**Transcriptome and genome sequencing uncovers functional variation in human populations**

Tuuli Lappalainen1,2,3#, Michael Sammeth4,5\*, Marc R Friedländer5,6\*, Peter AC ‘t Hoen7\*, Jean Monlong5\*, Manuel A Rivas8\*, Mar Gonzàlez-Porta9, Natalja Kurbatova9, Thasso Griebel4, Pedro G Ferreira5,6, Matthias Barann10, Thomas Wieland11, Liliana Greger9, Maarten van Iterson7, Jonas Almlöf12, Paolo Ribeca4, Irina Pulyakhina7, Daniela Esser10, Thomas Giger1, Andrew Tikhonov9, Marc Sultan13, Gabrielle Bertier5,6, Daniel G MacArthur14,15, Monkol Lek14,15, Esther Lizano5,6, Henk PJ Buermans7,16, Ismael Padioleau1,2,3, Thomas Schwarzmayr11, Olof Karlberg12, Halit Ongen1,2,3, Helena Kilpinen1,2,3, Sergi Beltran4, Marta Gut4, Katja Kahlem4, Vyacheslav Amstislavskiy13, Matti Pirinen8, Stephen B Montgomery1†, Peter Donnelly8, Mark I McCarthy8,17, Paul Flicek9, Tim M Strom11,18, The Geuvadis Consortium, Hans Lehrach13,19, Stefan Schreiber10, Ralf Sudbrak13,19, Ángel Carracedo20, Stylianos E Antonarakis1,2, Robert Häsler10, Ann-Christine Syvänen12, Gert-Jan van Ommen7, Alvis Brazma9, Thomas Meitinger11,18, Philip Rosenstiel10, Roderic Guigó5,6, Ivo G Gut4, Xavier Estivill5,6, Emmanouil T Dermitzakis1,2,3#

\* 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. Interpretation of their functional effects is still a challenge, but essential for understanding the genetic basis of variation in human traits. Towards this goal, we sequenced mRNA and small RNA 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 mRNA and miRNA variation in several human populations. Our result show that regulatory variants are extremely widespread and affect expression and transcript structure of most genes. By integration of functional annotation we inferred putative causal regulatory variants allowing us to also predict causal variants for dozens of disease-associated loci. 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 understanding and predicting the molecular genomic interactions of causal variants in the human genome and their cellular effects.

**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.{Abecasis, 2012 #88} While many studies have successfully linked genetic loci to various human phenotypes, e.g. by genome-wide association studies,{, 2007 #89} 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{Dunham, 2012 #168}. One approach to address this challenge has been to analyze cellular phenotypes, such as gene expression, resulting in large catalogs of regulatory variants{Emilsson, 2008 #169;Stranger, 2007 #11;Liang, 2007 #90;Grundberg, 2012 #113} known to affect many human diseases and traits.{Nica, 2010 #7;Nicolae, 2010 #91} 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.{Montgomery, 2010 #8;Pickrell, 2010 #92}

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{Abecasis, 2012 #88} 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{Abecasis, 2012 #88} 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, and transcriptome sequencing was performed in seven European laboratories on 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{Marco-Sola, 2012 #96} and miraligner{Pantano, 2010 #95} 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 with good-quality mRNA and miRNA sequencing data, respectively: 91 & 87 (mRNA & miRNA) in CEU, 95 & 93 in FIN, 94 & 94 in GBR, 93 & 89 in TSI, and 89 & 89 in YRI (Table S1), with CEU, FIN, GBR and TSI pooled as EUR (373 & 363) in many analyses. The samples had uniform clustering and distributions of gene and read counts 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 × 10-16 for mRNA, p = 1.34 × 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 a 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 normalized quantifications. Permutations were used to adjust FDR to 5% (Supplementary Methods).

**Transcriptome variation in human populations**

Individual differences in transcript expression can manifest in overall gene 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{Gonzalez-Porta, 2012 #100} shows that variation in alternative transcript usage as the dominating factor (Fig 1b, S10, S11a). The proportion of variation explained by expression levels differs markedly among genes but in a manner that is highly consistent 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 (Mann-Whitney test compared to random grouping p < 2.2 × 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 {Anders, 2012 #186;Barbosa-Morais, 2012 #189}.

We quantify a total of 644 autosomal miRNAs in >50% individuals, with the 29 highest expressed accounting for 90% of the total miRNA reads (SFig 12). 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{Parts, 2012 #192}. Thus far, miRNA function has been mostly studied in cell pertubation experiments (reviewed in {Huntzinger, 2011 #193}), 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{Xiao, 2009 #195}. Interestingly, 45% of the significant associations were positive – consistently to previous results {Parts, 2012 #198} – even though miRNAs are believed to mostly downregulate genes. To understand the direction of causality, we further 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 small 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 × 10-7 for negative targets and 26% p=4.0 × 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 feed-back circuit, stabilizing the expression of both the miRNA and the transcription factor {Wu, 2009 #196}. Observing several such circuits supports the idea that a function of miRNAs under steady-state conditions is to confer robustness to expression programs{Ebert, 2012 #197}. {Murata, 2010 #166}

**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 in the text, 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 with another independent eQTL, while as much as 34% of the genes have additional eQTLs 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 low 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 obtained for overall gene expression 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, χ2 p < 2.2 × 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 EUR trQTLs involve splicing changes only in a minority of cases, with 15% of the transcripts differing in exon skipping and as much as 48% and 43% varying in 5’ and 3’ ends, respectively (note that one trQTL gene can belong to several categories; Fig 2b). This implies that predicting the cellular effects of these variants that rarely change protein structure will be challenging, but likely highly relevant for post-transcriptional regulation. Altogether, these results demonstrate substantial regulatory complexity within individual genes, as well as ubiquitous regulatory variation 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.

For detection of cis-eQTLs in unannotated transcript elements, we quantified expressed retrotransposon-derived elements outside genes that are known to be an important source for evolution of new exons and transcripts.{Cordaux, 2009 #108} We detected a total of 5,763 elements outside annotated genes with significant cis-eQTLs, with an overrepresentation of long terminal repeat retrotransposons (Fig S16). We detected widespread sharing of eQTLs for exons and repeats: 49% and 6% of the best repeat eQTLs variants in EUR were also significant and the best exon eQTLs, respectively (3.8× and 26× enrichment; Fisher exact test p<2.2 × 10-16 for both) compared to set of null variants matched to repeat eQTL frequency and distance from genes. Furthermore, the repeats with eQTLs had significantly higher coexpression with exons (Fig S16; KS p<2.2 × 10-16). This suggests that repeat elements can contribute to transcript structure variation by providing alternative 3’- and 5’-UTRs, and that regulatory elements of nearby genes and genetic variants in them often affect repeat element expression too.

**Characterization of regulatory variants**

Genome sequencing data gives us a unique opportunity to characterize causal regulatory variants and their properties. In these analyses, we focused on our large data set of exon eQTLs from Europeans and describe results from these unless otherwise mentioned; results from Yoruba and from trQTLs are shown in Supplementary Figures. To characterize the properties of eQTLs and estimate our ability to find causal variants, we compared the most significant eQTL variant per gene – the most likely causal variant – to the 2nd, 5th and 10th most significant variants, and to a null distribution of non-eQTL variants matched for distance from transcription start site and minor allele frequency (Fig. 2a; Supplementary Methods).

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 × 10-3; Fig S17), suggesting that indels are more likely to have functional effects than SNPs. eQTLs are highly enriched in many 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 × 10-9), DNase1 hypersensitive sites (3.4×, p = 1.00 × 10-20), and a high enrichment in chromatin states of active promoters (3.5x, p = 1.08 × 10-36) and strong enhancers (median 2.4×, median p = 1.14 × 10-5) (Fig 2a, S18). Within genes, the most significant eQTL variants were highly enriched in 5’UTRs (3.6x, p = 3.23 × 10-17) consistently with the neighboring promoter activity, while splice-site (3.8×, p = 1.65 × 10-5) and nonsynonymous (2.3×, p = 4.84 × 10-6) enrichments point to coding variants having also regulatory functions. Transcript ratio QTLs are overrepresented in splice sites (6.8×, p = 2.44 × 10-7), as expected, but also for example in 3’UTRs (2.5×, p = 1.83 × 10-6) and H3K9ac sites (2.9×, p = 6.57 × 10-6) and promoters (2.4×, p = 5.79 × 10-6) (Fig. S19). Altogether, these analyses add to previous understanding{Degner, 2012 #172;Gaffney, 2012 #182} of the functional consequences of genetic variants to transcription.

These analyses show a characteristic pattern of highest enrichment for the most significant eQTL variant, and a rapidly decreasing trend towards lower ranks. With this information, we were able to estimate how likely the most significant variant is also the causal 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 the best variant is the causal variant in 55% of EUR eQTLs and 74% of YRI eQTLs, with more conservative estimates being 34% and 41%, respectively (Fig S20). The higher percentage in YRI is expected due to lower LD confounding; altogether these results indicate reasonable power to find causal regulatory variants. In order to validate the putative causal effects of a subset of our best eQTL variants from both EUR and YRI, we analyzed allele-specific binding in a published dataset of CTCF CHiP-seq data from 6 individuals{McDaniell, 2010 #190}, and observe that our best eQTL variants overlapping CTCF binding peaks show significantly more allele-specific binding compared to our matched null variants (p = 2.0 × 10-3, Fig S21). Furthermore, our best eQTLs showed a high degree of overlap with DNase1 sensitivity QTLs{Degner, 2012 #172} (3.3×, p = 2.51 × 10-6 in EUR, 7.9×, p < 2.2 × 10-16 in YRI). Not relying on SNP array data{Veyrieras, 2008 #184} 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 have significant enrichment of GWAS{Hindorff, 2010 #31} variants (16% of 6,473 GWAS variants are eQTLs in EUR or YRI, versus 11% of a frequency-matched GWAS null distribution; chi2 p < 2.2 × 10-16; for trQTLs the percentages are 1.8% and 0.84%, respectively, p = 7.2 × 10-9). The large proportion of overlap even under the null warrants caution against inference of causality from a mere overlap of QTL-GWAS signals; however, the significant enrichment of GWAS SNPs in the top eQTL ranks (p=1.18 × 10-7; Fig. S23) suggests that eQTL effects indeed have increased probability to be causal GWAS mechanisms, as observed before{Nicolae, 2010 #91;Nica, 2010 #7}. We can now apply our ability to predict causal eQTL variants to find causal disease-associating variants for these loci. To this end, we selected 78 strictly filtered eQTL regions that have been statistically estimated to underlie 92 GWAS SNPs using the RTC method,{Nica, 2010 #7;Grundberg, 2012 #113} and of these, the best eQTL variant in our data – i.e. the most likely causal variant – is not the GWAS SNP in all but two cases (Fig S24). Altogether, we pinpoint novel putative causal variants in 78 GWAS loci, listed in Table S5. Figure 2d shows an example of the DGKD gene locus where an intronic SNP rs838705 is associated to calcium levels{O'Seaghdha, 2010 #165}, 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 *de novo* approach that captures differences in the exonic distribution of reads from the two haplotypes (Fig S2, S25, 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 S31d) indicates 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 show a clear population structure (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 measure concordance between allelic ratios over an aseSNP, and the genotypes of putative rSNPs (prSNPs) in the cis-regulatory region (100kb window from TSS): a heterozygote for a true rSNP is expected to have large deviation of the null allelic ratio of 0.5, whereas a homozygote should have a ratio close to 0.5 (see Supplementary Methods, Fig. S2, S31). Based on this principle, we calculated a concordance score for aseSNPs–prSNP pairs after applying stringent filters for ASE data to maximize power (read coverage sampled to 30 in >=80 individuals of which >1 with significant ASE), and for each 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. S32a), and assigning the prSNPs with empirical p<0.01 to p<0.001 as likely rSNPs gives 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{Stranger, 2007 #11} (Fig. S32c). eQTL variants are highly enriched in low prSNP p-values, thus independently replicating the majority of tested eQTL signals (pi1 = 0.55; Fig 3c). 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). Together, these results suggest that the vast majority of allele-specific expression is likely to be genetically driven, and that ASE-based approaches can capture also additional effects not seen in eQTL analysis. ASE analysis, with its measurement of regulatory effects within an individual rather than quantitatively in a population, may provide a powerful approach for 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.{MacArthur, 2012 #127} 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. Most of these variants are rare and seen in only one or two individuals. First, we analyzed premature stop variants that may induce nonsense-mediated decay.{Nagy, 1998 #131} As expected, we observe a high enrichment for ASE with loss of the variant allele (Fig, 4a, S33) as in previous studies{MacArthur, 2012 #119;Montgomery, 2011 #52}. Variants close to the end of the transcript are predicted to escape NMD{Nagy, 1998 #131}, which is generally the case in our data. 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 × 10-16, Fig. S34), 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 × 10-16, Fig 4d), indicating that the variants indeed lead to exon skipping.

**Conclusions**

TThis study illustrates the power of combining genome sequence analysis with a high-depth functional readout such the transcriptome. 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. Both gene expression and transcript structure variation contribute to intra- and inter-population differences, with alternative transcript processing dominating population differences and allele specific expression. We often find multiple independent regulatory variants in a gene, demonstrating both independent and shared effects on total expression levels and alternative transcript processing. 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. This will accelerate the development of methods for functional interpretation of personal genomes.

**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{Marco-Sola, 2012 #96} to hg19, Gencode v12{Harrow, 2012 #128} exons were quantified from read overlaps, and FluxCapacitor{Montgomery, 2010 #102} was used for transcript quantification. Small RNA data was mapped and quantified with miraligner{Pantano, 2010 #95} using miRBase v18{Kozomara, 2011 #132}. 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{Stegle, 2010 #134} for the QTL and miRNA-mRNA correlation analyses.

**References**

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://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. S35). For visualisation of the results we created the Geuvadis Data Browser ([www.ebi.ac.uk/Tools/geuvadis-das](http://www.ebi.ac.uk/Tools/geuvadis-das)) where quantifications and QTLs can be viewed, searched, and downloaded (Fig. S36). The project webpage [www.geuvadis.org](http://www.geuvadis.org) provides full documentation and links to all files, and the project wiki is open to the public in geuvadiswiki.crg.es.

Reprints and permissions information is available at www.nature.com/reprints

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{O'Seaghdha, 2010 #165} (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) The prSNP p-value indicating the concordance between genotypes and allele-specific expression for all the best eQTL variants, with the strong enrichment of low p-values demonstrating the power of ASE-based mapping of regulatory variants.

**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.