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

**Supplementary Methods**

Study design 2

RNA-sequencing data production 2

Cell line processing 2

RNA extraction 2

RNA sequencing 3

Raw data processing 3

Genotype data 3

Variant annotation 4

Imputation 4

Quality control 4

mRNA read mapping 4

mRNA quantifications 5

Exons and genes 5

Transcripts, splice junctions, and introns 5

Exon inclusion 6

Links between exons 6

Transcribed repeats 6

small RNA (sRNA) data processing 7

Improved miRNA gene annotations 7

sRNA read data processing 7

sRNA mapping and quantification 7

RNA-seq quality control 8

Outlier detection 8

Sample swap and contamination analysis 8

miRNA data quality control 8

Normalization of quantifications 9

Quantitative versus qualitative variation 9

Differentially transcribed genes 10

Differential expression 10

Differential isoform usage 10

Fusion genes 10

RNA editing 11

miRNA effects on the transcriptome 11

miRNA family and target definition 11

Integrated analysis of mina and mRNA expression 12

Trans-eQTL effects of cis-mirQTLs 13

Transcriptome QTL analysis 13

Transcriptome QTL mapping with linear regression 13

Isoform QTL mapping with variance test 14

Independence of QTLs 14

Null variant distribution 14

Allele-specific analysis 14

Allele-specific expression (ASE) analysis 14

Allele-specific transcript structure (ASTS) analysis 15

Loss-of-function analysis 15

Nonsense-mediated decay 15

Splice scores 15

# Study design

Transcriptome sequencing was performed in seven European laboratories, each processing 48-113 randomly assigned samples (Fig. S1b). 5 samples were sequenced in replicate in each of the labs for both mRNA and miRNA, and twice in UNIGE for mRNA. Additionally, in University of Geneva, 168 samples sequenced in other laboratories were mRNA-sequenced at 2/3 of the standard coverage.

# RNA-sequencing data production

## Cell line processing

EVB transformed lymphoblastoid cell lines (LCLs) from Coriell Cell Repositories (GBR, FIN, TSI) and University of Geneva (CEU, YRI) were shipped to ECACC (European COllection of Cell Cultures) as live cultures, in batches of ~ 30 samples from Coriell (GBR/FIN/TSI more or less randomized) and 2 x ~90 samples (by population) from Geneva.

In ECACC, these cell lines were cultures to approximately 1.2 x 10e8 cells. These cultures were split to produce 8 x cell banks of the samples, and a snap frozen pellet of 2 x 10e7 cells from a proliferating culture. The cell pellets were shipped from ECACC to University of Geneva in three batches, the first batch consisting of CEU/GBR/FIN/TSI samples, and the second and third batch with YRI and some CEU samples.

Total RNA was extracted from the cell pellets in Geneva (see below), first extracting about 2/3 of the first batch with full randomization, then adding the second batch and randomizing among that and the remaining 1/3 of the first batch, and finally extracting the third batch.

## RNA extraction

Total RNA was extracted from cell pellets using the TRIzol Reagent (Ambion). The pellets had been frozen at ECACC without any additives like RNAlater or Trizol. In Geneva they were thawed, 1mL of trizol was added in each sample, and the samples were transferred to eppendorf tubes. The rest of the protocol followed the manufacturer's guidelines. No DNAse treatment was done to the RNA samples.

RNA quality was assessed by Agilent Bioanalyzer RNA 6000 Nano Kit according to the manufacturer's instructions. The sample information includes a visual inspection of RNA quality (degradation), in the scale 0, 0.5, 1.

RNA quantity was measured by Qubit 2.0 (Invitrogen) using the RNA Broad range kit according to the manufacturer's instructions.

## RNA sequencing

Each of the sequencing laboratories were sent a minimum of 4 ug of total RNA of the samples allocated to them, and total RNA Bioanalyzer was ran for 10-20% of the RNA samples before library preparation to confirm sample quality after shipping. No further purification steps were done to the RNA samples other than that specified in the sequencing protocols. Library preps were done in random order in every laboratory.

mRNA sequencing was done on the Illumina HiSeq platform with 75 bp paired-end sequencing with fragment size of 280 bp. TruSeq RNA Sample Prep Kit v2 (the high-throughput protocol) was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation, and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired coverage of minimum 10M mapped and properly paired read pairs from any standard mapper, without filtering for mapping quality.

Small RNA sequencing was done on the Illumina HiSeq platform with 36 bp single-end sequencing with fragment size of 145-160 bp. TruSeq Sm RNA Sample Prep kit was used for library preparation, TruSeq PE Cluster Kit v3 for cluster generation, and TruSeq SBS Kit v3 for sequencing. The laboratories were allowed to choose freely how to pool the samples to get the desired coverage of approximately 3M total reads.

Extensive information of sample processing was collected from all the laboratories for both mRNA and miRNAseq in order to enable control of batch effects.

## Raw data processing

Each lab submitted one demultiplexed fastq file per sample per mRNAper and miRNAseq, produced by CASAVA 1.8 or 1.8.2 allowing one mismatch in the index. Reads failing Illumina quality filtering were removed. The fastq files are named as: SAMPLE\_ID.SeqLabNumber.M/MI\_YYMMDD\_Lane\_Read.fastq.gz, where M/MI stands for mRNA or miRNA sequencing, and YYMMDD is the sequencing date. All the data were submitted and initially stored in the project ftp site.

General tools used in data processing in this project were Samtools (REF) and R (REF).

# Genotype data

Since not all 1000 Genomes variants have rs-identifiers, we renamed all the variants as follows: SNPs had an identifier of type snp\_chr\_pos e.g. (snp\_21\_357682), and indels and structural variants were of type indel/sv:lengthI/D\_chr\_startpos (indel:3D\_1\_10523).

## Variant annotation

The Variant Effect Predictor (VEP v2.5; http://useast.ensembl.org/info/docs/variation/vep/vep\_script.html) tool from Ensembl was modified to produce custom annotation tags (STable) and additional loss of function (LoF) annotations. The additional LoF annotation was applied to variants that were annotated as STOP\_GAINED, SPLICE\_DONOR\_VARIANT, SPLICE\_ACCEPTOR\_VARIANT, and FRAME\_SHIFT (Figure 1) and flagged if any filters failed (Table 2). A LoF variant is predicted as high confidence (HC) if there is at least one transcript that passes all filters, otherwise it is predicted as low confidence (LC). The custom annotation tag is a comma-separated ordered list of features corresponding to each of the transcripts that overlap the variant. This modified version of VEP was applied to the 1KG phase1 data. Tuuli: additions to the annotation

## Imputation

From Natalja

Since IMPUTE2 did not handle multiallelic genotypes well, we kept only biallelic genotypes in the analysis. Furthermore, xx variants with >30% frequency difference between Phase1 and Phase2 data were removed. Additionally, the genotype calls of imputed genotypes with posterior probability <0.9 were marked as missing, keeping the genotype dosage.

## Quality control

PCA (REF) of genotype data from chr20 showed clear clustering to Phase1 and Phase2 individuals (SFig). Additionally, within Europe we see a clear clustering to populations. To make sure that our findings are not driven by biases from imputation or from population structure, we included the imputation status (0|1) and principal components 1-3 for Europeans and 1-2 for Yoruba as covariates in QTL analyses.

In QTL analysis, we used variants with >5% MAF in either EUR or YRI, which gives us 10,785,347 variants in total, of which 9,836,718 are SNPs, 945,987 are indels, and 2642 are SVs. QTL analysis was done with genotype dosage values.

# mRNA read mapping

We employed the JIP pipeline (Griebel & Sammeth submitted) to map RNA-Seq reads and to quantify mRNA transcripts. For alignment to the human reference genome sequence (GRCh37, autosomes + X + Y + M), we used the GEM mapping suite (v1.349 which corresponds to publicly available pre-release 2) (*REF*) to first map (max. mismatches = 4%, max. edit distance = 20%, min. decoded strata = 2 and strata after best = 1) and subsequently to split-map (max.mismatches = 4%, Gencode v12 and *de novo* junctions) all reads that did not map entirely. Both mapping steps are repeated for reads trimmed 20 nucleotides from their 3’-end, and then for reads trimmed 5 nucleotides from their 5’-end additionally to earlier 3’-trimming—each time considering exclusively reads that have not been mapped in earlier iterations. Finally, all read mappings were assessed with respect to the mate pair information: valid mapping pairs are formed up to a max. insert size of 100,000 bp, extension trigger = 0.999 and minimum decoded strata = 1. The mapping pipeline and settings is described below, and can also be found in http://github.com/gemtools, where the code as well as an example pipeline are hosted.

The GEM output format was converted to sam/bam format, with following mapping quality scores and flags:

1. Matches which are unique, and do not have any subdominant match: 251 >= MAPQ >= 255, XT=U
2. Matches which are unique, and have subdominant matches but a different score: 175 >= MAPQ >= 181, XT=U
3. Matches which are putatively unique (not unique, but distinguishable by score): 119 >= MAPQ >= 127, XT=U
4. Matches which are a perfect tie: 78 >= MAPQ >= 90, XT=R.

Furthermore, the NM flag contains the number of total mismatches (read1+read2). In most analyses, we used reads in categories 1 and 2 and with NM<=6.

The analysis of fusion genes was based on read mapping with bwa-0.5.9 (REF) with default parameters

# mRNA quantifications

The gene annotation used in this project was Gencode v12 (REF).

## Exons and genes

Exon quantifications were calculated for protein-coding and linc-RNA transcripts. All overlapping exons of a gene were into meta-exons with identifier of type ENSG000001.1\_exonstartcoord\_exonendcoord. Read counts over these elements were calculated without using information of read pairing, except that we exclude reads where the pairs map to two different genes. We count a read in an exon if either it's start or end coordinate overlaps an exon. For split reads, we count the exon overlap of each split fragment, and add counts per read as 1/(number of overlapping exons per gene). Gene counts are calculated as the total number of reads overlapping exons of each gene.

## Transcripts, splice junctions, and introns

Quantifications of transcripts, introns and splice-junctions by the Flux Capacitor approach (*REF*) are based on the annotation-mapped genomic mappings considering transcript structures of the Gencode transcriptome annotation: mappings of read pairs that were completely included within the annotated exon boundaries and paired in the expected orientation have been taken into account. For introns, we used all-intronic regions that are not retained in any mature annotated transcript, and reported mapped reads in different bins across the intron in order to distinguish reads stemming from retained introns from those produced by not yet annotated exons. Reads belonging to single transcripts were predicted by deconvolution according to observations of paired reads mapping across all exonic segments of a locus.

Annotated splice junctions were quantified using split read information, counting the number of reads supporting a given junction. Novel splice sites were quantified by split-reads that are overlapping a window of +/-30 bp around known exon boundaries. Each different pattern of split-reads was regarded as a potential splice-isoform, if they had insert-size and split-distance of at most 10,000 bp.

In addition to read counts in genes (see Exon quantifications), we calculated gene RPKMs as the sum of all transcript RPKMs per gene.

## Exon inclusion

Exon inclusion levels were calculated as the Percentage Splice Index (PSI) [PMID:21876675,18978772], defined as the ratio between inclusion reads and inclusion reads plus exclusion reads.

## Links between exons

For quantification of splicing variation, we used the Altrans method (Ongen & Dermitzakis submitted). The method utilizes the paired end nature of the RNA-seq experiment as well as split reads by quantifying links where two exons are joined either through a split read or by the mate pairs. The first exon in a link is referred to as the “primary exon”. Overlapping exons are grouped into “exon groups” and unique portions of each exon in an exon group are identified. These unique portions are subsequently used to assign reads to an exon. These raw links counts were normalized as described in the “Normalization of quantifications” section. The normalized link counts ascertained from unique regions of exons, which can be derived from parts of the linked exons rather than the whole exons, are divided by the probability of observing such a link given the empirically determined insert size distribution for each sample and unique portions of the exons in question, which is referred to as “link coverage”. Finally, the quantitative metric that can be used in QTL analysis is the fraction of one link’s coverage over the sum of the coverages of all the links that the primary exon makes. We calculated this metric in 5’-to-3’ (forward) and 3’-to-5’ (reverse) directions to capture splice acceptor and donor effects respectively. In the association analyses, we only included links where the primary exon’s exon group made at least 10 links in the analyzed direction in at least 80% of the individuals and where the primary exon made at least 5 links in the analyzed direction in at least 30% of the individuals.

## Transcribed repeats

We quantified transcription on repeat elements using the following approach. First, we extracted all repetitive elements from UCSC's repeat masker table, and excluded all elements that overlapped UCSC or Gencode genes by at least one nucleotide. This left us with 2.5M regions, in which we have then counted the number of overlapping RNA-seq reads in each region for each sample. Reads that were only partially overlapping are only counted for the part which is overlapping. Since we observed that rRNA elements had strong differences between laboratories, we excluded them from further analysis.

# small RNA (sRNA) data processing

## Improved miRNA gene annotations

Our annotation builds on miRBase version 18 (PMID 21037258) but with important improvements. In the cases where only one miRNA strand was annotated, the position and sequence of the other strand was estimated using RNA structure prediction. Further, for the mature and hairpin miRNAs which overlapped SNP or indel variants that were polymorphic in our genotype data, sequences carrying the nonreference alleles were generated and used for downstream analyses. This is important for avoiding allelic mapping bias especially for short sequences such as miRNAs.

## sRNA read data processing

Since some laboratories sequenced 50 bp reads of sRNA libraries, these datasets were trimmed using the FASTX suite from the Greg Hannon lab (http://hannonlab.cshl.edu/fastx\_toolkit/), using the following command: fastx\_trimmer -f 1 -l 36 -Q33. Homo-polymer reads were removed with this command: fastx\_artifacts\_filter -Q33 and reads with low PHRED scores removed with this line: fastq\_quality\_filter -q 10 -p 50 -Q33. Ligation adapters were clipped using the AdRec.jar program from the seqBuster suite (PMID 20008100) with the following options: java -jar AdRec.jar 1 8 0.3. A custom search subsequently clipped shorter adapters: if there were no matches to the first 8 nts, then matches to the first 7 nts of the adapter were searched in the last 7 nts of the read, then matches of the first 6 to the last 6 positions and so on. Reads that had no matches were retained, but not clipped. Last, reads shorter than 18 nts were discarded. The remaining reads were each mapped against several databases, depending on the purpose of the mapping.

## sRNA mapping and quantification

For tracing the reads to their genomic source for quality control purposes, reads were mapped to the hg19 genome concatenated with unassembled parts of the human genome and genomes of known human viral pathogens (available upon demand) with this command line: bowtie -f -v 1 -a --best --strata. sRNA reads were assigned to annotations based on the genome mappings. Annotations used are GENCODE 8 (PMID 22955987) supplemented with rRNA and LINE and Alu transposon annotations from RepBase (PMID 16093699) and snoRNA and miRNA annotations from the UCSC table browser (PMID 14681465). Annotations were first resolved so that each nucleotide on each strand had exactly one annotation. In case of nucleotides with more than one annotation, conflicts were resolved using a confidence-based floating hierarchy (as in PMID 18158128). The hierarchy used is: mitochondrion > virus > miRNA > snoRNA > rRNA > tRNA > snRNA > misc\_RNA > lincRNA > processed\_transcript > pseudogenes > protein\_coding > LINE > Alu > intron\_coding > intron\_non\_coding > intergenic. Each read mapping was weighted inversely to the number of genome mappings for the read, e.g. a read mapping to two genomic locations would get an assigned weight of 0.5. Each mapping was counted towards the annotation of the nucleotide in the middle of the mapping.

miRNA quantifications for analysis were calculated as read counts using miraligner.jar from the seqBuster suite using the following options: java -jar miraligner.jar 1 3 1, and using the improved annotations as the reference. Reads which map equally well to two or more miRNAs are counted fully towards each miRNA.

# RNA-seq quality control

A more detailed analysis of quality control of this dataset can be found in t’ Hoen et al. (submitted).

## Outlier detection

The read and gene count distribution of mRNA-seq data were very uniform (Fig. Sx, Table Sx). To further estimate sample quality, we calculated Spearman rank correlation between all samples using exon counts and transcript RPKMs. From these data, we calculated the so-called D-statistic for each sample – the median correlation of one sample against all the other samples. 2 samples in mRNA data and 4 samples in miRNA data were excluded from analysis due to low correlation with other samples.

We also used multidimensional scaling to visualize the sample correlation matrix. The samples clustered relatively uniformly but with some separation by the sequencing lab (Fig. Sx), but this effect was completely removed by normalization (see below).

## Sample swap and contamination analysis

Allele-specific expression analysis of mRNA-seq data was used to detect sample swaps, which we did not find. ASE analysis as well as analysis of sex chromosome specific genes were used to find potential sample contamination in RNA-sequencing data, and 5 samples were excluded because of possible contamination.

## miRNA data quality control

The total small RNA read count and the number of miRNA reads were relatively similar across samples, but the proportion of miRNA reads per sample showed large variation from close to 0 to 60% (Fig S). This is likely caused by variation in the library preparation step, and sequencing of a large number of non-miRNA reads in some samples. However, the number of quantified miRNAs is very uniform, and is not correlated to the proportion of miRNA reads (Fig S), and only 8 samples were excluded due having low mapping rate, coverage, or gene count. This indicates that while in some samples sequencing depth is lost on non-miRNA reads, this hardly affects our miRNA detection and quantification. Notably, correlations between miRNA samples were high, and population clustering clearly more pronounced than clustering by laboratory even before normalization (Fig S).

# Normalization of quantifications

All read count quantifications were corrected for variation in sequencing depth between samples by normalizing the reads to the median number of well-mapped reads (45M) for mRNA, and to the median number of miRNA reads (1.2M) for miRNA. In our analyses, we used only elements quantified in >50% of individuals (>90% for QTL analysis).

Additionally, all expression quantifications are affected by technical noise that reduces power. It has been shown in many studies that correcting for such sources of variance improves eQTL discovery dramatically. We normalized quantification data using PEER (REF), which finds synthetic covariates from quantification data that can then be regressed out from the data. These normalizations were done for the total sample set of 462 individuals together.

First, for each type of quantifications, we estimated the best number of covariates (K) to correct. PEER was ran for a subset of the data (chr20, or chr20-22) using k=0,1,3,5,7,10,13,15,20, and sequencing lab and population as additional covariates, the results are transformed to standard normal distribution, and cis-eQTL analysis was performed for each K. The number of genes with an eQTL (p<10e-8 and p<10e-6) was calculated, since finding more eQTL genes is a good indicator of power to find biological effects. These results can be seen in Figure Sx.

Based on these results, we chose K=10 as the number of covariates to correct for, except for transcribed repeats where we did not use PEER correction at all. To normalize the final data sets, we ran PEER for 20 000 quantification units (e.g. exons) using sequencing laboratory and population as additional covariates and adding the mean to the model. Covariates from this analysis were regressed out from the quantifications, and the population mean was added to the residuals. Correlation of samples after this normalization showed no clustering according to sequencing laboratory for mRNA data (Fig. Sx, Fig Sx, Fig Sx), and while these effects were not completely removed for miRNA (probably due to smaller number of genes which allows less efficient calculation of synthetic covariates), they are hardly visible in sample clustering. In eQTL analysis and miRNA-mRNA correlation analysis, these values were further transformed to standard normal distribution.

# Quantitative versus qualitative variation

We estimated the contribution of alternative splicing and gene expression on the total transcript abundance variation using approaches in Gonzales-Porta et al. 2012 (REF). Briefly, for each gene, the samples are represented in the RT space using transcript expression levels (T=number of expressed transcripts for this gene), from which we can calculate the total variability (Vt) in this space. Projecting the samples in a model of constant splicing ratios gives us an estimate of expression level variation (Vls). The ratio Vls/Vt estimate the contribution of gene expression in the transcript abundance variability, where Vls/Vt ≈ 1 implies that only gene expression contributes to transcript variability, and Vls/Vt ≈ 0 implies that only isoform variation contributes to transcript variability. In this analysis, we used only protein-coding genes expressed in at least 20 individuals per population with at least two expressed (RPKM >=0.0) isoforms expressed.

We further extended this model to between-population variation. Representing the samples in the space of the transcript expression, between-group variation is computed removing the within-group variation from the total variation. Then all the samples are projected on a line, which represents the model of constant splicing ratios. The between-group variation of these projected points is computed, and the estimator of gene expression level variation between populations is the ratio of between-group variation of the projected points over between- group variation of the original points. A value close to one means that the projection didn’t remove variation, so gene expression is the one mainly contributing to between-population variation.

# Differentially transcribed genes

## Differential expression

We performed gene differential expression (DE) using tweeDEseq (Juan R Gonzalez and Mikel Esnaola (). tweeDEseq: RNA-seq data analysis using the Poisson-Tweedie family of distributions. R package version 1.0.14. http://www.creal.cat/jrgonzalez/software.htm) a method that uses a Poisson-Tweedie family of distributions and is well suited to compare groups with more than 15 samples. After filtering genes with less than 5 counts per million in all samples but one a set of 16 583 genes remained for analysis. We performed pairwise population comparisons and population specific comparisons (one population against the remaining four). Genes with FDR < 0.05 and log2 fold change greater than 3 were considered significant.

## Differential isoform usage

To identify the genes with differences in isoform usage in the populations, for each gene for every pairwise population comparison, the transcript ratios were compared using non-parametric multivariate analysis of the variance (Gonzalez-Porta).

# Fusion genes

We discovered fusion genes from our data set using the following pipeline: Unmapped reads were extracted from all bam files mapped with bwa, converted to fastq files with Hydra (PMID:20308636) and subsequently subjected to fusion genes analysis with FusionMap (PMID: 21593131). Fusion genes that were located on the same chromosome and strand were retained for further analysis. The list of conjoined genes was further filtered by the number of split reads (>2) supporting the fusion.

To ensure that the observed population specific fusions are not due to the lack of expression of the partner genes in the other populations, only genes with positive RPKM values observed in all populations were considered in analysis. The RPKM values were estimated with the RPKM\_count.py script implemented in the RSeQC tool (PMID: 20308636). Finally, the novelty of the fusions was assessed via queries against known annotation databases for read through events such as ConjoinG (PMID:20967262), AceView (PMID:16925834**) published read through events (**PMID: 21261984**)** and literature search.

# RNA editing

RNA editing is a modification of RNA transcripts that might result in alterations of coding or non-coding sequence. Here we assessed population variation of RNA editing events in RNA-seq data by calling variants at 42,039 known editing sites from the DARNED database (Kiran, A. and P. V. Baranov (2010). "DARNED: a DAtabase of RNa EDiting in humans." Bioinformatics **26**(14): 1772-1776). We performed multi-sample variant calling over the 462 Geuvadis samples using SAMtools. Altogether, SAMtools called non-reference variants (i.e. at least one sample had a non-reference “genotype”) at 24,680 sites. To reduce the number of false positive RNA editing events we applied a set of very stringent filters: (1) we required a minimum median coverage of 10 at all called sites; (2) At least 10 samples had to have a non-reference “genotype” at each site; (3) all variants had to pass the SAMtools varFilter script; (4) the variant quality at all sites had to be above 100. Furthermore, to ensure that the observed variants are true RNA editing events and not due to unknown genetic variants, we required two things: (5) there should not be a corresponding variant in the 1000 Genomes Phase 1 data set; (6) all variants had to be located within the set of accessible regions defined by the 1000 Genomes project to ensure that a variant would be present in the genetic variant data if it was present at the DNA level. In the analysis, we used only the 422 of our samples that were part of 1000 Genomes Phase 1 and had full genome information.

# miRNA effects on the transcriptome

## miRNA family and target definition

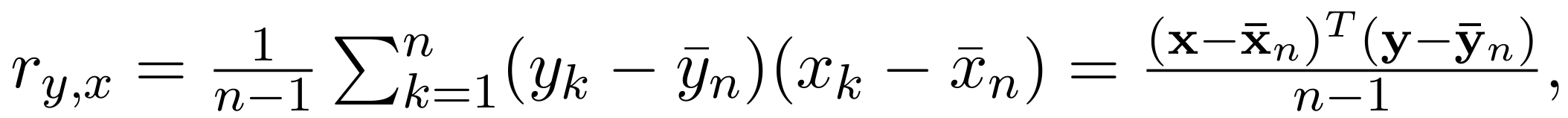
In the analysis of association of miRNA-mRNA quantifications we used 449 samples with both miRNA and mRNA expression data. For defining miRNA-targets we used the TargetScan version 5.2 predictions (PMID 15652477). Specifically, we downloaded the seed families of all known miRNAs conserved in vertebrates or mammals, and the corresponding conserved target sites (http://www.targetscan.org/). The target sites were lifted from REFSEQ annotations by mapping the 3'UTR sequences to the hg19 genome and intersecting the coordinates with our merged exon annotations (see mRNA Quantifications). The validity of the lift was confirmed at the sequence level by matching the seed sites of targets with the reverse complement of the miRNA seeds. For quantifying miRNA seed expression, we summed up read counts for all miRNAs with the TargetScan seed sequences. E.g. the expression of the miR-141/200a seed was found by summing the read counts from hsa-miR-141-3p and hsa-miR-200a-3p.

For mRNA expression data, we used the count data for the exon containing the predicted miRNA binding site. The matched microRNA-mRNA expression data consisted of 449 samples. Both microRNA and mRNA expression data were corrected for hidden confounding factors with PEER (PMID: 2046387) and the resulting residuals transformed to standard normal. The final analysis included 100 microRNA-families and 126,698 exons.

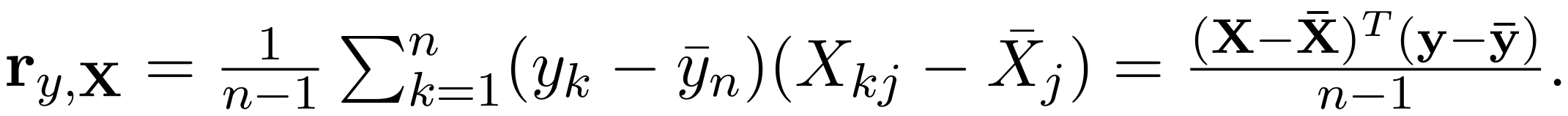
## Integrated analysis of mina and mRNA expression

The integrated analysis is based on the globaltest (PMID: 14693814) and is further described in (Iterson et al., Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions, submitted). The globaltest allows testing of the association of a group of genes -the predicted targets- with a microRNA expression profile. The global test is specifically designed for the situation of more samples then genes (p>>n). Furthermore, the test overcomes the large multiple testing problem that arises when each target is tested individually for association with a microRNA expression profile. P-values for a set of target mRNAs sharing a predicted miRNA seed sequence were obtained by 100,000 permutations of the sample labels and corrected for multiple testing using Holm’s procedure. Within each set of predicted mRNA targets, P-values for individual associations between expression of predicted mRNA targets and miRNA expression levels were corrected by the Bonferroni multiple testing procedure.

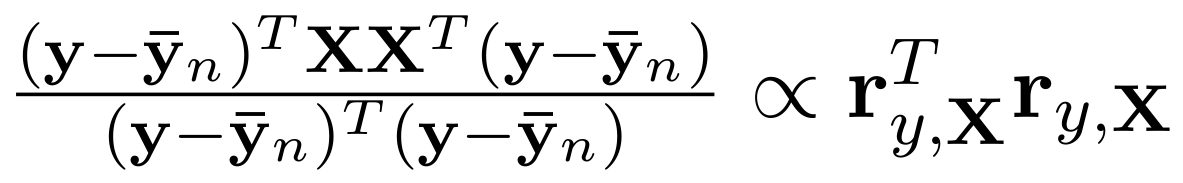
A useful interpretation of the global test is as a sum of squared covariances between a set of predictors Xn×p, and responses, yn×1 (see section 5 of PMID: 14693814). Consider the sample covariance, ry,x between a miRNA expression profile yn×1 and a single target xn×1 given by:



where bar yn and bar xn denote the sample means of miRNA and mRNA expression profiles, bar yn and bar xn are vectorized versions (note that ry,x = rx,y). For multiple mRNA profiles Xn×1 the p×1 vector of the sample covariances, ry,X can be expressed as:



Note that this expression is valid even when the number of targets exceeds the number of samples p > n, and again rTy,X = rX,y. Now the global test test-statistics,



is proportional to the squared sample covariance.

## Trans-eQTL effects of cis-mirQTLs

Variants that associate to miRNA expression levels can potentially be trans-eQTLs for the target genes of these miRNAs. This effect was sought using the European data set. The hypothesis was that a mi-eQTL variant should have a stronger trans-effect on the targets of the miRNA that it affects rather than on genes that are not targets of the miRNA. This analysis is highly dependent on the accuracy of target predictions and can be conservative.

We selected all miRNA-target exon pairs based on TargetScan predictions. From these, we selected only those exons that were included in eQTL analysis (expressed in >90% samples) , and only the 60 miRNAs that had a cis-eQTL. This left us with only 12 miRNA - mRNA exon pairs. For the best-associating variant of each of the 12 mi-eQTLs, we collected trans association p-values (>5MB from the site) with exons that have a target site of the miRNA affected by the eQTL (6392 variant-exon pairs in total, 125-1003 exons per eQTL), and with exons not in genes that have a target site of the miRNA with an eQTL (4842061 variant-exon pairs in total, 81063-82518 exons per eQTL). We compared these p-value distribution, separately for negative and positive associations, i.e. those where the miRNA-cis-eQTL allele increasing the miRNA expression has negative or positive correlation to the exon.

# Transcriptome QTL analysis

## Transcriptome QTL mapping with linear regression

The details of sample sets, data filtering and normalization are discussed above. Briefly, we did transcriptome QTL mapping separately for European (n=373) and Yoruba (n=89) populations. We used genetic variants with MAF>5% in either EUR or YRI <1MB from transcription start site, with covariates of imputation status (0|1), PCs 1-3 for Europeans and PCs 1-2 for Yoruba. For the different quantitative phenotypes, we used normalized quantification units (e.g. exons) with quantification >0 in >90% of all the individuals.

QTLs were mapped using a linear model implemented in Matrix eQTL (REF), and FDR was estimated by permutations as follows: For exon eQTLs, we permuted the quantifications of each exon 2000 times, keeping the best p-value per exon from each round. From these data, we adjusted the FDR to 5% according to the most stringent exon of each gene, having a separate p-value threshold for each gene. For genes, miRNAs and RNA editing sites, we ran 8000 permutations for each quantification unit, and calculated a p-value for each of them. For exon links, we permuted all the links of randomly selected 1000 genes in both forward and reverse direction, selected the most stringent link per gene, and used the median of this distribution as the genome-wide p-value limit. For repeats, we permuted randomly selected 1000 repeats and calculated a genome-wide p-value limit based on the distribution of these.

## Isoform QTL mapping with variance test

We also analyzed QTLs for transcript ratios with a non-parametric multivariate analysis of the variance test, using a factorial model, capable of detecting additive, dominant or recessive effect. Variants <50 kb from gene boundaries and with >=5 individuals in >=2 genotype groups were tested for association with genes with >=2 isoforms. Permutations were done at the gene level and FDR was adjusted to 5%. Only the GBR population was used in this analysis.

For each association, the two transcripts whose ratios change the most were used to characterize the QTL effect. Giving the annotation of these two transcripts, *AStalavista (REF)* was used to classify the event for each sQTL.

## Independence of QTLs

TEXT

## Null variant distribution

To compare QTL variants to a null distribution of similar variants but without regulatory association, we sampled genetic variants in cis-regions of genes expressed in our data set based on the QTL variant distribution of distance from the gene (taking upstream and downstream distance into account) and minor allele frequency. An additional sampling was done matching also whether the variant was exonic or not.

# Allele-specific analysis

## Allele-specific expression (ASE) analysis

Allele-specific expression analysis was based on binomial testing of allelic ratios over heterozygous sites of each individual. First, we excluded sites that are susceptible to allelic mapping bias: 1) sites with 50bp mapability <1 (REF) implying that the 50bp flanking region of the site is non-unique in the genome, and 2) simulated RNA-seq reads overlapping the site show >5% difference in the mapping of reads that carry the reference or non-reference allele. In all the analyses, we filtered for mapping quality >150, NM<5, and base quality >10.

Next, we calculated the expected reference allele ratio for each individual by summing up reads across all sites separately for each SNP allele combination after down-sampling reads of sites in the top 25th coverage percentile in order to avoid the highest covered sites having a disproportionally large effect on the ratios. These expected REF/TOTAL ratios correct for any remaining genome-wide mapping bias as well as GC bias in each individual (SFig).

Finally, for all the sites covered by >=8 reads in each individual, we calculated a binomial test of the REF/NONREF allele counts, using the expected ratio described above. Except for the NMD analysis (see below), we used only sites with >=16 reads, and sites where both alleles are observed in RNA-sequencing data in order to verify that the genotype is a true heterozygote (SFig).

In many analyses, differing coverage between sites creates noise due to difference in power to call ASE. To correct for this, in many analyses we used only sites with >=30 reads (SFig), and sampled all sites to exactly 30 reads. In a further analysis of ASE differences between individuals, we calculated allelic expression distances between all sample pairs as the median of absolute REF/TOTAL ratio differences of all the shared heterozygous sites between individuals after sampling the reads to 30.

## Allele-specific transcript structure (ASTS) analysis

Allele-specific alternative splicing (ASAS) is a novel sister method of ASE, and aims at detecting differences in isoforms between the two haplotypes of an individual. As in ASE, we look at reads overlapping heterozygous coding sites, and the allele of this site in the RNAseq data tells which read (and its mate) comes from which haplotype. The distribution of these reads to exons is then quantified.

For every sample, we first retrieved all heterozygous sites that are covered by >= 20 RNAseq reads, after mapability filter as in ASE analysis. Using the pysam package (REF), we scanned the bam file to extract all the reads and their mates that overlap the site, separated them to reads with REF or ALT allele, and printed out a pseudo-sam file that contains information of which SNP each read overlaps, and if it carries the REF or NONREF allele.

For this file, we can run our standard exon quantification, and calculate the number of REF and ALT read overlaps in all the exons. We kept only exons with >=10 reads of each allele, and required a total of >=20 REF and NONREF reads in the remaining exons. We used Fisher test to estimate whether the read counts in exons are different for REF and NONREF reads. For each site, we calculated a quantitative measure analogous to ASE allelic ratio (maximum imbalance for all exons of a site compared to the total REF/NONREF ratio).

# Loss-of-function analysis

## Nonsense-mediated decay

ASE data with coverage >=8… Manny to write additional methods

## Splice scores

Log-odd scores of variant effect in splice motifs were been computed employing the 1st order Markov Models for splice donor and acceptor sites of human U2-dependent introns from the gene prediction program GeneID (PMID: 18428791). The scoring has been applied to the ~478,000 splice sites currently included in the Gencode v12 reference annotation, and alternative sites have been distinguished from constitutive ones if at least one transcript included the respective site’s sequence but has been annotated to not splice at the site. Additionally, ~790,000 novel sites have been identified by variations of annotated introns, i.e., split-mappings between an annotated splice site and an hitherto un-annotated site in the region of ±30nt of the site(s) annotated in Gencode to be spliced together with aforementioned known site.

Splice site variants have been inferred from the 1000 Genomes Phase 1 genotype data, considering all SNPs and indels that modify the splice site sequence relevant for the Markov model.