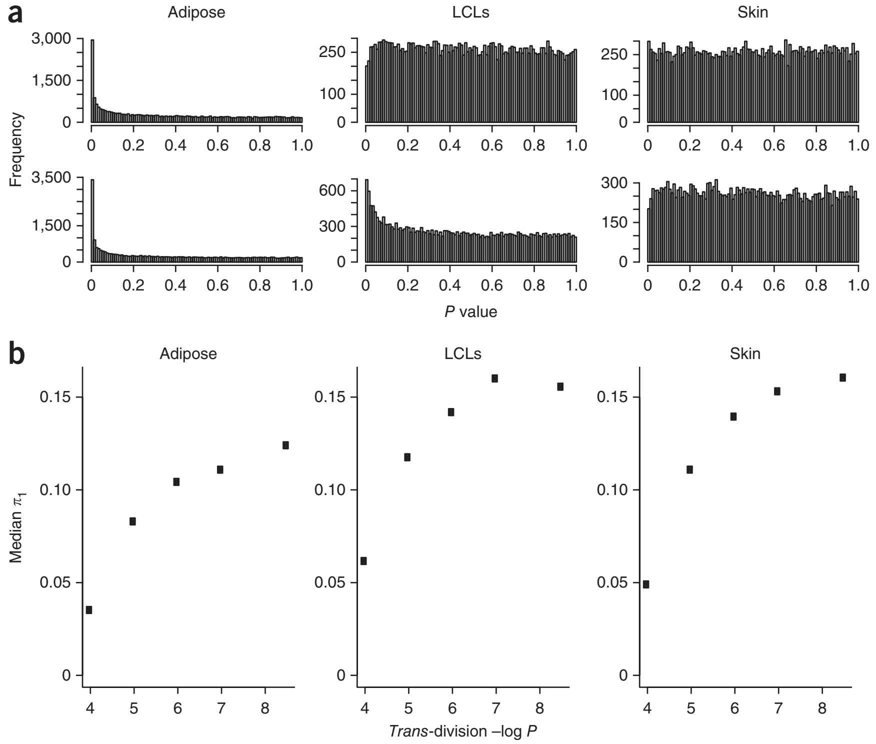

Figure 3 *Trans* variants regulating expression of multiple transcripts. (**a**) *P*-value distributions (*x* axis) of genome-wide associations (count *N*, *y* axis) between two potential adipose multigene regulators (rs1752223 on chromosome 1 (top) and rs7595947 on chromosome 2 (bottom)) and transcript levels in adipose (left), LCLs (middle) and skin (right). (**b**) Plots showing the median *trans* π_1 value (π_1 calculated from the *P*-value distribution of a *trans* SNP versus all probes) at increasing levels of *trans*-SNP significance in adipose (left), LCLs (middle) and skin (right). The top *trans* SNP for each probe was included, and *trans* SNPs were divided into nonoverlapping bins on the basis of the *P* value of the top *trans* association. The median π_1 value (*y* axis) is plotted against the –log *P* value (*x* axis) of the lower limit of each bin.

Table 3 Replication in independent cohorts of *cis* and *trans* associations

Tissue	Cohort	N	*Trans* hits tested/total hits at $P < 5 \times 10^{-8}$	$P < 0.05$[a]	π_1	*Cis* hits tested/total hits at 1% FDR	$P < 0.05$[a]	π_1
Adipose (subcutaneous)	deCODE	585	586/639	25/586 (4.3%)	0.10	N/A	N/A	N/A
Adipose (subcutaneous)	MGH	701	514/639	27/514 (5.3%)	0.096	2,980/3,332	1,751/2,980 (59%)	0.79
LCLs	ALSPAC	931	544/557	33/544 (6.1%)	0.13	6,181/6,289	4,154/6,181 (67%)	0.76
LCLs	Oxford (TwinsUK)	331	361/557	23/361 (6.4%)	0.13	4,608/6,289	2,745/4,608 (60%)	0.75
Skin (fibroblasts)	GenCORD	68	442/609	15/442 (3.4%)	0.00024	2,241/3,416	455/2,241 (20%)	0.34

MGH, Massachusetts General Hospital. N/A, data not available

[a]Consistent direction.

the distribution of P values from each *trans* SNP versus all transcripts and compared it with similar calculations for *trans* SNPs associated with only one transcript at $P < 5 \times 10^{-8}$ (single-gene regulators). We found that multigene regulators were enriched for greater numbers of true positives compared to single-gene regulators (**Supplementary Fig. 10** and **Supplementary Table 9**). We further investigated π_1 values at *trans* SNPs beyond our threshold of $P < 5 \times 10^{-8}$ and found that the median π_1 increased with increasing significance of the top association per *trans* SNP (**Fig. 3b**). As *trans* SNPs with more significant association should be enriched for true positives, this confirms that a general property of true *trans* SNPs might be regulation of multiple transcripts.

We then sought to study the genome-wide regulatory behavior of our LCL *trans* SNPs ($P < 5 \times 10^{-8}$), using the calculated π_1 values in our replication cohorts (The Avon Longitudinal Study of Parents and their Children (ALSPAC) and Oxford-TwinsUK (Oxford)) (**Supplementary Table 4**). In total, 314 *trans* SNPs with π_1 of ≥0.10 in the MuTHER discovery cohort were tested, of which 61 (19%) and 43 (14%) also had π_1 estimates of ≥0.10 in the ALSPAC and Oxford LCL replication cohorts, respectively. When comparing not only the π_1 values in the replication cohorts but also the top 1,000 associated transcripts to the bottom 1,000 associated transcripts for each of the 61 and 43 *trans* SNPs in the ALSPAC and Oxford replications studies, respectively, we found a highly significant enrichment for lower P values in the top 1,000 transcripts (Mann-Whitney; $P_{\mathrm{ALSPAC}} = 2.6 \times 10^{-6}$ and $P_{\mathrm{Oxford}} = 3.8 \times 10^{-5}$). This indicates that, not only genome-wide regulatory behavior, but also the ranking of associated genes for a subset of *trans* SNPs is consistent across studies, but larger sample sizes are needed to confirm the observed effect of gene regulation.

We also performed replication studies of each *trans* association identified at $P < 5 \times 10^{-8}$ in independent data sets (**Supplementary Table 4**) and noted that the replication rate of *trans* associations was markedly lower compared to that of *cis* effects (**Table 3** and **Supplementary Fig. 3**), with π_1 estimates ranging from 0.0002 to 0.13 compared to π_1 of 0.34 to 0.76 for replication of *cis* effects. However, using a P-value cutoff of <0.05 and taking direction of effect into account, we found up to threefold enrichment of replicated *trans* eQTLs (**Table 3** and **Supplementary Table 10**). Taken together, these data show that *trans*-regulatory effects of gene expression are highly complex, with small effect sizes indicating that sample sizes are required to be larger than previously expected.

DISCUSSION

We undertook a large-scale genetic association study of human gene expression traits in multiple disease-targeted tissue samples (subcutaneous fat, LCLs and whole skin) derived from 856 mono- and dizygotic female twins, as part of the MuTHER project. This is the first study performed to date using the twin design for the dissection of genetic and non-genetic components underlying population differences in tissue-independent and tissue-dependent expression profiles.

A study using family data sets aiming to partition the heritability of gene expression into *cis* and *trans* components recently estimated that 37% of the heritability in blood and 24% in adipose tissue are in fact due to *cis* regulation[16]. Here, we confirm these estimates but decompose the *cis* component further using IBD estimates in our dizygotic subjects. We found that 30–36% of heritability is due to *cis* components but that up to 40% of the heritable *cis* effect or 12% of total heritability is missed when only considering common SNPs from *cis*-eQTL mapping. However, as our analyses were conducted on heritable transcripts in each tissue for which we observed a significant *cis* association from the *cis*-eQTL mapping approach, the estimate of the contribution of undetected regulatory effects to *cis* genetic variance is most likely an underestimate. Although we acknowledge that common SNPs may in some instances tag low-frequency variants[25,26], we expect that a subset of the missing *cis* heritability still will be accounted for by low-frequency and rare variants, supporting the development of large-scale exome and genome resequencing initiatives for complex trait mapping. The missing *cis* heritability also has implications for GWAS signals in cases where the effect of the lead SNP is mediated via a *cis* eQTL; if a known GWAS variant is an eQTL and therefore affects disease risk by modulating expression of a gene, then any additional rare variant modulating expression of the same gene in the same tissue should also affect the same trait. This leads us to predict that, on average, an additional 40% or more of signal remains to be discovered at *cis*-eQTL GWAS loci (of which we identify 358 in this study). These estimates are based on calculations within each tissue and thus do not represent tissue-independent *cis* heritability. In agreement with a previous study[27], we do not find an enrichment of genetic correlations unequal to zero (data not shown), which is expected given the high degree of tissue independence of *cis* effects seen in our *cis*-eQTL mapping approach. However, as our sample size is limited, measurement error in genetic covariance cannot be ruled out. The finding that the majority (>60%) of the genetic effect of expression traits is regulated by components other than those acting in *cis* indicates the need for studies of the *trans*-regulatory landscape. *Trans*-regulatory variants are known to have small effect sizes and thus have previously been difficult to map, given limited sample sizes and lack of appropriate replication studies[4]. However, the recent findings of disease-related *trans* variants regulating the expression of multiple genes are promising[28,29]. A dilemma with genome-wide eQTL analysis is that only a small proportion of the variants survive multiple testing corrections and that, by restricting analysis to signals solely on the basis of arbitrary cutoffs, many true hits are likely to be missed. This can be circumvented by analytical methods, such as studying the global effect of *trans* variants using the proportion of true positives (π_1), as presented here. By applying this approach and with thorough replication, we found evidence of multiple *trans* variants acting as multigene regulators, predominantly in a tissue-dependent manner, similar to our previously reported example of the *KLF14* locus in adipose tissue[29]. For instance, the rs7595947 SNP on chromosome 2 was associated with 27 transcripts in the MuTHER adipose samples and was successfully replicated in independent cohorts. In skin, the rs1215608 *trans* SNP located within the *NUAK1* gene, defined as a multigene regulator and associated with three genes (*FMO6P*, *PPM1F* and *LECT1*) in the MuTHER discovery sample, was successfully replicated in the fibroblast cohort. Notably, the rs1215608 SNP is also a *cis*-acting SNP that regulates *NUAK1* expression. The *NUAK1* gene was recently identified as a key player in cellular senescence and cellular ploidy, mechanisms that are known to be important in aging[30]. These examples underscore the potential in using full-transcriptome architecture to understand biology. However, as shown here by the relatively low replication rate of *trans* SNPs, the dissection of *trans* effects and their characteristics, such as tissue dependency, are indeed challenging, as they are highly complex and require larger sample sizes to be discovered than was previously expected.

In conclusion, we present unique twin data using thousands of eQTLs in multiple tissues, extending understanding of the architecture and regulation of gene expression in multiple ways. We highlight the importance of studying low-frequency and rare regulatory variants in complex traits by detecting and mapping missing heritability of gene expression beyond the common *cis* variants. We also show that a substantial proportion of gene expression heritability is *trans* to structural genes and identify several replicating *trans* variants that seem to act predominantly in a tissue-restricted manner and are potentially regulators of many genes.

URLs. MuTHER Resource, http://www.muther.ac.uk/; ArrayExpress, http://www.ebi.ac.uk/arrayexpress/; Genevar database, http://www.sanger.ac.uk/resources/software/genevar/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Microarray data have been deposited in the ArrayExpress archive under accession E-TABM-1140, and the full data set of MuTHER *cis* eQTLs is available through the Genevar database.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The MuTHER Study was funded by a program grant from the Wellcome Trust (081917/Z/07/Z) and by core funding for the Wellcome Trust Centre for Human Genetics (090532). Additional funding came from the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project and grant agreement HEALTH-F4-2007-201413, the Swiss National Science Foundation and the NCCR Frontiers in Genetics, the Louis-Jeantet Foundation and a US National Institutes of Health–NIMH grant (GTEx project). Additional details on the funding for the participating studies and investigators are provided in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

K.R.A., M.I.M., P.D., E.T.D. and T.D.S. conceived the study. E.G., K.S.S., Å.K.H., A.C.N., A. Buil and S.K. analyzed data. T.-P.Y., E.M., S.-Y.S., J.L.M., K.T.Z., S.R., K.H., G.T., A.K., U.T., S.P., N.S., E.E.S., K.S. and G.D.S. contributed reagents, materials, or analysis tools. A. Barrett, J.N., M.S., A.W., D.G., M.T., N.H., C.I., M.K. and G.S. performed wet lab experiments or collected samples. J.T.B., C.A., A.S.D., D.K., C.E.L., P.D.M., S.B.M., L.P., L.T., S.T., V.B., R.D., F.O.N., S.O. and C.M.L. contributed experimental and technical support as well as discussion. E.G. prepared the manuscript, with contributions from K.S.S., Å.K.H., A.C.N., A. Buil, M.I.M., P.D., E.T.D. and T.D.S. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2394.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Eichler, E.E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
2. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
3. Göring, H.H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208–1216 (2007).
4. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19**, 1942–1952 (2009).
5. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
6. Myers, A.J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39**, 1494–1499 (2007).
7. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
8. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
9. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246–1250 (2009).
10. Greenawalt, D.M. *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* **21**, 1008–1016 (2011).
11. Nica, A.C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003 (2011).
12. Zeller, T. *et al.* Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS ONE* **5**, e10693 (2010).
13. Ding, J. *et al.* Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in *cis*-eQTL signals. *Am. J. Hum. Genet.* **87**, 779–789 (2010).
14. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
15. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39**, 1202–1207 (2007).
16. Price, A.L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).
17. Spector, T.D. & Williams, F.M. The UK Adult Twin Registry (TwinsUK). *Twin Res. Hum. Genet.* **9**, 899–906 (2006).
18. Andrew, T. *et al.* Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin Res.* **4**, 464–477 (2001).
19. Visscher, P.M., Benyamin, B. & White, I. The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. *Twin Res.* **7**, 670–674 (2004).
20. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
21. Aulchenko, Y.S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.* **41**, 47–55 (2009).
22. Freathy, R.M. *et al.* Variants in *ADCY5* and near *CCNL1* are associated with fetal growth and birth weight. *Nat. Genet.* **42**, 430–435 (2010).
23. Brown, K.M. *et al.* Common sequence variants on 20q11.22 confer melanoma susceptibility. *Nat. Genet.* **40**, 838–840 (2008).
24. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **39**, 1443–1452 (2007).
25. Anderson, C.A., Soranzo, N., Zeggini, E. & Barrett, J.C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.* **9**, e1000580 (2011).
26. Dickson, S.P. *et al.* Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
27. Powell, J.E. *et al.* Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* **22**, 456–466 (2012).
28. Heinig, M. *et al.* A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* **467**, 460–464 (2010).
29. Small, K.S. *et al.* Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011).
30. Humbert, N. *et al.* Regulation of ploidy and senescence by the AMPK-related kinase NUAK1. *EMBO J.* **29**, 376–386 (2010).

ONLINE METHODS

Sample collection. The study included 856 female individuals of European descent recruited from the TwinsUK Adult twin registry[17] (**Supplementary Table 1**). Punch biopsies (8 mm) were taken from a photo-protected area adjacent and inferior to the umbilicus. Subcutaneous adipose tissue was dissected from each biopsy, weighed and immediately stored in liquid nitrogen. Similarly, the remaining skin tissue was weighed and stored in liquid nitrogen. Peripheral blood samples were collected, and LCLs were generated through EBV-mediated transformation of the B-lymphocyte component by the European Collection of Cell Cultures agency. The project was approved by the local ethics committees of all institutions involved, and all samples were collected after obtaining written and signed informed consent.

RNA extraction. RNA was extracted from homogenized adipose and skin samples and lysed LCLs using TRIzol Reagent (Invitrogen) according to the protocol provided by the manufacturer. RNA quality was assessed with the Agilent 2100 BioAnalyzer (Agilent Technologies), and concentrations were determined using the NanoDropND-1000 (NanoDrop Technologies).

Expression profiling. Expression profiling of the samples, each with either two or three technical replicates, was performed using Illumina Human HT-12 V3 BeadChips (Illumina), including 48,804 probes with 200 ng of total RNA processed according to the protocol supplied by Illumina. All samples were randomized before array hybridization, and replicates were hybridized on different BeadChips. Raw data were imported to Illumina BeadStudio software, and probes with less than three beads present were excluded. Log_2-transformed expression signals were normalized separately per tissue, with quantile normalization of the replicates of each individual followed by quantile normalization across all individuals, as previously described[11]. We acknowledge that quantile normalization does not adjust for shared covariance due to technical factors that may influence subsequent analysis, but previous efforts[5] indicate that the impact on the result seems to be minor. After quality control, expression profiles were obtained for 825 (adipose and LCL) and 705 (skin) individuals. Illumina probe annotations were cross-checked by mapping probe sequences to the NCBI Build 36 genome with MAQ[31]. Only uniquely mapping probes with no mismatches and either an Ensembl or RefSeq ID were kept for analysis. Probes mapping to genes of uncertain function (LOC symbols) and those encompassing a common SNP (1000 Genomes Project release June 2010) were further excluded, leaving 23,596 probes for the analysis.

Genotyping and genotype imputation. Genotyping of the TwinsUK data set (N = ~6,000) was performed with a combination of Illumina HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M chips. Intensity data for each of the arrays were pooled separately (with 1M-Duo and 1.2MDuo 1M data pooled together), and genotypes were called with the Illuminus[32] calling algorithm, setting the threshold at a maximum posterior probability of 0.95, as previously described[29].

Imputation was performed using the IMPUTE software package (v2)26 using two reference panels: P0 (HapMap 2, release 22, combined Utah residents of Northern and Western European ancestry (CEU), Yoruba from Ibadan, Nigeria (YRI) and Asian (ASN) panels) and P1 (610k+, including the combined HumanHap610k and 1M arrays). After imputation, SNPs were filtered for MAF of >5% and IMPUTE info value of >0.8, resulting in a total of 2,029,988 SNPs available for testing.

Heritability analysis. The classical twin design was applied, comparing the similarity of mono- and dizygotic twins using the ACE model, which partitions the variance into additive genetic (A), common environment (variance due to environmental effects shared within twin pairs; C) and unique environment (environmental effects not shared within twin pairs; E). As all twin pairs included in the study visited the clinic in pairs and because monozygotic twins share 100% of their genes, any differences arising between them in these circumstances are unique (E). The correlation observed between monozygotic twins thus provides an estimate of A + C. In contrast, dizygotic twins have a common shared environment but share on average only 50% of their genes, such that the correlation between dizygotic twins is a direct estimate of 0.5 A + C. Consequently, twice the difference between mono- and dizygotic twins gives the genetic additive effect (A), and the common environment (E) is the monozygotic correlation minus the estimate of the genetic effect (A). A standard linear mixed model was used to estimate these variance components, as previously described[19]. The following covariates were included in the model: (i) age and experimental batch in adipose and LCL analysis and (ii) age, experimental batch and sample processing in the skin analysis. All available complete twin pairs were included, corresponding to 143 monozygotic and 214 dizygotic pairs with adipose profiles, 138 monozygotic and 221 dizygotic pairs with LCL profiles and 108 monozygotic and 162 dizygotic pairs with skin profiles.

eQTL analysis. Associations of expression levels with probabilities of imputed genotypes were tested in samples of related individuals using a two-step mixed model–based score test developed by Aulchenko *et al.*[33] and Chen and Abecasis[34] and implemented in the GenABEL/ProbABEL packages[35,36]. Briefly, the approach is an approximation of a full linear mixed model, where the first step includes a mixed model containing all terms but those involving SNPs fitted by maximum likelihood (fixed effects as well as the kinship matrix are based on genomic data). Fixed effects included age and experimental batch in the adipose and LCL analysis, and age, batch and sample processing were used in the skin analysis. This step was performed using GenABEL software[35] with the polygenic() function. The resulting object contains the inverse variance–covariance matrix of the estimates and expression trait residuals, which are used in the second step together with posterior genotypic probabilities to perform a score test in ProbABEL[36] using the –mmscore option. In total, 776 adipose, 777 LCL and 667 skin samples had both expression profiles and imputed genotypes and were included in the analysis. *Cis* analysis was limited to SNPs located within 1 Mb of either side of the transcription start or end site or within the gene body. FDR for the *cis* analysis was calculated from the complete list of P values, using the qvalue package[20] implemented in R2.11 (ref. 37). To characterize likely independent regulatory effects, the identified *cis* eQTLs were mapped to recombination hotspot intervals[38]. For each gene, the most significant SNP per hotspot interval was selected, and additional linkage disequilibrium filtering was performed (for each remaining SNP pair with $D' > 0.5$ and $r^2 > 0.1$, the least significant SNP was ignored).

Trans analysis was limited to SNPs located on a different chromosome than the tested transcript. After quality control filtering, analysis of the *trans* eQTLs revealed 52 probes with extreme outlier effects, which were filtered from further *trans* analysis. Transcripts associated with a *trans* SNP at $P < 5 \times 10^{-8}$ were used for calculations of transcript-wise FDR from the complete list of P values, using the qvalue package[20] implemented in R2.11 (ref. 37).

The score test is known to slightly underestimate additive effect sizes[34]; therefore, the top association per probe was validated with a linear mixed-effects model in R, using the lmer() function in the lme4 package[39], fitted by maximum likelihood (**Supplementary Fig. 2**). The linear mixed-effects model was adjusted for both fixed (age, experimental batch effect and sample processing effect (skin tissue only)) and random effects (family relationship and zygosity). A likelihood ratio test was applied to assess the significance of the SNP effect. The P value of the SNP effect in each model was calculated from the χ-squared distribution with 1 degree of freedom, using −2log(likelihood ratio) as the test statistic.

eQTL analysis using a matched co-twin design. eQTL analysis was performed separately for each tissue, as previously described[11]. Within each tissue, twins from the same pair were separated by ID into two samples that were analyzed independently. Related individuals (sister pairs) within a twin set were also removed. This separation resulted in the following sample sizes for adipose, LCLs and skin, respectively: twin 1 (390, 340 and 337) and twin 2 (384, 338 and 328). For each of the twin-by-tissue sets, associations between genotypes and normalized expression values were conducted using Spearman rank correlation (SRC). Age and experimental batch were included as cofactors in the adipose and LCL analysis, and age, batch and sample processing were included in the skin analysis. We considered a window of <1 Mb from the TSS for testing SNPs in *cis*. *cis* eQTLs were filtered at a nominal SRC P value of $<2.5 \times 10^{-6}$, which corresponds to a 10^{-3} permutation threshold[11]. We contrasted the eQTLs (same SNP-probe combinations) and expression fold changes (difference in mean expression of homozygous genotype classes) between twin sets of the same tissue and then performed comparison between tissues from the same twins (for example, twin 1 LCL versus twin 2 LCL, twin 1 LCL versus twin 1 adipose and twin 1 LCL versus twin 1 skin).

Global regression. To estimate the proportion of the genetic variance that is due to *cis* effects, we performed quantitative linkage analysis for the subset of transcripts that had $h^2 > 0.1$ and associated with a common *cis* SNP at 1% FDR. IBD sharing between every pair of dizygotic twins was calculated by phasing all the SNPs in every *cis* region (~2 Mb) using MERLIN 1.1.2 (ref. 40) and then counting haplotypes identical in both twins (IBD = 0, both haplotypes different; IBD = 1, one identical haplotype; IBD = 2, both haplotypes identical). IBD sharing in *cis* for all the probes tested for linkage was calculated.

To estimate the average heritability at *cis* regions, a modification of the Haseman-Elston regression method was used that analyzes the data of several traits in a single linear regression[16]. Briefly, y_{gi} represents the expression for gene g and individual i, normalized to have mean 0 and variance 1 across all the individuals. $Y_{gij} = (Y_{gi} \times Y_{gj})$ is a measure of phenotypic similarity between twins i and j in gene g, and π_{gij} is the IBD sharing between the dizygotic pair (i and j) at gene g calculated as described above. Y_{gij} was regressed on the IBD sharing π_{gij} over all genes and all dizygotic twin pairs. The coefficient of this regression is an estimate of the average variance explained in *cis*. The quotient of this value with the average of total heritability for the same set of transcripts represents a measure of the proportion of the heritability that is explained by variants in *cis*. Next, each gene's expression value was corrected by *cis*-eQTL effects and calculated as the residual of the linear regression of the original gene expression level on the independent *cis* eQTL for each gene. The global regression procedure was then repeated but, in this case, using the gene expression values corrected by the common *cis*-eQTL effects. The coefficient of this regression represents the estimate of the average gene expression variance explained by variants not discovered in our eQTL analysis, which are most likely rare variants. By subtracting this value from the total heritability in *cis*, we obtained an estimate of the genetic variance at the *cis* regions that is due to common SNPs. All three tissues were analyzed separately, and linear regressions were adjusted using R version 2.13.0 (ref. 37).

Replication cohorts. Characteristics of the replication cohorts are presented in **Supplementary Table 4**. The deCODE replication sample consisted of 585 subcutaneous adipose samples from healthy Icelandic individuals, as previously described[8].

The MGH replication sample consisted of 701 subcutaneous adipose samples from obese individuals undergoing Roux-en Y gastric bypass surgery at Massachusetts General Hospital, as previously described[10].

The Oxford replication sample consisted of 331 LCLs independently derived from the TwinsUK Adult twin registry and thus does not overlap with MuTHER samples, as recently described[41].

The ALSPAC replication sample consisted of 931 LCLs derived from The Avon Longitudinal Study of Parents and their Children (ALSPAC)[42]. Expression profiling of the samples, each with two technical replicates, was performed using Illumina Human HT-12 V3 BeadChips (IlluminaInc) and processed as for the MuTHER samples. ALSPAC individuals were genotyped using the Illumina HumanHap550 genome-wide SNP genotyping platform. Markers with <1% MAF or >5% missing genotypes or that failed an exact test of Hardy-Weinberg equilibrium ($P < 5 \times 10^{-7}$) were excluded from further analysis. Any individuals who did not cluster with the CEU individuals in MDS analysis or who had >3% missing data, minimal or excessive heterozygosity (>33% or <31%, respectively), evidence of cryptic relatedness (>10% IBD) or incorrect gender assignment were also excluded. After data cleaning, 315,807 SNPs were left. Imputation was carried out using MACH 1.0.16 (Markov chain haplotyping)[43], using Centre d'Etude du Polymorphisme Humain (CEPH) individuals from phase 2 of the HapMap project as a reference set. Associations between SNP genotypes and normalized expression values were conducted using a linear model.

The GenCORD replication sample consisted of 68 primary fibroblasts derived from the umbilical cord of newborns of Western European ancestry who were born at the maternity ward of the University of Geneva Hospital, with pregnancies being full term or near full term (38–41 weeks), as previously described[9].

Integration of eQTL data with GWAS hits. The likelihood of a shared functional effect between a GWAS SNP (NHGRI database, accessed 21 December 2010) and an eQTL was assessed by quantifying the change in the statistical significance of the eQTL after correcting for the effect of the GWAS SNP, as previously described[14]. ProbABEL association analysis of eQTL genotype with residuals from standard linear regression of the 'corrected-for' SNP against normalized expression was performed again. The linkage disequilibrium structure in each hotspot interval was individually accounted for by ranking (RankGWAS SNP) impact on the eQTL (quantified by the adjusted association P value after correction) of the GWAS SNP correction to that of correcting for all other SNPs in the same interval. By taking into account the total number of SNPs in the interval (N SNPs), this ranking across different genes and intervals can then be compared. For this purpose, we define the RTC score as RTC = (N SNPs – RankGWAS SNP)/N SNPs, ranging from 0 to 1, with values closer to 1 indicating causal regulatory effects.

31. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
32. Teo, Y.Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741–2746 (2007).
33. Aulchenko, Y.S., de Koning, D.J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577–585 (2007).
34. Chen, W.M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
35. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
36. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).
37. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (R Foundation for Statistical Computing, Vienna, 2010).
38. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
39. Bates, D.M. *lme4: Linear Mixed-Effects Models Using S4 Classes.* (R Foundation for Statistical Computing, Vienna, 2010).
40. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
41. Min, J.L. *et al.* The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PLoS ONE* **6**, e22070 (2011).
42. Golding, J., Pembrey, M. & Jones, R. ALSPAC—the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr. Perinat. Epidemiol.* **15**, 74–87 (2001).
43. Li, Y. *et al.* MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).

doi:10.1038/ng.2394