Transcriptome and genome sequencing uncovers functional variation in humans

Tuuli Lappalainen[1,2,3#], Michael Sammeth[4,5*], Marc R Friedländer[5,6*], Peter AC 't Hoen[7*], Jean Monlong[5*], Manuel A Rivas[8*], Mar Gonzàlez-Porta[9], Natalja Kurbatova[9], Thasso Griebel[4], Pedro G Ferreira[5,6], Matthias Barann[10], Thomas Wieland11[1], Liliana Greger[9], Maarten van Iterson[7], Jonas Almlöf[12], Paolo Ribeca[4], Irina Pulyakhina[7], Daniela Esser[10], Thomas Giger[1], Andrew Tikhonov[9], Marc Sultan[13], Gabrielle Bertier[5,6], Daniel G MacArthur[14,15], Monkol Lek[14,15], Esther Lizano[5,6], Henk PJ Buermans[7,16], Ismael Padioleau[1,2,3], Thomas Schwarzmayr[11], Olof Karlberg[12], Halit Ongen[1,2,3], Helena Kilpinen[1,2,3], Sergi Beltran[4], Marta Gut[4], Katja Kahlem[4], Vyacheslav Amstislavskiy[13], Matti Pirinen[8], Stephen B Montgomery[1†], Peter Donnelly[8], Mark I McCarthy[8,17], Paul Flicek[9], Tim M Strom[11,18], The Geuvadis Consortium, Hans Lehrach[13,19], Stefan Schreiber[10], Ralf Sudbrak[13,19], Ángel Carracedo[20], Stylianos E Antonarakis[1,2], Robert Häsler[10], Ann-Christine Syvänen[12], Gert-Jan van Ommen[7], Alvis Brazma[9], Thomas Meitinger[11,18], Philip Rosenstiel[10], Roderic Guigó[5,6], Ivo G Gut[4], Xavier Estivill[5,6], Emmanouil T Dermitzakis[1,2,3#]

* These authors contributed equally to this work

Corresponding authors

1 Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland

2 Institute for Genetics and Genomics in Geneva (1G3), University of Geneva, 1211 Geneva, Switzerland

3 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland

4 Centro Nacional d'Anàlisi Genòmica, 08028 Barcelona, Spain

5 Center for Genomic Regulation (CRG, 08003 Barcelona, Spain

6 Pompeu Fabra University (UPF, 08003 Barcelona, Spain

7 Department for Human and Clinical Genetics, Leiden University Medical Center, 2300 RC Leiden, the Netherlands

8 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdon

9 European Bioinformatics Institute, EMBL-EBI, Hinxton, United Kingdom

10 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany
11 Institute of Human Genetics, Helmholtz Zentrum M√ºnchen, 85764 Neuherberg, Germany
12 Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, 751 85 Uppsala, Sweden
13 Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany
14 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA
15 Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge MA 02142, USA
16 Leiden Genome Technology Center, 2300 RC Leiden, the Netherlands
17 Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford OX3 7BN, United Kingdom
18 Institute of Human Genetics, Technische Universit√§t M√ºnchen, 81675 Munich, Germany
19 Dahlem Centre for Genome Research and Medical Systems Biology, 14195 Berlin, Germany
20 Fundacion Publica Galega de Medicina Xenomica SERGAS, Genomic Medicine Group CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
† Present address: Departments of Pathology and Genetics, Stanford University, Stanford, CA 94305-5324, USA

Summary

Genome sequencing projects are discovering millions of genetic variants in humans, and interpreting their functional effects is essential for understanding the genetic basis of variation in human traits. Towards this goal, we sequenced mRNA and miRNA from lymphoblastoid cell lines of 465 individuals from the 1000 Genomes Project. The integration of RNA and DNA sequencing data allowed us to link gene expression and genetic variation, and to characterize transcriptome variation in several human populations. Our results show that regulatory variants are extremely widespread, affecting expression and transcript structure of most genes. By integration of functional annotation we inferred putative causal variants for regulatory effects as well as for dozens of disease-associations. Analysis of transcriptome effects of predicted loss-of-function variants uncovered mechanisms for splicing and nonsense-mediated decay. Altogether, this study takes us beyond cataloguing putative functional loci

towards predicting the molecular genomic effects of causal variants in the human genome.

Introduction

Interpreting functional consequences of genetic variants is one of the biggest challenges in human genomics, as large genome sequencing studies have revealed tens of millions of variants with mostly unknown effects.[1] While many studies have successfully linked genetic loci to various human phenotypes, e.g. by genome-wide association studies,[2] this usually gives little insight about causal variants and biological mechanisms underlying phenotypic variability and disease susceptibility, despite the improving annotation of the human genome[3]. One approach to address this challenge has been to analyze cellular phenotypes, such as gene expression, resulting in large catalogs of regulatory variants[4-7] known to affect many human diseases and traits.[8,9] A major technical advance in these studies has been RNA sequencing, not only allowing characterization of gene expression levels but also uncovering more complex features of transcriptome variation.[10,11]

The aim of the present study is to characterize transcriptome variation in several human populations by RNA-sequencing of hundreds of individuals from the 1000 Genomes collection. We integrate these data with existing high-quality genome sequencing data[1] of the same individuals in order to uncover not only loci with regulatory variation in the human genome but also putative causal functional variants. This is the biggest mRNA and small RNA sequencing data set of multiple human populations to date, with the appropriate quality and power for a comprehensive characterization regulatory diversity in the genome.

Study design and data set

In this study, we combined transcriptome and genome sequencing data by performing mRNA and small RNA sequencing on 465 lymphoblastoid cell line (LCL) samples from 5 populations of the 1000 Genomes Project: the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI) (Figure S1, Table S1).

Of these samples, 423 were part of the 1000 Genomes Phase 1 dataset[1] with genome sequencing data, and the remaining 42 were imputed from Omni 2.5M SNP array data (Supplementary Methods); furthermore we did functional reannotation for all the 1000 Genomes variants (Table S2). Thus, we obtained a powerful dataset to study transcriptome variation and its genetic causes (Fig. S2, S3). Our openly accessible RNA-seq data also provides a valuable reference data set to the human genomics community (see Data Access section).

As a parallel goal, we sought to assess the feasibility of distributed RNA sequencing. Thus, we performed transcriptome sequencing in seven different European laboratories with randomly allocated RNA samples, using the Illumina HiSeq2000 platform, with paired-end 75bp reads for mRNA-seq and single-end 36bp reads for small RNA-seq. We mapped the reads with GEM[12] and miraligner[13] for mRNA and small RNAs, respectively, resulting in an average of 48.9M well-mapped mRNA-seq reads and 1.2M good-quality micro-RNA (miRNA) reads per sample (Fig. S4; Supplementary Methods). Numerous transcript features were quantified: protein-coding and lincRNA genes (16,084 detected in >50% of samples), transcripts (67,603), exons (146,498), annotated splice junctions (129,805; analyzed in detail in Ferreira et al. submitted), transcribed repetitive elements (47,409), and mature miRNAs (715) (Table S3, Supplementary methods). In the final data set after quality control (Supplementary Methods), we had 462 and 452 individuals (89-95 samples per population and 373 and 363 total European samples) with good-quality mRNA and miRNA sequencing data, respectively (Table S1). The samples had uniform clustering both before and after normalization (Fig. S4-8; 't Hoen et al. submitted). Five samples were prepared and RNA-sequenced in replicate in each of the seven laboratories and demonstrated significantly less technical variation among laboratories than biological variation for both mRNA and miRNA data (Mann-Whitney $p = < 2.2 \times 10^{-16}$ for mRNA, $p = 1.34 \times 10^{-10}$ for miRNA; Fig. 1a, S9; 't Hoen et al. submitted). This indicates that, if proper attention is paid to harmonization of protocols and assignment of samples to avoid confounding, RNA sequencing is a mature and robust technology ready for big studies with distributed data production.

Combining transcriptome and genome sequencing data enables powerful analysis of regulatory genetic variants. To this end, we mapped cis-QTLs to different transcriptome traits of protein-coding and miRNA genes using linear regression separately in the European (EUR) and Yoruba (YRI) populations using genetic variants with >5% frequency in 1MB window and properly normalized quantifications. Permutations were used to adjust FDR to 5%. See Supplementary Methods for details.

Transcriptome variation in human populations

Individual differences in transcription can manifest in overall expression levels or relative abundance of transcripts from the same gene (transcript ratios). Statistical deconvolution of the relative contribution of these two to the overall variability of transcript levels[14] shows the variation in alternative transcript usage as the as the dominating factor in most genes (Fig.1b, S10, S11a). The proportion of variation explained by expression levels differs markedly among genes but in highly consistent manner in different populations (rho=0.84-0.87, Fig.S11b), which indicates that this pattern of variation is characteristic for each gene and not due to random noise. As expected, the vast majority of the total transcription variation is among individuals within populations, with a small but significant proportion of 3% explained by population differences (compared to random grouping MW $p < 2.2 \times 10^{-16}$). In addition to this genome-wide perspective to population variation, differential expression analysis allowed us to define 263-4379 genes with significant differences in expression levels and/or transcript ratios between population pairs (Supplementary Methods, Ferreira et al. in preparation). Interestingly, YRI-EUR population pairs have much higher proportions of genes with different transcript usage than European population pairs (75-85% versus 6-40%; Fig. 1c supported by two independent analyses in Fig. S11c,d). This suggests that transcript structure variation may contribute disproportionally to differences between continental human populations, which is consistent with splicing patterns capturing phylogenetic differences between species better than expression levels [15,16].

We quantify a total of 644 autosomal miRNAs in >50% individuals, with the 29 highest expressed accounting for 90% of the total miRNA reads (Fig. S12). We find significant cis-mirQTLs for 60 out of the 644 miRNAs (Table 1), showing that genetic effects on miRNA expression are more common than reported in previous studies[17]. Thus far, miRNA function has been mostly studied in cell pertubation experiments (reviewed in [18]), but our data set allowed us to correlate miRNA and target mRNA expression under steady-state conditions in a population sample. To this end, the expression levels of miRNAs with the same seed sequence were first summed into 100 miRNA families, of which 32 were significantly associated with the expression of their set of predicted target exons in a highly connected network ($P<0.001$, global test with Holm's correction for multiple testing, Fig. 1d, Table S4). These include miRNA families with important functions in lymphocytes, such as miR-150, miR-155 and miR-181, and associations involved in immune reactions such as miR-146 and IRAK1[19]. Interestingly, 45% of the significant associations were positive – consistently to previous results [17] – even though miRNAs are believed to mostly downregulate genes. To understand the direction of causality, we analyzed trans-eQTL effects of cis-mirQTLs on all the predicted targets of respective miRNAs (Fig.S13). While we have limited power to pinpoint specific trans-eQTL variants, we observe a small enrichment of low p-values only for negative trans-effects (pi1 = 0.11, versus pi1 = 0 for all other target sets). This suggests that miRNAs indeed downregulate their targets while the positive correlations may be driven by other causal mechanisms, such as transcription factors upregulating the expression of miRNAs. Consistent with this, significant targets in our data are enriched for regulators of transcription (29%, Fisher $p= 2.1 \times 10^{-7}$ for negative targets and 26% $p=4.0 \times 10^{-4}$ for positive targets), and even more when a target is associated with multiple miRNAs (43% vs 23% for single-miRNA targets, Fisher p = 0.0057). When the target of a given miRNA induces transcription of the miRNA, the expected effect is a negative feedback circuit, stabilizing the expression of both the miRNA and the transcription factor [20]. Observing several such circuits supports the idea that a function of miRNAs under steady-state conditions is to confer robustness to expression programs[21]. [22]

Transcriptome QTLs in coding genes

We discovered 7,825 genes with an expression QTL (eQTL) using exon quantifications as the quantitative trait and correcting for the number of exons per gene (out of total 12,981 genes; Table 1), and we refer to these as eQTLs unless otherwise specified. Regressing out the most significantly associated variant from the EUR eQTL data leaves 7.0% of the exons of the first association with another independent eQTL, while as much as 34% of the genes have a second eQTL for any of their exons, independent from the first association. This indicates substantial genetic independence between exons of the same gene (Fig.S14), likely driven by splicing variation, which exon eQTLs can capture in addition to overall gene expression level changes. Indeed, using total gene expression level (RPKM) as the quantitative trait yields only 3,773 gene eQTLs out of 13,703 genes. To characterize genetic effects specifically on splicing, we mapped transcript ratio QTLs (trQTLs) using the ratio of each transcript to the total gene quantification, finding a total of 639 genes with a trQTL out of 7,855 analyzed genes. The lower number relative to gene eQTLs is likely caused by the trQTL analysis relying on inferred model-based transcript quantifications, which are noisier and thus result in lower power than total observed expression level quantifications. We further analyzed the overlap of gene eQTL and trQTL signals in the EUR sample for the 7,855 genes. Even though the quantification of total gene expression level and transcript ratios are independent, we find a significant enrichment of genes with both types of QTL (279 genes = 45% of trQTL genes = 2.15× enrichment, χ^2 $p < 2.2 \times 10^{-16}$). Interestingly, regressing out the best trQTL from the gene eQTL analysis showed that in a minimum of 57% of the shared genes, the gene eQTL and trQTL causal variants are not correlated (Fig. S15). These analyses imply that while genes may have variants affecting both transcriptional activity and transcript choice, these traits are usually controlled by different genetic elements.

Importantly, the transcript differences driven by trQTLs involve splicing changes only in 15% of genes, with as much as 48% and 43% varying in 5' and 3' ends, respectively (in EUR; note that one trQTL gene can belong to several categories; Fig.2b). To further analyze transcript modifications through

unannotated transcript elements, we performed cis-eQTL analysis on expressed retrotransposon-derived elements outside genes, known to be an important source for evolution of new transcripts.[23] We detected a total of 5,763 repeat elements with significant cis-eQTLs (Fig.S16, Table 1) that were frequently shared with eQTLs of exons: of the best repeat eQTLs variants in EUR, 49% were also significant exon eQTLs, and 6% were the best exon eQTL variants (3.8× and 26× enrichment; Fisher exact test $p<2.2 \times 10^{-16}$ for both, compared to a set of null variants matched to repeat eQTL frequency and distance from genes). This suggests that repeat elements can be controlled by the same regulatory elements and genetic variants as nearby genes. Together, these results imply that genetic effects on transcript structure through annotated and unannotated 3' and 5' changes are widespread and likely much more common than true splicing QTLs. Predicting the cellular effects of these variants will be challenging as they rarely change protein structure but are likely to be relevant for post-transcriptional regulation.

Altogether, the results of the transcriptome QTL analysis demonstrate substantial regulatory complexity of individual genes. Regulatory variation in cis is extremely widespread in the genome, with the majority of the genes in our sample – 8,329 out of 13,970 analyzed genes – having QTLs either for exon or gene expression level or for transcript ratio.

Characterization of regulatory variants

Genome sequencing data gives us a unique opportunity to characterize causal regulatory variants and functional mechanisms underlying eQTL effects on expression. To understand why eQTLs affect gene expression, we compared the best, i.e. the most significant eQTL variant per gene to a null distribution of non-eQTL variants (matched for distance from transcription start site and minor allele frequency). In an ideal scenario the best eQTL variant would be the causal variant, but this is not always true due to the incomplete list of variants, variance in genotype calling methods, and noise in the phenotypic data. Thus, to estimate how close we are to finding causal variants we contrasted the properties of the

1st eQTL to the 2nd, 5th and 10th best eQTL variants (Fig. 2a; Supplementary Methods). In the following, we describe results of exon eQTLs from Europeans unless otherwise mentioned; results from Yoruba and from trQTLs are shown in Supplementary Figures.

First, comparing the eQTL with the best p-value to the matched null showed that in 13% of the genes the strongest eQTL variant is an indel, which is 1.22× more than for the matched null variants (Fisher exact test $p = 1.9 \times 10^{-3}$; Fig.S17). This suggests that indels are more likely to have functional effects than SNPs. eQTLs are highly enriched in several noncoding elements from the Ensembl Regulatory Build, such as many transcription factor peaks (median enrichment 3.3×, median p = 0.009), histone marks (median 2.3×, median $p = 8.72 \times 10^{-9}$), DNase1 hypersensitive sites (3.4×, $p = 1.00 \times 10^{-20}$), and a high enrichment in chromatin states of active promoters (3.5x, $p = 1.08 \times 10^{-36}$) and strong enhancers (median 2.4×, median $p = 1.14 \times 10^{-5}$) (Fig.2a, S18). Within genes, best eQTL variants were highly enriched in 5'UTRs (3.6x, $p = 3.23 \times 10^{-17}$) consistently with the neighboring promoter activity, while splice-site (3.8×, $p = 1.65 \times 10^{-5}$) and nonsynonymous (2.3×, $p = 4.84 \times 10^{-6}$) enrichments point to coding variants having also regulatory functions. Transcript ratio QTLs are overrepresented in splice sites (6.8×, $p = 2.44 \times 10^{-7}$), as expected, but also for example in 3'UTRs (2.5×, $p = 1.83 \times 10^{-6}$) and promoters (2.4×, $p = 5.79 \times 10^{-6}$) (Fig. S19). Altogether, these analyses add to previous understanding[24,25] of the functional consequences of genetic variants to transcription.

These analyses show a characteristic pattern of highest enrichment for the best eQTL variant, and a rapidly decreasing trend towards lower ranks. With this information, we were able to estimate how likely the first variant is to be also the causal regulatory variant: we calculated the annotation enrichment of the best eQTL variants relative to the matched null for (1) all eQTL loci, and (2) those loci where we reasoned that the best eQTL variant is the causal due to having a log 10 p-value >1.5 higher than the second variant; this threshold was based on saturation of the functional enrichment of the best (1st) versus the 2nd eQTL variant (Fig.S20; Supplementary Methods). From the ratio of the enrichments (1) and (2), we obtain an approximate estimate of the best variant

being the causal variant in 55% of EUR eQTLs and 74% of YRI eQTLs, with more conservative estimates being 34% and 41%, respectively (Fig.S20), indicating reasonable power to pinpoint causal regulatory variants. To validate the putative causal effects of a subset of eQTL variants from both EUR and YRI, we analyzed allele-specific binding in a published dataset of CTCF ChIP-seq data from 6 individuals[26], observing that the best eQTL variants overlapping CTCF binding peaks show significantly more allele-specific binding compared to matched null variants ($p = 2.0 \times 10^{-3}$, Fig.S21). Furthermore, the best eQTLs showed a high degree of overlap with DNase1 sensitivity QTLs[24] (3.3×, $p = 2.51 \times 10^{-6}$ in EUR, 7.9×, $p < 2.2 \times 10^{-16}$ in YRI). Not relying on SNP array data[27] is an important advantage in the characterization of causal variants: 15% of the eQTL genes do not have a single Omni 2.5M array SNP among the significant variants, and in 81% the best (i.e. most likely causal) variant is not on the Omni 2.5M chip (Fig.2c, Fig.S22).

Our eQTL variants also have significant enrichment of GWAS variants [28] (16% of 6,473 GWAS variants are eQTLs in EUR or YRI, versus 11% of a frequency-matched GWAS null distribution; chi^2 $p < 2.2 \times 10^{-16}$; for trQTLs the percentages are 1.8% and 0.84%, respectively, $p = 7.2 \times 10^{-9}$). The large proportion of overlap even under the null warrants caution against inference of causality from a mere overlap of eQTL-GWAS signals; however, the significant enrichment of GWAS SNPs in the top eQTL ranks ($p=1.18 \times 10^{-7}$; Fig. S23) suggests that eQTL effects indeed have increased probability to be causal GWAS mechanisms[8,9]. In such loci, the causal eQTL variant is also the likely to be causal for the disease association. To find such loci, we selected 78 strictly filtered eQTL regions that have been statistically estimated to underlie 92 GWAS SNPs using the RTC method,[7,8] and for each of these loci we pinpoint the best eQTL variant in our data as the putative disease-causing causal variant (Fig.S24, Table S5). Figure 2d shows an example of the DGKD gene locus where an intronic SNP rs838705 is associated to calcium levels[29], and 21 kb downstream we find a top eQTL variant – an insertion of two nucleotides – located close the start site of two transcripts and in the middle of regulatory regions for H3K27ac, MEF2A, MEF2C, and DNase1.

Allele-specific transcription

RNA sequencing enables analysis of potential regulatory differences between the two haplotypes of an individual. We analyzed allele-specific expression (ASE) – imbalance between the two heterozygous alleles – as well as allele-specific transcript structure (ASTS), which is a novel approach that captures differences in the exonic distribution of reads from the two haplotypes (Fig.S2, S25, S26, Supplementary Methods). In our dataset, we analyzed ASE and ASTS in a median of 8,420 and 2,135 sites per individual, respectively, of which 6.5% and 5.6% have a significant signal ($p<0.005$) of allelic difference. The substantial overlap of ASE and ASTS signals (Fig.3a, see also Fig. S31d) suggests that ASE is actually often driven by transcript structure variation rather than different overall expression levels of the two haplotypes. The vast majority of significant ASE (Fig. 3b) and ASTS (Fig. S27) events are rare in the population, and ASE data shows a clear population structure (Fig. S28, S29).

In order to quantify how well ASE effects capture genetic regulatory variation (as opposed to direct loss-of-function effects, Fig. S30), we developed a novel approach to map cis-regulatory variants (rSNPs) underlying ASE signals. Here, we search for putative rSNPs (prSNPs) in the cis-regulatory region with genotypes concordant with allelic ratios of an aseSNP: an individual heterozygous for a true rSNP should have differential expression of the two haplotypes in the affected gene, i.e. a significant deviation from the RNAseq allelic ratio of 0.5 over an aseSNP, whereas an rSNP homozygote should have two equally expressed haplotypes and an aseSNP allelic ratio close to 0.5 (see Supplementary Methods, Fig. S2, S31). Based on this principle, for each aseSNP (passing stringent filters, see Supplementary Methods), we calculated for all prSNPs (+/-100kb window from TSS) if the prSNP genotypes were concordant with the aseSNP allelic ratios across individuals. For each aseSNP-prSNP pair, an empirical p-value was calculated by permuting prSNP genotype status up to 1000 times. The results show a clear enrichment of low p-values (pi1=0.16, Fig. S32), and assigning the prSNPs with empirical p-value<0.01 to $p<0.001$ as likely rSNPs yields a total of 224,640 rSNPs out of the 3,044,486 tested (7.4%, Table S6), with clustering close to TSS as expected for true cis-regulatory variants[5]

(Fig. S32c). Of the 5,479 analyzed aseSNPs, 5,216 (95%) had more observed rSNPs than expected, and this signal as well as the TSS enrichment remain even after excluding eQTL genes (Table S6, Fig.S32). eQTL variants are highly enriched in low prSNP p-values, thus independently replicating the majority of tested eQTL signals (pi1 = 0.55; Fig.32). eQTLs that are also rSNPs show an increased enrichment in functional annotations (Fig. S3c, S33), suggesting that regulatory variants with both eQTL and ASE-based evidence are more likely to be causal variants. Together, these results suggest that the vast majority of allele-specific expression is likely to be genetically driven, and that our ASE-based approach only appears to add resolution to eQTL results but also captures additional effects not seen in eQTL analysis. With the measurement of regulatory effects at the level of an individual rather than quantitatively in a population, this ASE approach may be powerful in identifying low-frequency regulatory variants in future studies with bigger sample sizes or family-based study designs.

Interpretation of loss-of-function variation

While QTL and prSNP analyses aim at identifying previously unknown regulatory variants, we can also observe and quantify functional effects of Loss-of-Function (LoF) variants.[30] Our samples have 2,987 premature stop codon variants and 4,090 splice-site variants of which 839 and 849, respectively, are captured by our RNA-seq data. As expected, premature stop variants show a high enrichment for ASE with loss of the variant allele (Fig, 4a, S34), nonsense-mediated decay[31] as in previous studies[30,32]. Variants close to the end of the transcript are predicted to escape NMD[31], which we generally observe. However, of the variants predicted to cause NMD, only 32% (54% of rare variants MAF<1%) are ASE outliers, while in 68% (46%) the ASE results do not support NMD (Fig.4b,c). This suggests currently unknown mechanisms of NMD escape.

To quantify the effect of splicing variants, we calculated a score of the predicted change in splicing efficiency for all genetic variants overlapping an annotated splicing motif (Ferreira et al. submitted). On average, nonreference variants have lower splicing affinity ($p<2.2 \times 10^{-16}$, Fig. S35), and 10% are predicted to destroy the motif. This effect can be validated by our RNA

sequencing data: individuals carrying two predicted splice-motif-destroying alleles have 29% lower median inclusion rates of the affected exon ($p<2.2 \times 10^{-16}$, Fig.4d), indicating that the variants indeed lead to exon skipping.

Conclusions

By integrated analysis of RNA and DNA sequencing data we were able to obtain a unique high-precision view to variation of the transcriptome and its genetic causes. Comparing the contribution of gene expression *versus* transcript structure variation in the transcriptome, in QTL discoveries, and in allele-specific effects shows that a large proportion of genes exhibit variation in both dimensions, with both independent and shared genetics effects on different features of a gene. Altogether, detection of thousands of transcriptome QTLs and widespread allelic expression that is mostly driven by genetic effects shows that it is the rule rather than the exception that a gene is affected by genetic regulatory variation segregating in human populations. Furthermore, we demonstrated how transcriptome analysis of mostly rare loss-of-function variants can improve our understanding of this class of variants with high functional relevance. For the first time, we were able to predict large numbers of causal regulatory variants, which is a key to real understanding of the cellular mechanisms of regulatory variation of the genome, and its potential downstream effects on phenotypic variability. Furthermore, our search for causal variants for cellular phenotypes from genome sequencing data brings us closer to developing methods for functional interpretation of personal genomes also for phenotypes above the cellular level. Ultimately, this study illustrates the power of combining genome sequence analysis with a high-depth functional readout such the transcriptome.

Methods summary

Total RNA was extracted from EVB transformed lymphoblastoid cell line pellets by the TRIzol reagent (Ambion), and mRNA and small RNA sequencing of 465

unique individuals was performed on the Illumina HiSeq2000 platform in seven European laboratories, each processing 48-113 randomly assigned samples. Five samples were sequenced in replicate in each of the labs for both mRNA and miRNA, and twice in UNIGE for mRNA. The mRNA reads were mapped with GEM[12] to hg19, Gencode v12[33] exons were quantified from read overlaps, and FluxCapacitor[10] was used for transcript quantification. Small RNA data was mapped and quantified with miraligner[13] using miRBase v18[34]. Quantifications were normalized by the total number of well-mapped reads. 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.

References

1 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:nature11632 [pii]
10.1038/nature11632 (2012).

2 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:nature05911 [pii]
10.1038/nature05911 (2007).

3 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:nature11247 [pii]
10.1038/nature11247 (2012).

4 Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-428, doi:nature06758 [pii]
10.1038/nature06758 (2008).

5 Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-1224, doi:ng2142 [pii]
10.1038/ng2142 (2007).

6 Liang, D. M. & Qiao, J. J. [Recent advances in the study of macrolide glycosyltransferases]. *Yao Xue Xue Bao* **42**, 455-462 (2007).

7 Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-1089, doi:ng.2394 [pii]
10.1038/ng.2394 (2012).

8 Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895, doi:10.1371/journal.pgen.1000895 (2010).

9 Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888, doi:10.1371/journal.pgen.1000888 (2010).
10 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:nature08903 [pii]
10.1038/nature08903 (2010).
11 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772, doi:nature08872 [pii]
10.1038/nature08872 (2010).
12 Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, doi:nmeth.2221 [pii]
10.1038/nmeth.2221 (2012).
13 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, doi:gkp1127 [pii]
10.1093/nar/gkp1127 (2010).
14 Gonzalez-Porta, M., Calvo, M., Sammeth, M. & Guigo, R. Estimation of alternative splicing variability in human populations. *Genome Res* **22**, 528-538, doi:gr.121947.111 [pii]
10.1101/gr.121947.111 (2012).
15 Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017, doi:gr.133744.111 [pii]
10.1101/gr.133744.111 (2012).
16 Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587-1593, doi:338/6114/1587 [pii]
10.1126/science.1230612 (2012).
17 Parts, L. *et al.* Extent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genet* **8**, e1002704, doi:10.1371/journal.pgen.1002704
PGENETICS-D-12-00282 [pii] (2012).
18 Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* **12**, 99-110, doi:nrg2936 [pii]
10.1038/nrg2936 (2011).
19 Xiao, C. & Rajewsky, K. MicroRNA control in the immune system: basic principles. *Cell* **136**, 26-36, doi:S0092-8674(08)01633-4 [pii]
10.1016/j.cell.2008.12.027 (2009).
20 Wu, C. I., Shen, Y. & Tang, T. Evolution under canalization and the dual roles of microRNAs: a hypothesis. *Genome Res* **19**, 734-743, doi:19/5/734 [pii]
10.1101/gr.084640.108 (2009).
21 Ebert, M. S. & Sharp, P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell* **149**, 515-524, doi:S0092-8674(12)00464-3 [pii]
10.1016/j.cell.2012.04.005 (2012).

22 Murata, T. *et al.* miR-148a is an androgen-responsive microRNA that promotes LNCaP prostate cell growth by repressing its target CAND1 expression. *Prostate Cancer Prostatic Dis* **13**, 356-361, doi:pcan201032 [pii]
10.1038/pcan.2010.32 (2010).
23 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703, doi:nrg2640 [pii]
10.1038/nrg2640 (2009).
24 Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-394, doi:nature10808 [pii]
10.1038/nature10808 (2012).
25 Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol* **13**, R7, doi:gb-2012-13-1-r7 [pii]
10.1186/gb-2012-13-1-r7 (2012).
26 McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235-239, doi:science.1184655 [pii]
10.1126/science.1184655 (2010).
27 Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* **4**, e1000214, doi:10.1371/journal.pgen.1000214 (2008).
28 Hindorff, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. *www.genome.gov/gwastudies* (2010).
29 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, doi:ddq342 [pii]
10.1093/hmg/ddq342 (2010).
30 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:335/6070/823 [pii]
10.1126/science.1215040 (2012).
31 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, doi:S0968-0004(98)01208-0 [pii] (1998).
32 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, doi:10.1371/journal.pgen.1002144
PGENETICS-D-10-00589 [pii] (2011).
33 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:22/9/1760 [pii]
10.1101/gr.135350.111 (2012).
34 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157, doi:gkq1027 [pii]
10.1093/nar/gkq1027 (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, doi:10.1371/journal.pcbi.1000770 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions

Designed the study: TL, TGi, SBM, PACH, EL, HL, SS, RS, AC, SEA, RH, ACS, GJvO, AB, TM, Pro, RG, IGG, XE, ETD

Coordinated the project: TL, TGi, GB, XE, ETD

Participated in data production: TL, TGi, IPa, MSu, EL, SB, MG, VA, KK, DE, PR, OK

Analyzed the data: TL, MSa, MRF, PACH, JM, MAR, MGo, NK, TGr, PGF, MB, TW, LG, MvI, JA, PRi, IPu, DE, AT, MSu, DGM, ML, EL, HPJB, IPa, TS, OK, HO, HK, SB, MGu, KK, VA, MP, PD, MIM, PF, TMS

Drafted the paper: TL, ETD

Data access

The Geuvadis RNA-sequencing data, genotype data, and quantification and QTL files are freely and openly available with no restrictions. The main portal for accessing the data is EBI ArrayExpress (accessions E-GEUV-1, E-GEUV-2, E-GEUV-3; see the data access schema in Fig. S36). For visualisation of the results we created the Geuvadis Data Browser (www.ebi.ac.uk/Tools/geuvadis-das) where quantifications and QTLs can be viewed, searched, and downloaded (Fig. S37). The project webpage www.geuvadis.org provides full documentation and links to all files, and the project wiki is open to the public in geuvadiswiki.crg.es.

Correspondence and requests for materials should be addressed to tuuli.e.lappalainen@gmail.com or emmanouil.dermitzakis@unige.ch

Acknowledgements:

We would like to thank Emilie Falconnet, Luciana Romano, Alexandra Planchon, Deborah Bielsen, Alisa Yurovsky, Alfonso Buil, Julies Bryois, Alexandra Nica, Sebastian Waszak, Johan Rung, Nikolay Kolesnikov, Asier Roa, Eugene Bragin, Simon Brent, Justo Gonzalez, Marta Morell, Anna Puig, Emilio Palumbo, Marina Ventayol Garcia, Jeroen F.J. Laros, Julie Blanc, Rahnehild Birkelund, Gloria Plaja, Matt Ingham, Jordi Camps, Monica Bayes, Lidia Agueda, Anais Gouin, Marie-Laure Yaspo, Elisabeth Graf, Anett Walther, Carola Fischer, Sandy Loesecke, Bianca Schmick, Daniela Balzereit, Simon Dökel, Matthias Linser, Alexander Kovacsovics, Melanie Friskovec, Catharina von der Lancken, Melanie Schlapkohl, Anita Dietsch, Markus Schilhabel, the SNP&SEQ Technology Platform in Uppsala, Sascha Sauer (Max Plank Institute for Human Genetics (MPIMG, Berlin, Germany) for ESGI (European Sequencing and Genotyping Infrastructure) coordination, the Vital-IT high-performance computing center of the SIB Swiss Institute of Bioinformatics, Bernadette Goldstein and others at the Coriell Institute, and James Cooper, Edward Burnett, Karen 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, Project N. 261123 (GEUVADIS), as well as The Swiss National Science Foundation 31003A_130342 and 127375, the Louis Jeantet Foundation, NIH-NIMH MH090941, ERC 249968 grant, The Swedish Research Council (90559401), the Knut and Alice Wallenberg Foundation (2011.0073), the DFG Cluster of Excellence Inflammation at Interfaces, the INTERREG4A project HIT-ID, the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project, grant agreement HEALTH-F4-2007-201413, the Centre for Medical Systems Biology within the framework of The Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO), Spanish Plan Nacional funding SAF2008-00357 (NOVADIS), the Generalitat de Catalunya funding AGAUR 2009 SGR-1502, the Ministry of Economy and Competitiveness funding to CLL-Consortium project, ESGI – European Commission Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 262055, Academy of Finland, Emil Aaltonen Foundation, EMBO long-term fellowship ALTF 225-2011, European Community's FP7 75 CAGEKID

(grant agreement 241669), FIS PS09/02368, and the German Ministry of Education and Research (01GR0802, 01GM0867, 01GR0804).

Tables

Table 1: Quantitative trait locus discovery. Numbers of transcriptome features with a QTL (FDR 5%). Exons and transcripts are collapsed to the gene level. Union refers to nonredundant count in EUR and YRI.

	Total	EUR (n=373)	YRI (n=89)	Union
exon eQTL	12981 genes	7390	2369	7825
gene eQTL	13703 genes	3259	501	3773
transcript ratio QTL	7855 genes	620	83	639
mirQTL	644 miRNAs	57	15	60
repeat eQTL	43875 repeats	5763	1055	6069

Figure Legends

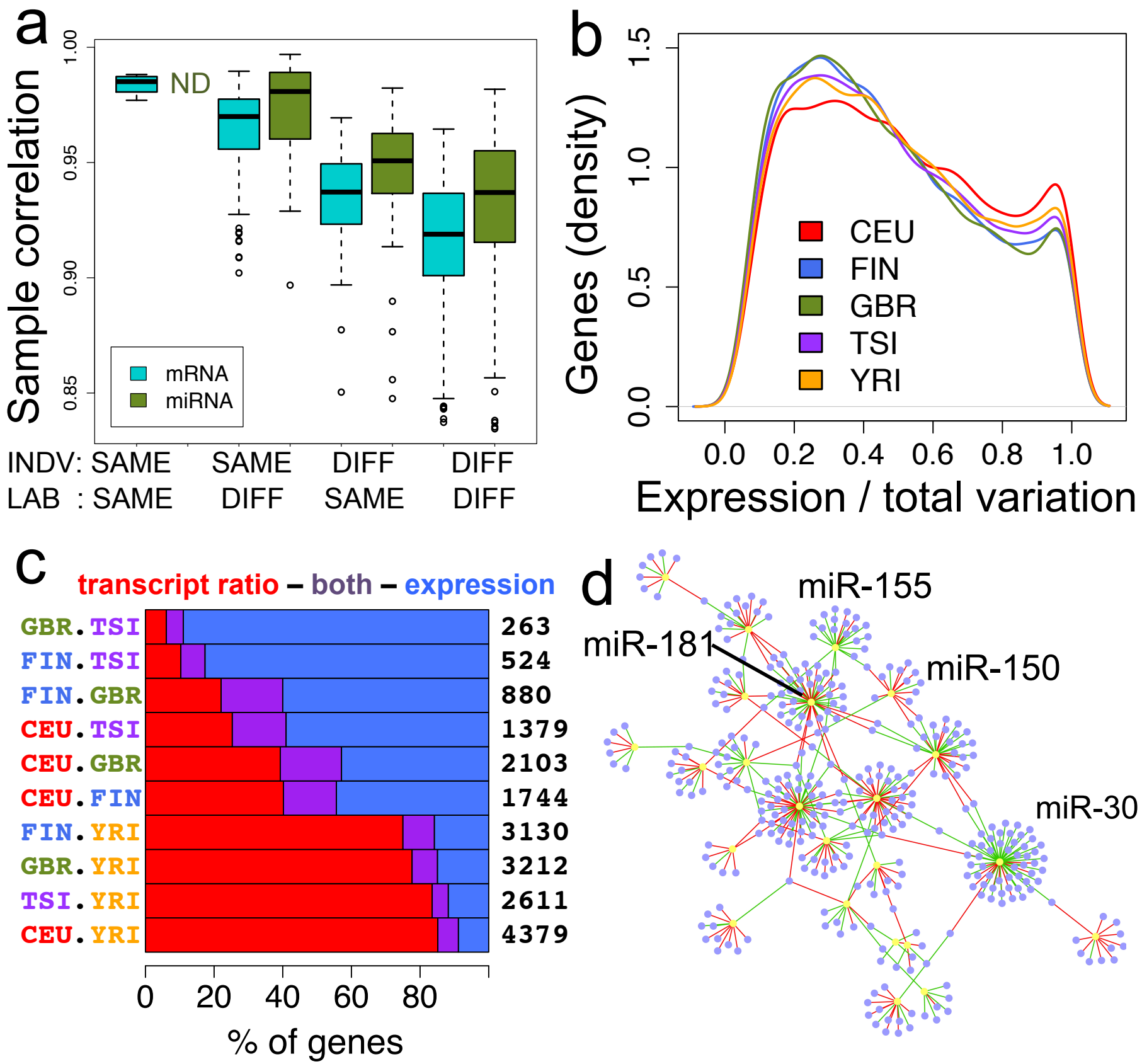
Figure 1: Transcriptome variation. a) Spearman rank correlation of replicate samples, based on mRNA exon and miRNA quantifications of 5 individuals sequenced 8 and 7 times for mRNA and miRNA, respectively. The boxplots are separated by the individual and the sequencing lab being the same or different. The quantifications have been normalized only for the total number of mapped reads (see Fig. S9 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 significant differential expression levels and/or transcript usage between population pairs. The numbers on the right denote the total number of differential genes. d) Network of significant (P<0.001) miRNA families (yellow) and their significantly associated mRNA targets (P<0.05; purple). The edges display negative (green) and positive (red) associations.

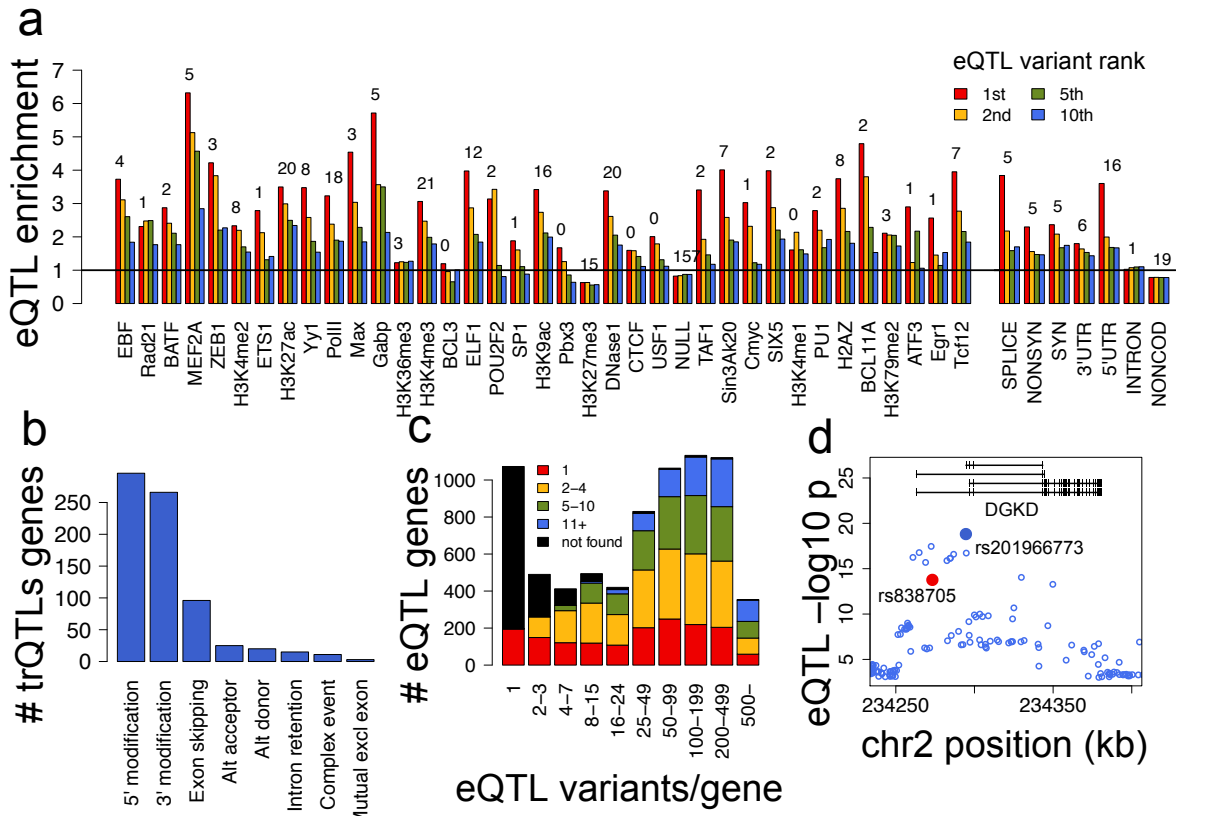
Figure 2. Transcriptome QTLs. Functional annotation of EUR eQTLs (a), with an enrichment of eQTLs in regulatory and coding annotations for the 1st, 2nd, 5th and 10th best associating eQTL variant per gene, relative to a matched null set of variants. The vertical line at 1 denotes the null, and the numbers are –log10 p-values of a Fisher test between the best eQTL and the null set for each category. b) Classification of changes caused by transcript ratio QTLs. c) The rank of the best Omni2.5M SNP among the significant EUR eQTL variants per gene, in bins on the x-axis according to the total number of significant variants. d) DGKD gene locus where an intronic SNP rs838705 is associated to calcium levels[29] (red), and the top eQTL variant 21 kb downstream (blue) is a very likely causal variant, being located close the start site of two transcripts and in the middle of regulatory regions for H3K27ac, MEF2A, MEF2C, and DNase1.

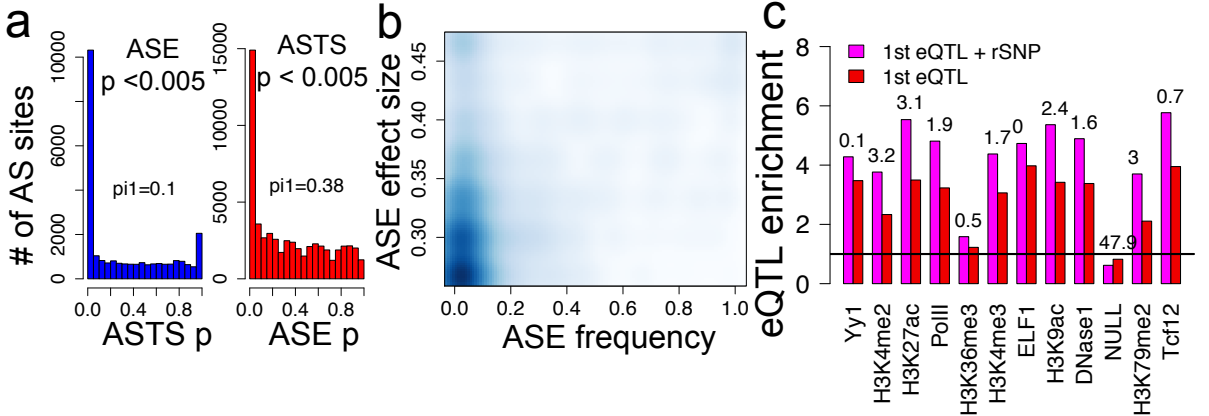
Figure 3: Allele-specific effects on expression and transcript structure. a) Sharing of allele-specific expression (ASE) and transcript structure (ASTS) signals: For significant (p<0.005) ASE sites, the distribution of ASTS p-value of

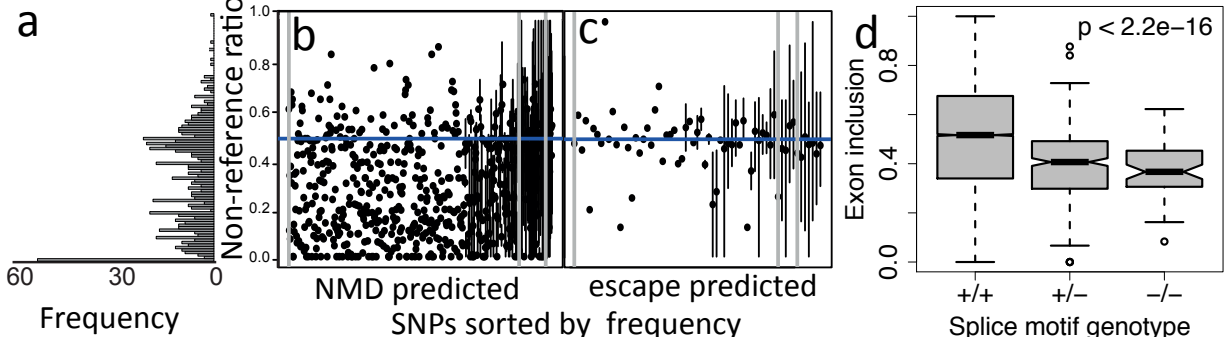
the sites in the same individual is plotted, and vice versa for significant ASTS sites. The ASE p-values are calculated from sites sampled to exactly 30 reads in order to avoid inflation of sharing due to shared high coverage. The numbers denote the pi1 statistic measuring the enrichment of low p-values. b) Frequency of significant ASE event in the population (x-axis) and the effect size (abs(0.5 – REF/TOTAL)) of the significant ASE events, calculated per ASE SNP. Only ASE SNPs with >=20 heterozygote individuals with >=30 reads were included, and the data was corrected for coverage biases and false positives by sampling and permutations (see Supplementary Methods). 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 Fig. S33 for all categories, see also Fig. 2a).

Figure 4: Transcriptome effects of loss-of-function variants. Nonsense-mediated decay due to premature stop codon variants was measured using allele-specific expression. a) shows the distribution of non-reference allele ratios (on the y-axis) for premature stop variants, and the scatterplots break this distribution to individual variants sorted on the x-axis according to derived allele frequency, with sites classified to those predicted to trigger (b) and escape (c) 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. d) Exon inclusion scores for variable exons for individuals that carry 0, 1 or 2 copies of variants that destroy a splice motif.


a
Sample correlation
1.00
0.95
0.90
0.85
ND
mRNA
miRNA
INDV: SAME SAME DIFF DIFF
LAB : SAME DIFF SAME DIFF
b
Genes (density)
1.5
1.0
0.5
0.0
CEU
FIN
GBR
TSI
YRI
0.0 0.2 0.4 0.6 0.8 1.0
Expression / total variation
c
transcript ratio – both – expression
GBR.TSI 263
FIN.TSI 524
FIN.GBR 880
CEU.TSI 1379
CEU.GBR 2103
CEU.FIN 1744
FIN.YRI 3130
GBR.YRI 3212
TSI.YRI 2611
CEU.YRI 4379
0 20 40 60 80
% of genes
d
miR-155
miR-181
miR-150
miR-30


a
eQTL enrichment
eQTL variant rank
1st
2nd
5th
10th
EBF
Rad21
BATF
MEF2A
ZEB1
H3K4me2
ETS1
H3K27ac
Yy1
PolII
Max
Gabp
H3K36me3
H3K4me3
BCL3
ELF1
POU2F2
SP1
H3K9ac
Pbx3
H3K27me3
DNase1
CTCF
USF1
NULL
TAF1
Sin3Ak20
Cmyc
SIX5
H3K4me1
PU1
H2AZ
BCL11A
H3K79me2
ATF3
Egr1
Tcf12
SPLICE
NONSYN
SYN
3'UTR
5'UTR
INTRON
NONCOD
b
trQTLs genes
5' modification
3' modification
Exon skipping
Alt acceptor
Alt donor
Intron retention
Complex event
Mutual excl exon
c
eQTL genes
1
2–4
5–10
11+
not found
eQTL variants/gene
d
eQTL –log10 p
DGKD
rs201966773
rs838705
chr2 position (kb)


a
# of AS sites
ASE
p <0.005
pi1=0.1
ASTS p
ASTS
p < 0.005
pi1=0.38
ASE p
b
ASE effect size
ASE frequency
c
eQTL enrichment
1st eQTL + rSNP
1st eQTL
0.1
3.2
3.1
1.9
0.5
1.7
0
2.4
1.6
47.9
3
0.7
Yy1
H3K4me2
H3K27ac
PolII
H3K36me3
H3K4me3
ELF1
H3K9ac
DNase1
NULL
H3K79me2
Tcf12


a
b
c
d
Non-reference ratio
1.0
0.8
0.6
0.4
0.2
0.0
60
30
0
Frequency
NMD predicted
escape predicted
SNPs sorted by frequency
p < 2.2e−16
Exon inclusion
0.0
0.4
0.8
+/+
+/−
−/−
Splice motif genotype