ARTICLE

doi:10.1038/nature12531

Transcriptome and genome sequencing uncovers functional variation in humans

Tuuli Lappalainen[1,2,3], Michael Sammeth[4,5,6,7]†*, Marc R. Friedländer[5,6,7,8]*, Peter A. C. 't Hoen[9]*, Jean Monlong[5,6,7]*, Manuel A. Rivas[10]*, Mar Gonzàlez-Porta[11], Natalja Kurbatova[11], Thasso Griebel[4], Pedro G. Ferreira[5,6,7], Matthias Barann[12], Thomas Wieland[13], Liliana Greger[11], Maarten van Iterson[9], Jonas Almlöf[14], Paolo Ribeca[4], Irina Pulyakhina[9], Daniela Esser[12], Thomas Giger[1], Andrew Tikhonov[11], Marc Sultan[15], Gabrielle Bertier[5,6], Daniel G. MacArthur[16,17], Monkol Lek[16,17], Esther Lizano[5,6,7,8], Henk P. J. Buermans[9,18], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[13], Olof Karlberg[14], Halit Ongen[1,2,3], Helena Kilpinen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[15], Oliver Stegle[11], Matti Pirinen[10], Stephen B. Montgomery[1]†, Peter Donnelly[10], Mark I. McCarthy[10,19], Paul Flicek[11], Tim M. Strom[13,20], The Geuvadis Consortium‡, Hans Lehrach[15,21], Stefan Schreiber[12], Ralf Sudbrak[15,21]†, Ángel Carracedo[22], Stylianos E. Antonarakis[1,2], Robert Häsler[12], Ann-Christine Syvänen[14], Gert-Jan van Ommen[9], Alvis Brazma[11], Thomas Meitinger[13,20,23], Philip Rosenstiel[12], Roderic Guigó[5,6,7], Ivo G. Gut[4], Xavier Estivill[5,6,7,8] & Emmanouil T. Dermitzakis[1,2,3]

Genome sequencing projects are discovering millions of genetic variants in humans, and interpretation of their functional effects is essential for understanding the genetic basis of variation in human traits. Here we report sequencing and deep analysis of messenger RNA and microRNA from lymphoblastoid cell lines of 462 individuals from the 1000 Genomes Project—the first uniformly processed high-throughput RNA-sequencing data from multiple human populations with high-quality genome sequences. We discover extremely widespread genetic variation affecting the regulation of most genes, with transcript structure and expression level variation being equally common but genetically largely independent. Our characterization of causal regulatory variation sheds light on the cellular mechanisms of regulatory and loss-of-function variation, and allows us to infer putative causal variants for dozens of disease-associated loci. Altogether, this study provides a deep understanding of the cellular mechanisms of transcriptome variation and of the landscape of functional variants in the human genome.

Interpreting functional consequences of millions of discovered genetic variants is one of the biggest challenges in human genomics[1]. Although genome-wide association studies (GWAS) have linked genetic loci to various human phenotypes and the functional annotation of the genome is improving[2,3], we still have a limited understanding of the underlying causal variants and biological mechanisms. One approach to addressing this challenge has been to analyse variants affecting cellular phenotypes, such as gene expression[4–8], known to affect many human diseases and traits[9,10].

In this study, we characterize functional variation in human genomes by RNA-sequencing hundreds of samples from the 1000 Genomes Project[1], the most important reference data set of human genetic variation, thus creating the biggest RNA sequencing data set of multiple human populations so far. We not only catalogue novel loci with regulatory variation, but also, for the first time, discover and characterize molecular properties of causal functional variants.

We performed mRNA and small RNA sequencing on lymphoblastoid cell line samples from five populations: the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI). After quality control, we had 462 and 452 individuals (89–95 per population) with mRNA and miRNA data, respectively (Supplementary Figs 1–11 and Supplementary Table 1). Of these, 421 are in the 1000 Genomes Phase 1 data set[1], and the remainder were imputed from single nucleotide polymorphism (SNP) array data (Supplementary Fig. 3 and Supplementary Table 2). High-throughput RNA sequencing (RNA-seq) was performed in seven laboratories, and the smaller amount of variation between laboratories than individuals demonstrated that RNA sequencing is a mature technology ready for distributed data production (Mann-Whitney $P < 2.2 \times 10^{-6}$ for mRNA, $P = 1.34 \times 10^{-10}$ for miRNA; Fig. 1a, Supplementary Fig. 11; for further details see ref. 11). To discover genetic regulatory variants, we mapped *cis*- quantitative trait loci (QTLs) to transcriptome traits of protein-coding and miRNA genes separately in the European (EUR) and Yoruba (YRI) populations (Table 1, Supplementary Fig. 12 and Supplementary Table 3). The RNA-seq read, quantification, genotype and QTL data are available open-access (see Author Information section).

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [2]Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, 1211 Geneva, Switzerland. [3]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. [4]Centro Nacional de Análisis Genómico, 08028 Barcelona, Catalonia, Spain. [5]Centre for Genomic Regulation (CRG), 08003 Barcelona, Catalonia, Spain. [6]Pompeu Fabra University (UPF), 08003 Barcelona, Catalonia, Spain. [7]CRG Hospital del Mar Research Institute, 08003 Barcelona, Catalonia, Spain. [8]CRG CIBERESP, 08003 Barcelona, Catalonia, Spain. [9]Department of Human Genetics, Leiden University Medical Center, 2300 RC Leiden, the Netherlands. [10]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [11]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK. [12]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany. [13]Institute of Human Genetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [14]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, 751 85 Uppsala, Sweden. [15]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany. [16]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [17]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [18]Leiden Genome Technology Center, 2300 RC Leiden, the Netherlands. [19]Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford OX3 7BN, UK. [20]Institute of Human Genetics, Technische Universität München, 81675 Munich, Germany. [21]Dahlem Centre for Genome Research and Medical Systems Biology, 14195 Berlin, Germany. [22]Fundacion Publica Galega de Medicina Xenomica (SERGAS), Genomic Medicine Group, CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [23]Deutsches Forschungszentrum für Herz-Kreislauferkrankungen (DZHK), Partner Site Munich Heart Alliance, 81675 Munich, Germany. †Present addresses: Bioinformatics Laboratory, National Laboratory of Scientific Computing (LNCC), Petropolis 25651-075, Rio de Janeiro, Brazil (M.S.); Departments of Pathology and Genetics, Stanford University, Stanford, California 94305-5324, USA (S.B.M.); Alacris Theranostics GmbH, 14195 Berlin, Germany (R.S.).
*These authors contributed equally to this work.
‡A list of authors and their affiliations appears in the Supplementary Information.

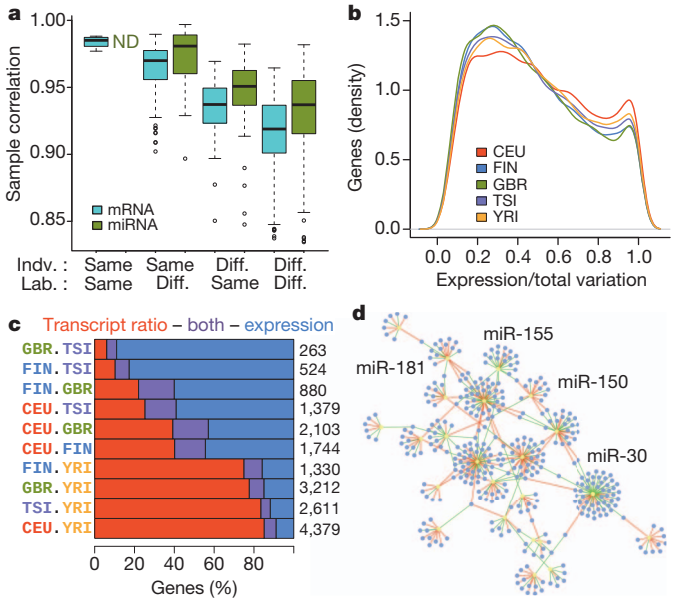

Figure 1 | Transcriptome variation. **a**, Spearman rank correlation of replicate samples, based on mRNA exon and miRNA quantifications of 5 individuals sequenced 8 and 7 times for mRNA and miRNA, respectively, and separated by the individual (indv.) or the sequencing laboratory (lab.) being the same or different (diff.). The quantifications have been normalized only for the total number of mapped reads (see Supplementary Fig. 11 for correlations after normalization). **b**, The proportion of expression level variation (as opposed to splicing) of the total transcription variation between individuals in each population, measured per gene. **c**, Proportion of genes with differential expression levels and/or transcript usage between population pairs, out of the total listed on the right-hand side. **d**, Network of significant miRNA families ($P < 0.001$; yellow) and their significantly associated mRNA targets ($P < 0.05$; purple). The edges display negative (green) and positive (red) associations.

Transcriptome variation in populations

This first uniformly processed RNA-seq data set from multiple human populations allowed high-resolution analysis of transcriptome variation. Individual and population differences in transcripts can manifest in (1) overall expression levels, and (2) relative abundance of transcripts from the same gene (transcript ratios). Deconvolution of the relative contribution of these[12] indicates that this ratio is characteristic for each gene, with transcript ratio being on average more dominant (Fig. 1b and Supplementary Figs 13 and 14). Population differences explain a small but significant proportion of 3% of the total variation (Mann-Whitney $P < 2.2 \times 10^{-16}$). In addition to this genome-wide perspective to population variation, we identified 263–4,379 genes with differential expression and/or transcript ratios between population pairs (P.G.F. *et al.* manuscript submitted). Notably, continental differences between YRI–EUR population pairs have a much higher contribution of genes with different transcript usage than European population pairs (75–85% versus 6–40%; Fig. 1c and Supplementary Fig. 14). This has not been observed before in humans, but it is consistent with splicing patterns capturing phylogenetic differences between species better than expression levels[13,14].

We quantify a total of 644 autosomal miRNAs in >50% individuals, of which 60 have significant *cis*-eQTLs for miRNA expression levels (*cis*-mirQTLs, Table 1 and Supplementary Fig. 15), showing that genetic effects on miRNA expression are much more widespread than the previously identified loci[15]. To complement previous studies of miRNA function in cell perturbation experiments, we analysed miRNA–mRNA interactions in our steady-state population sample. Of 100 miRNA families, 32 correlated with the expression of predicted target exons in a highly connected network ($P < 0.001$; Fig. 1d and Supplementary Table 4), including miRNA families with important immunological or lymphocyte functions, such as miR-150, miR-155, miR-181 and miR-146 (ref. 16). Interestingly, 45% of the associations were positive—consistent with previous results[15]—even though based on perturbation experiments miRNAs mostly downregulate genes. Analysing the direction of causality, *cis*-mirQTLs had small *trans*-eQTL effects to predicted targets only when effects were negative ($\Pi_1 = 1 - $ Storey's $\Pi_0 = 0.11$ versus $\Pi_1 = 0$; Supplementary Fig. 16), suggesting that miRNAs indeed downregulate their targets. Positive correlations may be driven by other effects, which is supported by overrepresentation of transcription factors in the network (29%, Fisher $P = 2.1 \times 10^{-7}$ for negative targets and 26% $P = 4.0 \times 10^{-4}$ for positive targets). This suggests feedback loops of both mRNA and miRNA genes affecting the expression of each other, and supports the idea that under steady-state conditions, miRNAs confer robustness to expression programs[17]. Altogether, these results highlight the added insight into the role of miRNAs in regulatory networks from analysis of population variation.

Table 1 | Numbers of transcriptome features with a QTL (FDR 5%)

	Total	EUR ($n = 373$)	YRI ($n = 89$)	Union
Exon eQTL	12,981 genes	7,390	2,369	7,825
Gene eQTL	13,703 genes	3,259	501	3,773
Transcript ratio QTL	7,855 genes	620	83	639
mirQTL	644 miRNAs	57	15	60
Transcribed repeat eQTL	43,875 repeats	5,763	1,055	6,069

FDR, false discovery rate.

Genetic effects on the transcriptome

Expression QTL (eQTL) analysis of protein-coding and long intergenic non-coding RNA (lincRNA) genes uncovered extremely widespread regulatory variation, with 3,773 genes having a classical eQTL for gene expression levels (Table 1). Although the potential of RNA-seq to discover other transcriptome traits such as splicing variation is widely known[7,8,18–20], a comprehensive analysis has been lacking. To this end, we first mapped eQTLs for exon quantifications that can capture both gene expression and splicing variation, discovering as many as 7,825 genes with an eQTL, referred to as eQTLs in this paper unless otherwise specified. Regressing out the most significantly associated variant from the EUR eQTL analysis showed that as many as 34% of the genes have a second, independent eQTL for any of their exons (of which 7% for the exon of the first association). Thus, there is substantial allelic heterogeneity for regulatory effects on a single gene and independence of exons of the same gene (Supplementary Fig. 17). To investigate genetic effects specifically on splicing, we discovered 639 genes with transcript ratio QTLs (trQTLs) affecting the ratio of each transcript to the gene total—the largest number of genetic effects on transcript structure identified so far. The lower number relative to gene eQTLs is probably caused by higher noise in model-based transcript quantifications than in gene counts. To characterize the relationship of genetic variants affecting expression versus splicing, we regressed out the best trQTL variant from the gene eQTL analysis in 279 genes with both types of QTL. The results showed that the causal variants are independent in ≥57% of these genes (Supplementary Fig. 18), suggesting that transcriptional activity and transcript usage are usually controlled by different regulatory elements of the genome.

The transcript differences driven by trQTLs involve exon skipping only in 15% of genes, with as much as 48% and 43% varying in 5′ and 3′ ends, respectively (in EUR; categories not mutually exclusive; Fig. 2b). To analyse transcript modifications further through unannotated transcript elements, we mapped *cis*-eQTLs for expressed retrotransposon-derived elements (repeat elements) outside genes, known to be an important source for evolution of new transcripts[21]. We detected widespread sharing between the 5,763 *cis*-eQTLs discovered for repeat elements (Table 1 and Supplementary Fig. 19) and nearby exon eQTLs: of the best repeat eQTLs variants in EUR, 49% were significant and 6% the top eQTL variants for exons of a nearby gene (3.8× and 26× enrichment; Fisher $P < 2.2 \times 10^{-16}$). This suggests that retrotransposon-derived elements can share regulatory elements with nearby genes. These results provide

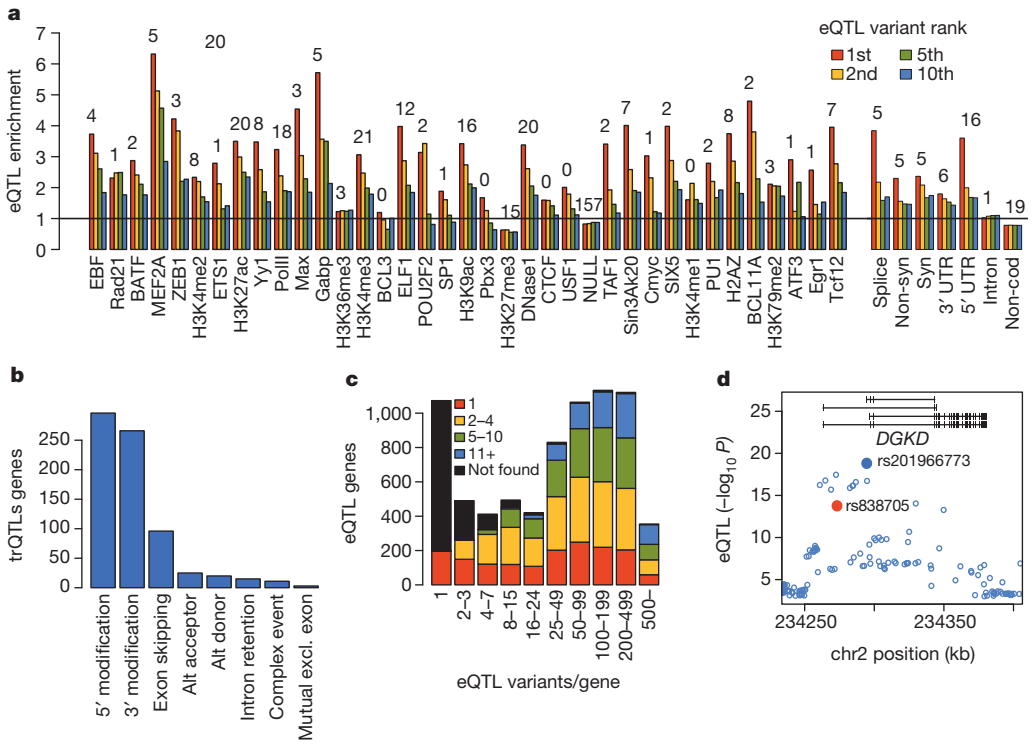

Figure 2 | Transcriptome QTLs. **a**, Enrichment of EUR exon eQTLs in functional annotations for the first, second, fifth and tenth best associating eQTL variant per gene, relative to a matched null set of variants denoted by the horizontal line. The numbers are $-\log_{10}(P$ values) of a Fisher test between the best eQTL and the null. UTR, untranslated region. **b**, Classification of changes caused by transcript ratio QTLs. **c**, The rank of the best Omni 2.5M SNP among the significant EUR eQTL variants per gene. **d**, The *DGKD* gene locus, in which an intronic SNP rs838705 is associated with calcium levels (red), and the top eQTL variant 21 kb downstream (blue) is a very likely causal variant, close to the TSS of two transcripts in the MEF2A,C binding region.

the first, to our knowledge, genome-wide characterization of genetic effects on transcript structure through annotated and unannotated 3′ and 5′ changes, which may predominate the exon skipping that previous studies have focused on[19]. This opens new perspectives for understanding their cellular and high-level effects, as end modifications will rarely change protein structure but may affect post-transcriptional regulation.

Altogether, we present the largest and the most diverse catalogue of *cis*-regulatory variants discovered in a single tissue so far. Most of the analysed genes—8,329 out of 13,970—have one or several QTLs for different transcript traits, a resolution enabled by in-depth analysis of high-quality transcriptome and genome sequencing data. These results highlight both allelic heterogeneity of regulatory variants and phenotypic heterogeneity of diverse transcriptome traits of individual genes.

Properties of regulatory variants

To understand how eQTLs affect gene expression, we compared the properties of the top (most significant) eQTL variant per gene to a null of non-eQTL variants (matched for distance from transcription start site (TSS) and minor allele frequency). The best eQTL variant may not always be the causal variant owing to noise in genotype and phenotype data, and to estimate our ability to pinpoint causal variants, we contrasted the properties of the first eQTL to the second, fifth and tenth best eQTL variants (Fig. 2a).

First, comparing the eQTL with the best P value to the matched null showed an enrichment of indels among top eQTLs (13% = 1.22× enrichment; Fisher $P = 1.9 \times 10^{-3}$ in EUR; Supplementary Fig. 20), suggesting that indels are more likely to have functional effects than SNPs. eQTLs are highly enriched in several non-coding elements from the Ensembl Regulatory Build, such as many transcription factor peaks (median enrichment 3.3×, median $P = 0.009$ in EUR; Fig. 2a and Supplementary Fig. 21), DNase1 hypersensitive sites (3.4×, $P = 1.00 \times 10^{-20}$), as well as in chromatin states of active promoters (3.5×, $P = 1.08 \times 10^{-36}$) and strong enhancers (median 2.4×, median $P = 1.14 \times 10^{-5}$). Within genes, splice-site (3.8×, $P = 1.65 \times 10^{-5}$) and non-synonymous (2.3×, $P = 4.84 \times 10^{-6}$) enrichments point to putative regulatory functions of coding variants. Transcript ratio QTLs are overrepresented in splice sites (6.8×, $P = 2.44 \times 10^{-7}$; Supplementary Fig. 22), as expected, but also, for example, in 3′ untranslated regions (2.5×, $P = 1.83 \times 10^{-6}$) and promoters (2.4×, $P = 5.79 \times 10^{-6}$). Altogether, the higher resolution of annotations and eQTLs relative to previous studies[22,23] provides important insight into the role of individual transcription factors and other regulatory elements mediating genetic regulatory effects.

Functional enrichment typically decreases rapidly from the best eQTL variant towards lower ranks. To estimate how often the first variant is likely to be the causal regulatory variant, we calculated the annotation enrichment of the best eQTL variants relative to the null for (1) all eQTL loci, and (2) loci in which the best eQTL variant is very likely causal owing to having a $\log_{10}$ P-value >1.5 higher than the second variant (Supplementary Fig. 23). The ratio of the enrichments (1) and (2) yields an approximation of the best variant being causal in 55% of EUR and 74% of YRI eQTLs, with more conservative estimates being 34% and 41%, respectively (Supplementary Fig. 23). Thus, we have reasonable power to pinpoint causal regulatory variants from unbiased P-value distributions alone without annotation priors[23]. This is enabled by not relying on SNP array data[22]: in 81% of the cases the best variant is not on the Omni 2.5M array (Fig. 2c and Supplementary Fig. 25). Validating the putative causal effects, we observed that the best eQTL variants in CTCF peaks showed more allele-specific binding compared to matched null variants ($P = 2.0 \times 10^{-3}$; Supplementary Fig. 24) using CTCF ChIP-seq data from six individuals[24], and the best eQTLs were enriched in DNase1 hypersensitivity QTLs[25] (3.3×, $P = 2.51 \times 10^{-6}$ in EUR, 7.9×, $P < 2.2 \times 10^{-16}$ in YRI). In conclusion, we not only identify broad eQTL loci but also substantially increase our confidence to pinpoint individual causal variants and their functional mechanisms.

Of the 6,473 variants in the GWAS catalogue[26], 16% are eQTLs and 1.8% are trQTLs in EUR or YRI, but a high overlap is observed also by

chance for a frequency-matched GWAS null (11% and 0.84%, respectively). The modest (albeit significant: eQTL χ^2 $P < 2.2 \times 10^{-16}$; trQTL $P = 7.2 \times 10^{-9}$) enrichment[9,10] is due to eQTLs being very ubiquitous, and consequently, a GWAS variant being an eQTL does not mean that the regulatory change is necessarily driving the disease association. Our data offers a unique opportunity to address the key question of whether the causal eQTL variant is also causal for the disease. The enrichment of GWAS SNPs in the top eQTL ranks ($P = 1.18 \times 10^{-7}$; Supplementary Fig. 26) is a genome-wide signal of shared causality. To characterize individual loci further, we selected 78 eQTL regions that are likely causal signals for 91 GWAS SNPs (estimated by the regulatory trait concordance method)[6,9], and in these loci our best eQTL variant is the putative disease-causing variant (Supplementary Fig. 27 and Supplementary Table 5). Figure 2d shows an example of the *DGKD* gene, in which an intronic SNP rs838705 is associated with calcium levels[27], and 21 kilobases (kb) downstream the top eQTL—a 2-base pair (bp) insertion—is the likely causal variant affecting calcium levels. Thus, the integration of genome sequencing and cellular phenotype data helps to not only understand causal genes and biological processes but also pinpoint putative causal genetic variants underlying GWAS associations.

Allelic and loss-of-function effects

Transcript differences between the two haplotypes of an individual allow quantification of regulatory variation even when eQTLs cannot be detected, for example, owing to low allele frequency. We analysed both allele-specific expression (ASE) and allele-specific transcript structure (ASTS), a novel approach based on exonic distribution of reads (Supplementary Figs 2 and 28–33). This first genome-wide quantification of allelic effects on transcript structure shows that it is almost equally common as ASE, with significant ($P < 0.005$) ASE and ASTS in a median of 6.5% and 5.6% sites (out of 8,420 and 2,135) per individual, respectively. Furthermore, the substantial overlap of ASE and ASTS signals (Fig. 3a) suggests that ASE may actually often be driven by transcript structure variation. The low population frequency of the vast majority of ASE (Fig. 3b) and ASTS (Supplementary Fig. 30) events points to widespread rare regulatory variation that is undetectable in eQTL analysis.

An important caveat in ASE analysis has been the possibility that it can be driven by purely epigenetic effects rather than *cis*-regulatory genetic variants. We investigated this by a novel approach to quantify concordance between ASE and putative regulatory variants (prSNPs), in which heterozygotes but not homozygotes for a true rSNP should have differential expression of the two haplotypes, that is, allelic imbalance in an aseSNP (Supplementary Figs 2 and 34). We calculated concordance of allelic ratios of 5,479 aseSNPs and genotypes of all variants ± 100 kb from TSS, with an empirical *P* value from 100–1,000 permutations. Assigning the prSNPs with empirical *P*-value < 0.01 to $P < 0.001$ as likely rSNPs yielded a total of 224,640 rSNPs (7.4% of tested; Supplementary Table 6) that clustered close to TSS as expected for regulatory variants[5] and replicated most eQTL signals (Supplementary Fig. 35). Nearly all aseSNPs (95%) had more observed rSNPs than expected; thus ASE seems to nearly always be genetic rather than driven by genotype-independent allelic epigenetic effects. rSNP signals are widespread and robust also outside eQTL genes (Supplementary Table 6 and Supplementary Fig. 35), indicating potential to capture novel effects. Variants that are both eQTLs and rSNPs show higher enrichment in functional annotations (Fig. 3c and Supplementary Fig. 36), suggesting that integrated analysis may improve resolution to find causal regulatory variants. Altogether, we show evidence that ASE effects are mostly rare and nearly always genetic, and ASE-based analyses may complement eQTL analysis in identification of especially low-frequency regulatory variants in future studies.

Although QTL and prSNP analyses aim at identifying previously unknown regulatory variants, we can also quantify functional effects of predicted loss-of-function variants[28]. Our RNA-seq data captures 839 premature stop codon and 849 splice-site variants, with the much higher number than in previous studies enabling proper quantification of their transcriptome effects. As expected, premature stop variants often show loss of the variant allele (Supplementary Fig. 37), indicating nonsense-mediated decay[29] (NMD) as in previous studies[28,30]. Variants close to the end of the transcript seem to escape NMD as predicted[29]. However, of the variants predicted to trigger NMD, in 68% (54% of rare variants with minor allele frequency < 1%) the ASE results do not support this (Fig. 4a), suggesting currently unknown mechanisms of NMD escape.

Finally, we modelled how genetic variants affect splicing affinity in the entire splicing motif rather than only the canonical splice site, which is the first comprehensive set of such predictions genome-wide (P.G.F. *et al.*, manuscript submitted). Non-reference alleles have a lower splicing affinity on average ($P < 2.2 \times 10^{-16}$; Supplementary Fig. 38). For the 10% of these variants predicted to destroy the motif, individuals carrying two motif-destroying alleles have 29% lower median inclusion rates of the affected exon ($P < 2.2 \times 10^{-16}$; Fig. 4b), indicating that our RNA-seq data are consistent with predictions of splicing effects.

Conclusions

By integrated analysis of RNA and DNA sequencing data we were able to obtain a unique view of variation of the transcriptome and its genetic causes, moving beyond eQTL catalogues to a high-resolution view of genetic regulatory variants. We deconvoluted the effect of gene expression and transcript structure in population differences of the transcriptome, in QTLs, and in allele-specific effects, and show that these two dimensions of transcript variation appear equally common but largely independent. Genetic regulatory variation is the rule rather than the

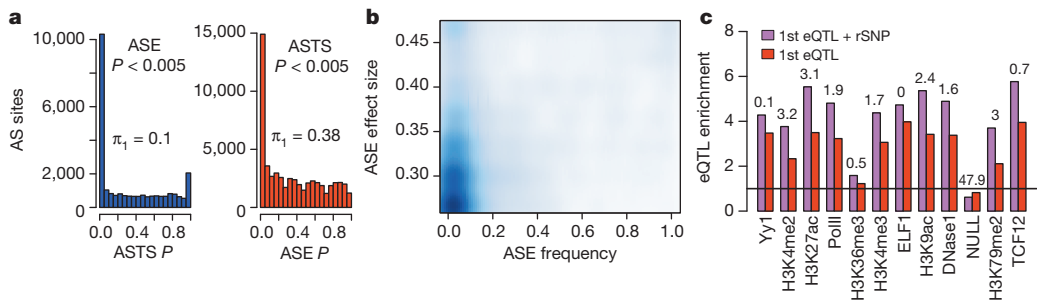

Figure 3 | Allele-specific effects on expression and transcript structure. **a**, Sharing of allele-specific expression (ASE) and transcript structure (ASTS) signals: the distribution of ASTS *P* value of the sites with significant ($P < 0.005$) ASE in the same individual, and vice versa. The ASE *P*-values are calculated from sites sampled to exactly 30 reads. The numbers denote the Π_1 (1 – Storey's Π_0) statistic measuring the enrichment of low *P* values. **b**, Frequency of significant ASE event in the population (*x* axis) and their effect size (|0.5 – ref/total|), calculated per ASE SNP. Only ASE SNPs with ≥20 heterozygote individuals with ≥30 reads were included, and the estimates were corrected for coverage bias and false positives by sampling and permutations. **c**, Enrichment of variants in regulatory annotations relative to a matched null distribution for the most significant eQTL variants, and for the subset of these that are also rSNPs. Categories with highest amount of data are shown (see Supplementary Fig. 36 for all categories, see also Fig. 2a).

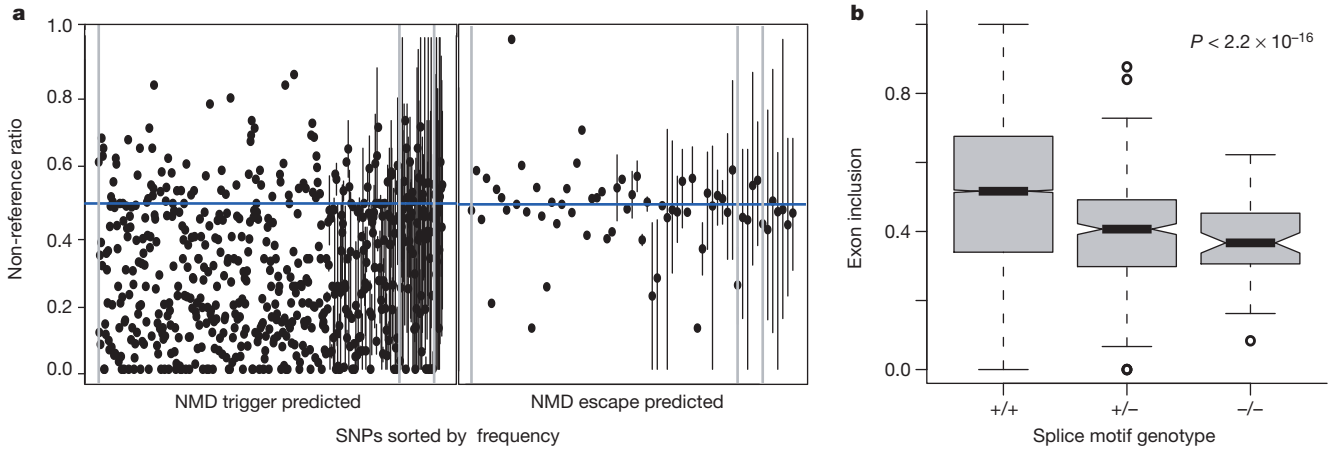

Figure 4 | Transcriptome effects of loss-of-function variants. a, NMD due to premature stop codon variants was measured using allele-specific expression. The distribution of non-reference allele ratios (on the *y* axis) for premature stop variants sorted on the *x* axis according to derived allele frequency, split to sites predicted to trigger and escape NMD. The dots denote the median across individuals, and the vertical lines show the range of ratios for variants carried by several individuals. The grey vertical lines denote derived allele frequencies of 0, 0.001 and 0.01. **b**, Exon inclusion scores for variable exons for individuals that carry 0, 1 or 2 copies of variants that destroy a splice motif, with *P*-value from Mann–Whitney test.

exception in the genome with widespread allelic heterogeneity, and is the major determinant of allelic expression. For the first time, we were able to predict large numbers of causal regulatory variants, and thus provide a detailed view into cellular mechanisms of regulatory and loss-of-function variation, which is essential for future functional prediction of variants discovered in personal genomes.

A subset of this functional variation at the cellular level will also have effects on higher-level traits. We demonstrate how eQTL data can be used to pinpoint putative causal GWAS variants of individual loci, which is important as a new model of how integration of cellular phenotypes and genome sequencing data can uncover both causal variants and biological mechanisms underlying diseases. The landscape of regulatory variation in this study adds a functional dimension to the 1000 Genomes Project data, which is used in effectively all disease studies, and together they form an important joint reference data set of variation and function of the human genome. Ultimately, this study illustrates the power of combining genome sequence analysis with a high-depth functional readout such as the transcriptome.

METHODS SUMMARY

Total RNA was extracted from Epstein–Barr-virus-transformed lymphoblastoid cell line pellets using TRIzol reagent (Ambion), and mRNA and small RNA sequencing of 465 unique individuals were performed on the Illumina HiSeq2000 platform, with paired-end 75-bp mRNA-seq and single-end 36-bp small-RNA-seq. Five samples were sequenced in replicate in each of the seven sequencing laboratories. The mRNA and small RNA reads were mapped with GEM[31] and miraligner[32], respectively, with an average of 48.9 million mRNA-seq reads and 1.2 million miRNA reads per sample after quality control. Numerous transcript features were quantified using Gencode v12 (ref. 33) and miRBase v18 (ref. 34) annotations: protein-coding and lincRNA genes (16,084 detected in >50% of samples), transcripts (67,603; with FluxCapacitor[7]), exons (146,498), annotated splice junctions (129,805; analysed in detail in P.G.F. *et al.*, manuscript submitted), transcribed repetitive elements (47,409), and mature miRNAs (715). Data quality was assessed by sample correlations and read and gene count distributions, and technical variation was removed by PEER normalization[35] for the QTL and miRNA–mRNA correlation analyses[11]. The samples clustered uniformly both before and after normalization. The genotype data was obtained from 1000 Genomes Project Phase 1 data set for 421 samples (80× average exome and 5× whole-genome read depth), and the remaining 41 samples were imputed from Omni 2.5M SNP array data. Furthermore, we did functional reannotation for all the 1000 Genomes Project variants using Gencode v12. QTL mapping was done with linear regression, using genetic variants with >5% frequency in 1-megabase window and normalized quantifications transformed to standard normal. Permutations were used to adjust the false discovery rate to 5%. Full details are provided in the Supplementary Methods.

Received 7 February; accepted 5 August 2013.
Published online 15 September 2013.

1. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
2. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
3. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
4. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
5. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nature Genet.* **39,** 1217–1224 (2007).
6. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nature Genet.* **44,** 1084–1089 (2012).
7. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).
8. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
9. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6,** e1000895 (2010).
10. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).
11. Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotech.* http://dx.doi.org/10.1038/nbt.2702 (in the press).
12. Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res.* **22,** 528–538 (2012).
13. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338,** 1593–1599 (2013).
14. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–1593 (2012).
15. Parts, L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genet* **8,** e1002704 (2012).
16. Xiao, C. & Rajewsky, K. MicroRNA control in the immune system: basic principles. *Cell* **136,** 26–36 (2009).
17. Ebert, M. S. & Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149,** 515–524 (2012).
18. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* **6,** e1001236 (2010).
19. Lee, Y. *et al.* Variants affecting exon skipping contribute to complex traits. *PLoS Genet.* **8,** e1002998 (2012).
20. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
21. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nature Rev. Genet.* **10,** 691–703 (2009).
22. Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4,** e1000214 (2008).
23. Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13,** R7 (2012).
24. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328,** 235–239 (2010).
25. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

26. Hindorff, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies; available at http://www.genome.gov/gwastudies (accessed 11 September 2012).
27. O'Seaghdha, C. M. *et al.* Common variants in the calcium-sensing receptor gene are associated with total serum calcium levels. *Hum. Mol. Genet.* **19,** 4296–4303 (2010).
28. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).
29. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci.* **23,** 198–199 (1998).
30. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7,** e1002144 (2011).
31. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods* **9,** 1185–1188 (2012).
32. Pantano, L., Estivill, X. & Marti, E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.* **38,** e34 (2010).
33. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
34. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39,** D152–D157 (2011).
35. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLOS Comput. Biol.* **6,** e1000770 10.1371/journal.pcbi.1000770 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank E. Falconnet, L. Romano, A. Planchon, D. Bielsen, A. Yurovsky, A. Buil, J. Bryois, A. Nica, I. Topolsky, N. Fusi, S. Waszak, C. Bustamante, J. Rung, N. Kolesnikov, A. Roa, E. Bragin, S. Brent, J. Gonzalez, M. Morell, A. Puig, E. Palumbo, M. Ventayol Garcia, J. F. J. Laros, J. Blanc, R. Birkelund, G. Plaja, M. Ingham, J. Camps, M. Bayes, L. Agueda, A. Gouin, M.-L. Yaspo, E. Graf, A. Walther, C. Fischer, S. Loesecke, B. Schmick, D. Balzereit, S. Dökel, M. Linser, A. Kovacsovics, M. Friskovec, C. von der Lancken, M. Schlapkohl, A. Hellmann, M. Schilhabel, the SNP&SEQ Technology Platform in Uppsala, S. Sauer, the Vital-IT high-performance computing centre of the SIB Swiss Institute of Bioinformatics, B. Goldstein and others at the Coriell Institute, and J. Cooper, E. Burnett, K. Ball and others at the European Collection of Cell Cultures (ECACC) and the 1000 Genomes Consortium. This project was funded by the European Commission 7th Framework Program (FP7) (261123; GEUVADIS); the Swiss National Science Foundation (130326, 130342), the Louis Jeantet Foundation, and ERC (260927) (E.T.D.); NIH-NIMH (MH090941) (E.T.D., M.I.M., R.G.); Spanish Plan Nacional SAF2008-00357 (NOVADIS), the Generalitat de Catalunya AGAUR 2009 SGR-1502, and the Instituto de Salud Carlos III (FIS/FEDER PI11/00733) (X.E.); Spanish Plan Nacional (BIO2011-26205) and ERC (294653) (R.G.); ESGI, READNA (FP7 Health-F4-2008-201418), Spanish Ministry of Economy and Competitiveness (MINECO) and the Generalitat de Catalunya (I.G.G.); DFG Cluster of Excellence Inflammation at Interfaces, the INTERREG4A project HIT-ID, and the BMBF IHEC project DEEP SP 2.3 (P.Ro.); German Centre for Cardiovascular Research (DZHK) and the German Ministry of Education and Research (01GR0802, 01GM0867, 01GR0804, 16EX1020C) (T.M.); EurocanPlatform (FP7 260791), ENGAGE and CAGEKID (241669) (A.B.); FP7/2007-2013, ENGAGE project, HEALTH-F4-2007-201413, and the Centre for Medical Systems Biology within the framework of The Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific and Research (NWO) (P.AC.H and G.-J.v.O.); The Swedish Research Council (C0524801, A028001) and the Knut and Alice Wallenberg Foundation (2011.0073) (A.-C.S.); The Swiss National Science Foundation (127375, 144082) and ERC (249968) (S.E.A.); Instituto de Salud Carlos III (FIS/FEDER PS09/02368) (A.C.); German Federal Ministry of Education and Research (01GS08201) (R.S.); Max Planck Society (H.L.); Wellcome Trust (WT085532) and the European Molecular Biology Laboratory (P.F.); ENGAGE, Wellcome Trust (081917, 090367, 090532, 098381), and Medical Research Council UK (G0601261) (M.I.M.); Wellcome Trust Centre for Human Genetics (090532/Z/09/Z, 075491/Z/04/B), Wellcome Trust (098381, 090367, 076113, 083270), the WTCCC2 project (085475/B/08/Z, 085475/Z/08/Z), Royal Society Wolfson Merit Award, Wellcome Trust Senior Investigator Award (095552/Z/11/Z) (P.D.); EMBO long-term fellowship EMBO-ALTF 2010-337 (H.K.); NIH-NIGMS (R01 GM104371) (D.G.M.); Marie Curie FP7 fellowship (O.S.); Scholarship by the Clarendon Fund of the University of Oxford, and the Nuffield Department of Medicine (M.A.R.); EMBO long-term fellowship ALTF 225-2011 (M.R.F.); Emil Aaltonen Foundation and Academy of Finland fellowships (T.L.).

Author Contributions Designed the study: T.L., T.Gi., S.B.M., P.AC.H., E.L., H.L., S.S., R.S., A.C., S.E.A., R.H., A.-C.S., G.-J.v.O., A.B., T.M., P.Ro., R.G., I.G.G., X.E. and E.T.D. Coordinated the project: T.L., T.Gi., G.B., X.E. and E.T.D. Participated in data production: T.L., T.Gi., I.Pa., M.Su., E.L., S.B., M.G., V.A., K.K., D.E., P.Ri. and O.K. Analysed the data: T.L., M.Sa., M.R.F., P.A.C.H., J.M., M.A.R., M.G.-P., N.K., T.Gr., P.G.F., M.B., T.W., L.G., M.v.I., J.A., P.Ri., I.Pu., D.E., A.T., M.Su., D.G.M., M.L., E.L., H.P.J.B., I.Pa., T.S., O.K., H.O., H.K., S.B., M.G., K.K., V.A., O.S., M.P., P.D., M.I.M., P.F. and T.M.S. Drafted the paper: T.L. and E.T.D. See Supplementary Note for Members of the Geuvadis Consortium.

Author Information The Geuvadis RNA-sequencing data, genotype data, variant annotations, splice scores, quantifications, and QTL results are freely and openly available with no restrictions. The main portal for accessing the data is EBI ArrayExpress, under accessions E-GEUV-1, E-GEUV-2 and E-GEUV-3 (see the data access schema in Supplementary Fig. 39). For visualization of the results we created the Geuvadis Data Browser (http://www.ebi.ac.uk/Tools/geuvadis-das) where quantifications and QTLs can be viewed, searched and downloaded (Supplementary Fig. 40). The project webpage (http://www.geuvadis.org) provides full documentation and links to all files, and the analysis group wiki is open to the public (http://geuvadiswiki.crg.es). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.L. (tuuli.e.lappalainen@gmail.com) or E.T.D. (emmanouil.dermitzakis@unige.ch).