nature genetics

Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins

Alfonso Buil[1–3] Andrew Anand Brown[4,5], Tuuli Lappalainen[1–3,6], Ana Viñuela[7], Matthew N Davies[7], Hou-Feng Zheng[8–10], J Brent Richards[7–10], Daniel Glass[7], Kerrin S Small[7], Richard Durbin[4], Timothy D Spector[7] & Emmanouil T Dermitzakis[1–3]

Understanding the genetic architecture of gene expression is an intermediate step in understanding the genetic architecture of complex diseases. RNA sequencing technologies have improved the quantification of gene expression and allow measurement of allele-specific expression (ASE). ASE is hypothesized to result from the direct effect of *cis* regulatory variants, but a proper estimation of the causes of ASE has not been performed thus far. In this study, we take advantage of a sample of twins to measure the relative contributions of genetic and environmental effects to ASE, and we find substantial effects from gene × gene (G×G) and gene × environment (G×E) interactions. We propose a model where ASE requires genetic variability in *cis*, a difference in the sequence of both alleles, but where the magnitude of the ASE effect depends on *trans* genetic and environmental factors that interact with the *cis* genetic variants.

Gene expression is a cellular phenotype that provides information on the functional state of the cell. It is used as an intermediate phenotype between genetic variation and complex traits to help in the identification of causal genes affecting variation in complex traits. Gene expression is itself a complex trait that depends on genetic and environmental factors. Many researchers have studied the genetics of gene expression, and thousands of expression quantitative trait loci (eQTLs) have been identified within different populations and tissues[1–3]. Recently, epistatic interactions affecting gene expression have been described[4], adding more complexity to the genetic architecture of gene expression. The use of RNA sequencing (RNA-seq) technologies to measure gene expression allows the estimation of ASE. ASE quantifies the difference in the expression of the two haplotypes of an individual at a specific genetic locus[5–7] (**Supplementary Fig. 1a**). Whereas eQTLs are population-based measures of the effect of genetics on gene expression, ASE is a more direct measure of how gene expression changes at the level of the individual. In addition, ASE is much less sensitive to technical parameters, as such effects would affect both alleles equally. Although ASE may occur in a stochastic way within single cells, measurements from a population of cells for each individual represent the average behavior of the two alleles and, theoretically, are expected to result from the direct effect of genetic regulatory variants in *cis*. ASE is therefore expected to be much less influenced by environmental and experimental variability, which account for approximately 70% of the variance[1], thereby allowing us to dissect in more detail the remaining 30% of genetic variability. In this study, we dissect the underlying causes of ASE, by measuring the relative contributions of genetic and environmental factors, and propose biological models of ASE action. To achieve these goals, we sequenced the mRNA fraction of the transcriptomes of ~400 female twin pairs (~800 individuals) from the TwinsUK cohort in 4 tissues—fat, skin, blood and lymphoblastoid cell lines (LCLs)—using 49-bp paired-end sequencing on an Illumina HiSeq 2000. We sequenced 766 fat samples, 814 LCL samples, 716 skin samples and 384 blood samples and obtained 28 million exonic reads per sample on average. Genotype information was imputed into the 1000 Genomes Project Phase 1 reference panel. By constructing a quantitative measure of ASE and exploiting the twin structure, we can dissect the proportions of variation in ASE that are due to distinct genetic and non-genetic causes.

Because our expectation was that *cis* eQTLs would have an important role in ASE, we looked for *cis* eQTLs in the four tissues. We used a linear regression approach with SNPs in a 1-Mb window centered on the transcription start site (TSS) for each gene (Online Methods). We identified 9,166 significant *cis* eQTLs in fat, 9,551 in LCLs, 8,731 in skin and 5,313 in blood (false discovery rate (FDR) of 1%).

We used the RNA-seq data to estimate ASE for every individual at every transcribed heterozygous SNP in the four tissues separately. First, we ran a test to identify statistically significant ASE sites. We then defined a quantitative phenotype that measured the amount of ASE at a site and looked for the variance components of that phenotype.

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [2]Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland. [3]Swiss Institute of Bioinformatics, Geneva, Switzerland. [4]Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [5]NORMENT, KG Jebsen Center for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [6]Department of Genetics, Stanford University, Stanford, California, USA. [7]Department of Twin Research, King's College London, London, UK. [8]Department of Medicine, McGill University, Montreal, Quebec, Canada. [9]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [10]Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada. Correspondence should be addressed to E.T.D. (emmanouil.dermitzakis@unige.ch) or A.B. (alfonso.buil@unige.ch).

Received 12 March; accepted 6 November; published online 1 December 2014; doi:10.1038/ng.3162

To assess whether a heterozygous site showed statistically significant ASE, we used a binomial test on the proportion of reference alleles versus total counts (Online Methods). Because ASE estimates are sensitive to read coverage and mapping bias, we restricted our analysis to sites with at least 30 reads that passed a rigorous filtering process to control for mapping bias and other confounders[7] (Online Methods). We tested an average of 1,582 sites per individual, 8% of which were statistically significant at an FDR threshold of 10% (**Supplementary Table 1**). We identified 8,013 ASE sites in fat, 10,751 in LCLs, 9,538 in skin and 6,827 in blood. About 80% of the ASE sites were in genes for which we also identified a *cis* eQTL (**Supplementary Table 1**). We assume that the genes with ASE without observed *cis* eQTLs also have genetic variants in *cis* causing the allelic imbalance in expression, but we did not have the power to find these, owing to small effect sizes or the variants being at low frequency in the population or being involved in epistatic or G×E interactions. We cannot exclude the possibility that, in some cases, homeostatic or feedback mechanisms act to constrain total expression such that an imbalance in allelic expression does not change the total output.

To quantify genetic and environmental sources of variation in ASE, we developed an extension of the classical variance components approach based on the correlations within monozygotic and dizygotic twin pairs. We defined a quantitative phenotype of ASE as the logit of the proportion of reference alleles. This measure is not dependent on the overall gene expression level and is not susceptible to giving false interactions due to *trans* or environmental effects that increase the overall level of expression. We jointly analyzed all the sites in genes with at least one eQTL, where both siblings had at least 30 reads overlapping the site and ASE was statistically significant for at least one of the siblings. We estimated the correlation of the ASE phenotypes within monozygotic and dizygotic twin pairs and observed that the correlation among the dizygotic twins was greater than half of the correlation among the monozygotic twins (**Fig. 1**). This finding could indicate a potential shared environment component, but, in our case, it is more likely to be due to the fact that the *cis* eQTL has a large effect on ASE and that our dizygotic twins are genetically more similar than random mating predicts at the ASE locus (the mean identical-by-state (IBS) coefficient at the eQTL for dizygotic twins was 0.9). Indeed, when we looked at the correlation between dizygotic twins who had IBS = 0.5 (and hence shared half of the contribution of the additive eQTL in monozygotic twins), we observed that this correlation was less than half of the correlation between monozygotic twins (**Fig. 1**), indicating the potential presence of non-additive genetic effects.

To incorporate these complexities in the model, we separated the twin pairs on the basis of the average genetic similarity across the genome (1 for monozygotic twins and 0.5 for dizygotic twins); the genetic similarity at the locus, based on the identity-by-descent (IBD) status of the *cis* region surrounding the gene; and the genetic similarity of the eQTL, based on the IBS status at the eQTL locus. We estimated the correlations of the ASE phenotype for each category of twins and modeled these correlations as a function of six variance components (Online Methods). These components represent the proportions of variance in ASE that can be explained by environmental variation, by the top eQTL, by other variants in *cis*, by variants in *trans* and by genetic interactions. As recombination is unlikely to have occurred within the *cis* window, interacting pairs of SNPs are similarly inherited by twin pairs, and their contribution to variance was thus effectively additive. Instead, we looked to calculate the proportion of variance explained by *cis*-*trans* interactions. We found that the heritability of ASE (the sum of all the genetic components:

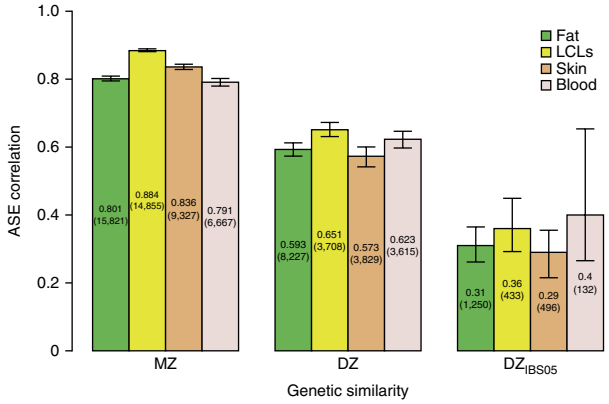

Figure 1 ASE correlation among twin pairs for different categories of genetic similarity. MZ, monozygotic twins; DZ, dizygotic twins; DZ_{IBS05}, dizygotic twins with IBS values equal to 0.5 at the eQTL locus. The 95% confidence intervals were calculated using 1,000 bootstrap permutations. For each bar, the top number is the ASE correlation and the number in parentheses is the number of twin pairs used in the calculation.

eQTL + *cis* + *trans* + interactions) ranged from 62% to 88% (**Fig. 2**). The effect of the most statistically significant *cis* eQTL for each gene accounted for 26–46% of the variance in ASE. This means that nearly half of the heritability of ASE is due to a common *cis* eQTL. The remainder of the variance is due to other genetic effects in *cis* (11–22% of the ASE variance) and genetic interaction effects (11–29% of the ASE variance). As expected from the biological assumptions, we did not observe significant additive *trans* effects. We found a significant effect from shared environment only in blood (11% of the ASE variance). This finding of a shared environmental effect only in blood could be due to the fact that blood is more heterogeneous than the other tissues, with variable proportions of the different cell types in individuals and shared environment affecting the counts of the different cell types. In the shared environment component, we are likely detecting cell type–specific effects. We used 1,000 bootstrap permutations to calculate the confidence intervals of our variance component estimates (**Fig. 2**). This approach is robust to different coverage thresholds and the presence of several ASE sites in the same gene (**Supplementary Figs. 2** and **3**). In summary, the main cause of ASE is genetic variants in *cis*, as expected, but between 38% and 49% of variance in the ASE ratio is due to genetic interactions and environmental factors.

Our variance components model showed that genetic effects did not explain all the observed variance in ASE and that environmental factors could have an effect on ASE. Given the nature of ASE—which, contrary to total gene expression, is intrinsically controlled by the gene locus—these environmental effects should mainly be due to true biological effects mediated by epigenetic mechanisms and are much less likely to be explained by technical and experimental effects. However, environmental and epigenetic effects alone cannot create imbalance in allelic expression. As ASE is averaged over a large population of cells, stochastic effects are equally distributed between the two alleles. In consequence, to observe ASE, an effect from a *cis* DNA sequence is required. We therefore postulated the existence of G×E interactions affecting ASE.

To identify cases of G×E interaction, we used an analysis inspired by the classical discordant monozygotic twins analysis[8–10]. We defined the phenotype as the absolute difference in measured ASE between monozygotic twins and looked for SNPs around the ASE site that were associated with this phenotype. Significant associations suggest the influence of environmental factors on ASE in a genotype-dependent

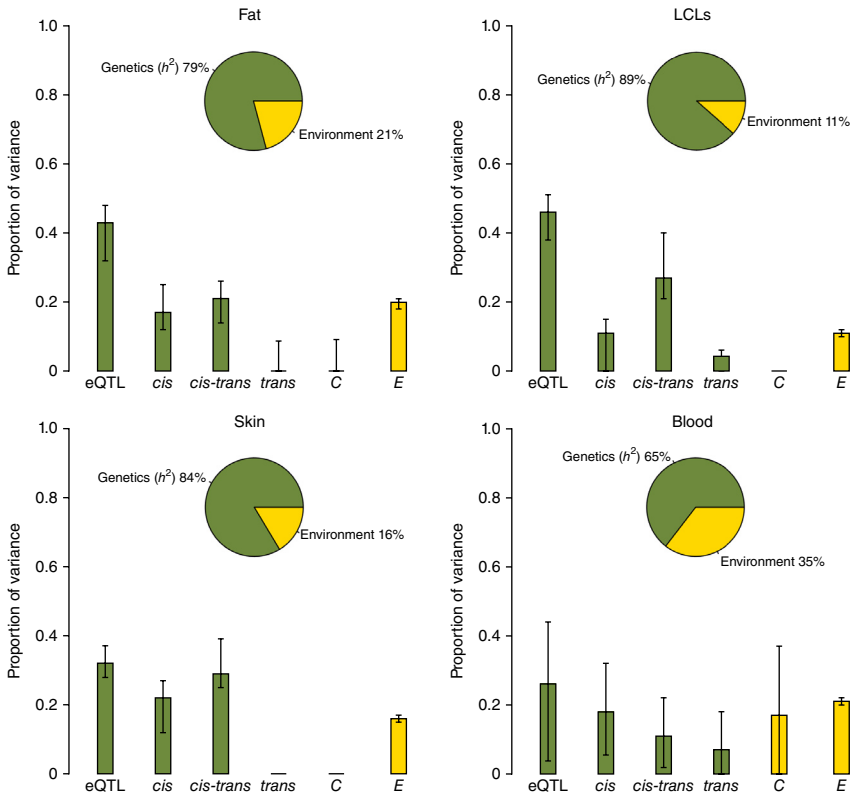

Figure 2 Variance components for ASE for the model with *cis* × *trans* interaction. h^2 (equal to eQTL + *cis* + *cis-trans* + *trans*) is the heritability of ASE, where "eQTL" is the proportion of variance explained by a common eQTL, "*cis*" is the proportion of variance explained by other variants in *cis*, "*cis-trans*" is the proportion of variance explained by interactions between *cis* and *trans* genetic variants and "*trans*" is the proportion of variance explained by genetic variants in *trans*. "*C*" is the proportion of variance explained by shared environment, and "*E*" is the proportion of variance explained by individual environment. The 95% confidence intervals were calculated using 1,000 bootstrap permutations.

manner. After correction for multiple testing, we found evidence of G×E interactions in fat and LCLs but no conclusive results in skin and blood (**Supplementary Tables 2–5**). One of the top hits in LCLs was *EBI3* (encoding Epstein-Barr virus (EBV)-induced 3) (**Fig. 3** and **Supplementary Fig. 4**). This finding means that ASE at the *EBI3* gene depends on the interaction of *cis* genetic variants with an environmental factor, in this case likely to be related to the transformation process of B cells with EBV. The two top hits in fat were *ADIPOQ* and *ACSL1*, two genes that encode the adiponectin and long-chain fatty acid–CoA ligase 1 proteins, respectively (**Fig. 3** and **Supplementary Fig. 4**). These two proteins are functionally related: both participate in the Gene Ontology (GO) biological processes 'response to fatty acid' and 'response to nutrient', and both are known to be regulated by environmental factors such as diet and exercise in a genotype-dependent manner[11–14]. Attempts to link the observed G×E interaction to environmentally affected phenotypes (body mass index (BMI), glucose levels and insulin levels) did not show any significant associations, which is not surprising as these are phenotypes affected by the environment and not direct environmental measures. The analysis above suggests that environment can modulate the effect of SNPs on gene expression.

In conclusion, these results show a complex genetic architecture for the *cis* regulation of gene expression as measured through ASE. We propose a model where allelic imbalance in expression (ASE) requires genetic variability in *cis*; however, the magnitude of the ASE effect depends on *trans* genetic and environmental factors that interact with the *cis* genetic variants (**Supplementary Fig. 1b,c**). Examples of interactions between *cis* and *trans* genetic variants affecting gene expression have been described recently[4]. Here we provide global quantification of the magnitude of these effects. About 38–49% of the variance in the observed ASE is not explained by additive genetic effects. This means that a substantial amount of the variance observed in ASE and, therefore, in the genetic regulation of gene expression is due to G×G and G×E interactions. It is worth noting that our results show no additive *trans* effects on ASE. This does not mean that there are no additive *trans* effects affecting gene expression; it means that the *trans* effects on ASE are not additive. We found an example of a G×E interaction affecting gene expression that has been widely described in the literature (adiponectin), supporting the validity of our approach. However, the

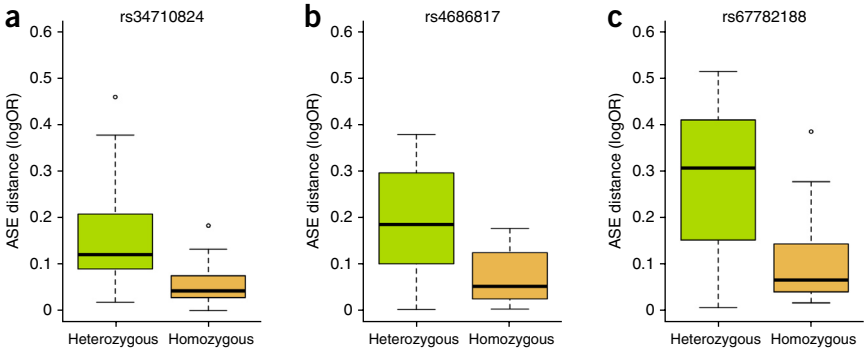

Figure 3 G×E examples discovered using analysis of discordant monozygotic twins. Monozygotic twin pairs show different ASE effects at some genes depending on the genotype of specific SNPs. Each box plot represents the median and interquartile range (IQR) for each distribution, whiskers represent data up to 1.5 times the IQR and outliers are shown as separate dots. The *y* axis shows the ASE difference between monozygotic twins, which is equal to the logarithm of the odds ratio (OR) between the two ASE measures. Because monozygotic twins are genetically identical, this association reflects the interaction of the SNP with an unknown environment. Box plots represent the difference in ASE in monozygotic twin pairs who are heterozygous (green) and homozygous (orange) at the SNP of interest. (**a**) ASE for the *ACSL1* gene shows G×E interaction with SNP rs34710824 in fat (32 homozygous and 23 heterozygous twin pairs). (**b**) ASE for the *ADIPOQ* gene shows G×E interaction with SNP rs4686817 in fat (26 homozygous and 36 heterozygous twin pairs). (**c**) ASE for the *EBI3* gene shows G×E interaction with SNP rs67782188 in LCLs (19 homozygous and 32 heterozygous twin pairs).

limitation on power due to sample size prevented us from discovering specific associations in most other cases. Allelic gene expression is the molecular phenotype closest to the action of genetic variation. The presence of widespread G×G and G×E interactions affecting this phenotype implies that G×G and G×E interactions can be important in other complex phenotypes, including diseases. The proposed model has implications for the interpretation of the effects of genome-wide association study (GWAS)-identified genetic variants on complex diseases. About 80% of GWAS signals are estimated to be regulatory variants. The search for G×G and G×E interactions conditioning on relevant biological models rather than whole-genome agnostic searches is likely to recover a substantial fraction of the genetic and non-genetic variance associated with disease risk.

METHODS

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

Accession codes. RNA-seq data have been deposited in the European Genome-phenome Archive (EGA) under accession EGAS00001000805.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the twins for their voluntary contribution to this project. This work has been funded by European Union Framework Programme 7 grant EuroBATS (259749), which also supports A.A.B., A.B., M.N.D., D.G., A.V. and T.D.S. A.A.B. is also supported by a grant from the South-Eastern Norway Health Authority (2011060). R.D. is supported by the Wellcome Trust (098051). The Louis-Jeantet Foundation, the Swiss National Science Foundation, the European Research Council (ERC) and the US National Institutes of Health/National Institute of Mental Health GTEx grant support E.T.D. T.D.S. is a National Institute of Health Research (NIHR) senior investigator and the holder of an ERC Advanced Principal Investigator award. J.B.R. and H.F.Z. are supported by the Canadian Institutes of Health Research, Fonds de Recherche Santé du Québec and the Quebec Consortium for Drug Discovery. Most computations were performed at the Vital-IT center for high-performance computing of the Swiss Institute of Bioinformatics (SIB; http://www.vital-it.ch/). The TwinsUK study was funded by the Wellcome Trust, European Community Framework Programme 7 (2007–2013), and the NIHR Clinical Research Facility at Guy's and St Thomas' National Health Service (NHS) Foundation Trust and the NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. SNP genotyping was performed by the Wellcome Trust Sanger Institute and National Eye Institute via US National Institutes of Health/Center for Inherited Disease Research (CIDR) funding.

AUTHOR CONTRIBUTIONS

A.B., R.D., T.D.S. and E.T.D. conceived the study. A.B., A.A.B., A.V. and M.N.D. analyzed the data. T.L. and K.S.S. contributed experimental and technical support as well as discussion. D.G. contributed to sample collection. H.F.Z. and J.B.R. contributed technical support and analyzed data. A.B. prepared the manuscript, with contributions from A.A.B. and E.T.D. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
2. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
3. Stranger, B.E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
4. Hemani, G. *et al.* Detection and replication of epistasis influencing transcription in humans. *Nature* **508**, 249–253 (2014).
5. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
6. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
7. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
8. Essaoui, M. *et al.* Monozygotic twins discordant for 18q21.2qter deletion detected by array CGH in amniotic fluid. *Eur. J. Med. Genet.* **56**, 502–505 (2013).
9. Souren, N.Y. *et al.* Adult monozygotic twins discordant for intra-uterine growth have indistinguishable genome-wide DNA methylation profiles. *Genome Biol.* **14**, R44 (2013).
10. Surakka, I. *et al.* A genome-wide association study of monozygotic twin-pairs suggests a locus related to variability of serum high-density lipoprotein cholesterol. *Twin Res. Hum. Genet.* **15**, 691–699 (2012).
11. Ferguson, J.F. *et al.* Gene-nutrient interactions in the metabolic syndrome: single nucleotide polymorphisms in *ADIPOQ* and *ADIPOR1* interact with plasma saturated fatty acids to modulate insulin resistance. *Am. J. Clin. Nutr.* **91**, 794–801 (2010).
12. Joseph, P.G., Pare, G. & Anand, S.S. Exploring gene-environment relationships in cardiovascular disease. *Can. J. Cardiol.* **29**, 37–45 (2013).
13. Pérez-Martínez, P. *et al.* Adiponectin gene variants are associated with insulin sensitivity in response to dietary fat consumption in Caucasian men. *J. Nutr.* **138**, 1609–1614 (2008).
14. Warodomwichit, D. *et al. ADIPOQ* polymorphisms, monounsaturated fatty acids, and obesity risk: the GOLDN study. *Obesity (Silver Spring)* **17**, 510–517 (2009).

ONLINE METHODS

Sample collection. The study included 856 female individuals of European ancestry recruited from the TwinsUK Adult twin registry. 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 by EBV transformation of the B-lymphocyte component by the European Collection of Cell Cultures agency.

The St. Thomas' Research Ethics Committee (REC) approved on 20 September 2007 the protocol for the dissemination of data, including DNA, with REC reference number RE04/015. On 12 March 2008, the St Thomas' REC confirmed that this approval extended to expression data. Volunteers gave informed consent and signed an approved consent form before the biopsy procedure. Volunteers were supplied with an appropriate detailed information sheet regarding the research project and biopsy procedure by post before attending for the biopsy.

Genotyping and imputation. Samples were genotyped on a combination of the HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo Illumina arrays. Samples were imputed into the 1000 Genomes Project Phase 1 reference panel (data freeze 10 November 2010)[15] using IMPUTE2 (ref. 16) and filtered (minor allele frequency (MAF) < 0.01 and IMPUTE info value < 0.8).

RNA processing. Samples were prepared for sequencing with the Illumina TruSeq sample preparation kit according to the manufacturer's instructions and were sequenced on a HiSeq 2000 machine. The 49-bp sequenced paired-end reads were mapped to the GRCh37 reference genome[17] with the Burrows-Wheeler Aligner (BWA) v0.5.9 (ref. 18). We used genes defined as protein coding in GENCODE 10 annotation[19]. We excluded samples that failed in the library preparation or sequence process. We also excluded samples with fewer than 10 million reads sequenced and mapped to exons. Finally, we excluded samples for which the sequence data did not correspond with the actual genotype data. We ended with 766 samples for fat, 814 for samples for LCLs, 716 samples for skin and 384 samples for blood (we had blood samples for only half of the individuals).

eQTL discovery. *Exon quantification.* All overlapping exons of a gene were merged into meta-exons with an identifier of the form 'geneID_start.pos_end.pos'. We counted a read as mapping to a meta-exon if either its start or end coordinate overlapped a meta-exon.

Normalization. All read count quantifications were corrected for variation in sequencing depth between samples by normalizing the number of reads to the median number of well-mapped reads. We used only exons that were quantified in more than 90% of the individuals. We removed the effects of technical covariates, regressing out the first 50 factors from PEER[20], including BMI and age in the model to preserve major biological sources of variation.

eQTL association. Because our data samples were from twins, they did not constitute independent observations, and we needed to take this into account in our models. We used the two-step strategy described in Aulchenko *et al.*[21]. First, we kept the residuals of a mixed model that removed the effects of the family structure using the implementation in the GenAbel R package. We then transformed these residuals using a rank normal transformation. Finally, we performed a linear regression of the transformed residuals on the SNPs in a 1-Mb window centered on the TSS for each gene, using the MatrixeQTL R package[22]. We examined association at the exon level and kept the best association for each gene.

Permutations. We permuted the quantifications of each exon 2,000 times, keeping the best *P* value for each exon from each round. From these data, we adjusted the empirical FDR to 1% according to the most stringent exon of each gene, stratifying the analysis on the number of exons for a given gene.

Site filtering for ASE. In all ASE analyses, we excluded sites that were susceptible to allelic mapping bias, removing (i) sites with 50-bp mappability <1 according to the UCSC mappability track, implying that the 50-bp region flanking the site is non-unique in the genome, and (ii) simulated RNA-seq reads overlapping the site that showed >5% difference in the mapping of reads that carried the reference and non-reference allele. To verify that genotype was truly heterozygous, we used only sites covered by at least 30 reads and where both alleles were observed in the RNA-seq data[7].

Binomial test for ASE. We identified statistically significant ASE sites using a binomial test. We performed a test for each heterozygous SNP in every individual to detect the presence of statistically significant allelic imbalance in expression. For each site-individual combination, we counted the number of reads covering each allele and calculated a binomial test comparing the observed proportion of reference allele counts with the expected proportion. In theory, the expected proportion should be 0.5, but mapping bias can alter it a little. To correct for systematic bias in allelic ratios, we calculated the overall reference to total allele ratio for each individual for each SNP base combination. These ratios were then used as the expected ratios in the binomial test. We called significant ASE sites using an FDR threshold of 10%. We assessed the robustness of our significant ASE calls in four ways. First, we evaluated the concordance of ASE among tissues by measuring the ASE of significant ASE sites from one tissue in another tissue in the same individual and observed a replication rate of about 70% in the three tissues with a complete sample size (**Supplementary Fig. 5**). We then analyzed five samples from the gEUVADIS Project that were sequenced between two and seven times in different laboratories[7,23]. We observed that the ASE ratio was quite stable with coverage of 30 reads or more (**Supplementary Fig. 6**). We also observed that the agreement in significant ASE calls was stable for different coverages (**Supplementary Fig. 7**). Finally, we analyzed two LCL samples (from the gEUVADIS Project[7]), following the same protocol and analysis pipeline as described in the present paper, and compared the results to the ASE ratios obtained from a new technique that uses microfluidic multiplex PCR and sequencing (mmPCR)[24]. Experimental and statistical analysis of the two samples was independently performed in the two different laboratories. We found a very good agreement between the results we obtained using RNA-seq and using the new mmPCR technique. The replication rate was about 80–82%, and the correlation among the ASE ratios for sites that were significant using RNA-seq data was 0.86 (**Supplementary Fig. 8** and **Supplementary Data Set**). These observations show a high degree of replicability of ASE measures.

Quantification of ASE. The measure we used for variance components analysis of ASE was the logit of the percentage of reference alleles. Calculating p, the percentage of the reference allele at a site for an individual, by p = REF_COUNT/TOTAL_COUNT, the measure of ASE was:

$$\mathrm{ASE} = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{\mathrm{REF\_COUNT}}{\mathrm{NONREF\_COUNT}}\right)$$

This measure is not dependent on the overall gene expression level and is thus not susceptible to giving false interactions due to *trans* effects or environmental effects that increase the overall level of expression (**Supplementary Figs. 9** and **10**).

IBD and IBS calculations. *IBD.* We calculated the haplotypes in a 1-Mb window centered on the TSS of each gene and counted the number of haplotype alleles that were shared by the twin pairs at each locus.

IBS. We estimated IBS for each twin pair at each locus on the basis of the eQTL–ASE site haplotype. For each site, we counted the number of alleles in the eQTL–ASE site haplotype that were equal for the pair.

The difference between the IBS and IBD estimates is that, for IBD, we took into account the information from a 1-Mb haplotype, whereas, for IBS estimates, we used only the haplotype with two SNPs—the eQTL and the ASE site.

Variance components models. Classical variance components models in twins model the phenotypic correlation between monozygotic (MZ) and dizygotic (DZ) twins as a function of the additive genetic variance and the shared environmental variance[25]:

$$\mathrm{cor}_{\mathrm{MZ}} = A + C$$

$$\mathrm{cor}_{\mathrm{DZ}} = \frac{1}{2}A + C$$

doi:10.1038/ng.3162

where A represents additive genetic effects and C represents effects due to the common environment for the twin pair (events that affect each member of a twin pair in the same way). The individual environmental effect (events that occur for one twin but not the other) would be calculated as $E = 1 - \text{cor}_{MZ}$. From the two equations above, heritability can be estimated as:

$$h^2 = A = 2(\text{cor}_{MZ} - \text{cor}_{DZ})$$

Here, we extend this model to incorporate new sources of variation, including (i) A_{QTL}, the additive effect due to the best eQTL, (ii) A_{cis}, other genetic additive effects in *cis*, (iii) A_{trans}, additive genetic effects in *trans* and (iv) I, the epistatic interaction between *trans* and *cis* genetic effects, where:

$$A = A_{QTL} + A_{cis} + A_{trans} + I$$

Our model has six variance components: (i) variance due to the effect of the major *cis* eQTL (the IBS status at this locus), (ii) variance due to the rest of the genetic variants in *cis* (including the effect of rare variants; captured by the IBD status), (iii) variance due to genetic variants in *trans* (the genome-wide IBD), (iv) variance due to non-additive genetic effects (genetic interactions), (v) variance due to the shared environmental effect and (vi) variance due to the individual environmental effect.

The equations of the extended model are:

$$\text{cor}_{MZ} = A_{QTL} + A_{cis} + A_{trans} + I + C$$

$$\text{cor}_{DZ\_IBD1} = A_{QTL} + A_{cis} + \frac{1}{2}A_{trans} + \frac{1}{2}I + C$$

$$\text{cor}_{DZ\_IBD05\_IBS1} = A_{QTL} + \frac{1}{2}A_{cis} + \frac{1}{2}A_{trans} + \frac{1}{4}I + C$$

$$\text{cor}_{DZ\_IBD05\_IBS05} = \frac{1}{2}A_{QTL} + \frac{1}{2}A_{cis} + \frac{1}{2}A_{trans} + \frac{1}{4}I + C$$

$$\text{cor}_{DZ\_IBD0\_IBS1} = A_{QTL} + \frac{1}{2}A_{trans} + C$$

$$\text{cor}_{DZ\_IBD0\_IBS05} = \frac{1}{2}A_{QTL} + \frac{1}{2}A_{trans} + C$$

$$\text{cor}_{DZ\_IBD0\_IBS0} = \frac{1}{2}A_{trans} + C$$

where cor_{MZ} is the correlation between monozygotic twins, cor_{DZ_IBD1} is the correlation between dizygotic twins who have IBD = 1 at the gene, $\text{cor}_{DZ\_IBD05\_IBS1}$ is the correlation between dizygotic twins who have IBD = 0.5 at the gene and IBS = 1 at the eQTL, $\text{cor}_{DZ\_IBD05\_IBS05}$ is the correlation between dizygotic twins who have IBD = 0.5 at the gene and IBS = 0.5 at the eQTL, $\text{cor}_{DZ\_IBD0\_IBS1}$ is the correlation between dizygotic twins who have IBD = 0 at the gene and IBS = 1 at the eQTL, $\text{cor}_{DZ\_IBD0\_IBS05}$ is the correlation between dizygotic twins who have IBD = 0 at the gene and IBS = 0.5 at the eQTL, and $\text{cor}_{DZ\_IBD0\_IBS0}$ is the correlation between dizygotic twins who have IBD = 0 at the gene and IBS = 0 at the eQTL.

To calculate these correlations, we used sites covered by at least 30 reads that showed significant ASE in genes with at least one *cis* eQTL. Because the number of individuals that have ASE at a given site is small, we analyzed all the sites together to obtain a global estimate of the variance components. This strategy has been used previously with gene expression data[1,26].

To solve the system of equations, we used the nonlinear optimization package Rsolnp from the R statistical environment[27]. We estimated the solution that minimized quadratic errors, forcing the variance components to be positive.

Genotype-environment interaction. For every ASE site with data for at least 50 monozygotic twin pairs, we calculated the Mann-Whitney test:

$$\text{ASE\_distance} \sim \text{SNP}_i$$

for all the SNPs in a 1-Mb window centered on the TSS of the gene containing the ASE site. ASE_distance is the absolute value of the difference in the ASE phenotype between the two siblings in the monozygotic twin pair and SNP_i represents the genotype of one SNP. Because we were looking for an effect on ASE, we expected similar behavior for the two homozygous genotypes. Therefore, for the association analysis, we coded the genotypes in two categories: homozygous and heterozygous. To correct for multiple testing, we calculated the number of effective tests and applied Bonferroni correction on the basis of the number of tests. Because monozygotic twins are genetically identical, a difference in ASE for two monozygotic siblings has to be caused by environmental or epigenetic causes. A significant association in our tests suggests the existence a G×E interaction affecting ASE. It is worth noting that the associated SNP genotype is not equivalent to the existence of ASE, as other variants might be contributing to ASE as well. There were cases of homozygous pairs with ASE and heterozygous pairs without ASE; in all cases, the difference in ASE was greater for heterozygous pairs (**Supplementary Fig. 11**). Finally, the existence of ASE does not imply a significant G×E interaction, as shown in **Supplementary Figure 12**.

15. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
16. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
17. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
19. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
20. Parts, L., Stegle, O., Winn, J. & Durbin, R. Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genet.* **7**, e1001276 (2011).
21. 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).
22. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
23. 't Hoen, P.A. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
24. Zhang, R. *et al.* Quantifying RNA allelic ratios by microfluidic multiplex PCR and sequencing. *Nat. Methods* **11**, 51–54 (2014).
25. Falconer, D.S. & MacKay, T.F.C. *Introduction to Quantitative Genetics* (Longmans Green, 1996).
26. 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).
27. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2008).