Integration of transcriptome and genome sequencing uncovers functional variation in human populations

Tuuli Lappalainen[1], M Sammeth[2,3], N Kurbatova[4], J Monlong[3], M Friedländer[2], M Rivas[5], T Strom[6], PAC 't Hoen[7], M Barann[8], O Karlberg[9], M Sultan[10], T Griebel[3], T Wieland[6], E Lizano[2], I Padioleau[1], S Schreiber[8], H Lehrach[10], S Antonarakis[1], GJ van Ommen[7], R Sudbrak[10], Robert Häsler[8], A Brazma[4], AC Syvanen[9], P Rosenstiel[8], T Meitinger[6], R Guigo[3], I Gut[3], X Estivill[2], ET Dermitzakis[1], on behalf of the Geuvadis Consortium[1,2,3,4,6,7,8,9,10]

1 Dept of Genetic Medicine and Development, University of Geneva; 2 Center for Genomic Regulation and UPF, Barcelona; 3 Centro Nacional de Analisis Genomico, Barcelona; 4 European Bioinformatics Institute, Hinxton; 5 Wellcome Trust Centre for Human Genetics, Oxford; 6 Institute of Human Genetics, Helmholtz Zentrum München, Munich; 7 Center for Human and Clinical Genetics, Leiden University Medical Center; 8 Institute of Clinical Molecular Biology, University of Kiel; 9 Department of Medical Sciences, Uppsala University; 10 Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin

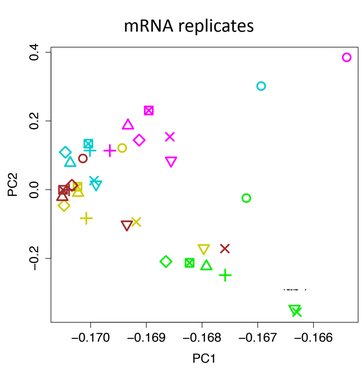
mRNA and smallRNA sequencing of 465 samples from the 1000 genomes project

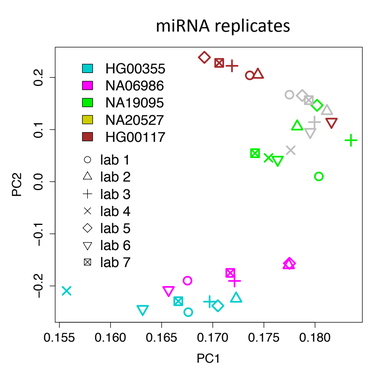
Aims of the study: (1) How to do distributed RNA sequencing? (2) What can we learn of transcriptome variation and its genetic component by integrating genome and transcriptome data from hundreds of individuals? (3) Create one of the biggest reference datasets for transcriptomics

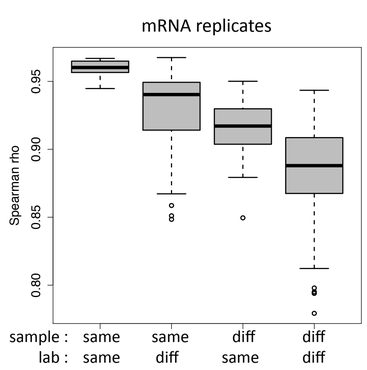
	1000G Phase1 + Phase2	RNAseq
TSI	92 + 1	93
GBR	86 + 10	96
FIN	89 + 6	95
CEU	79 + 13	92
YRI	77 + 12	89
TOTAL	423 + 42	465

RNA sequencing in 7 institutes with Illumina TruSeq protocol.
- Random distribution of samples
- Replicates: 5 samples in each lab + 168 samples in two labs.
- 32 billion total mRNAseq reads (median 48 M / sample)
- 700 M good-quality miRNAseq reads (median 1.5 M / sample)
- Lab effects do not overwhelm biological variation

From 1000 Genomes: 27 M total variants, of which 11 M >5% MAF







Quantitative trait loci for expression levels and splicing

We performed a cis-QTL analysis using genetic variants >5% MAF in 1MB window around genes, and Spearman rank correlation with (1) exon quantifications to find expression QTLs (eQTLs) (2) ratio of the most common transcript to find splicing QTLs (sQTLs). We ran permutations for CEU+GBR, and used a 0.01 permutation threshold for eQTLs and 0.001 for sQTLs.

Pop	N	Genes with eQTL (FDR)	*Best eQTL indel (null 8.9%)*	Genes with sQTL (FDR)
CEU+GBR	161	2608 (5.1%)	*375 (14.4%)*	121 (9.1%)
TSI	92	1748 (7.7%)	*242 (13.8%)*	102 (10.8%)
FIN	89	1822 (7.3%)	*255 (14.0%)*	104 (10.6%)
YRI	77	2138 (6.3%)	*242 (11.3%)*	108 (10.2%)
EUR union	342	3898	*NA*	230
ALL union	419	4895	*NA*	274

We discover:

- Thousands of eQTLs
- Hundreds of sQTLs
- High overlap between populations
- Enrichment of indel eQTLs ($p < 10^{-4}$ in all populations)

eQTL pi1 (low p value enrichment)

eQTL in	CEUGBR	FIN	TSI	YRI
CEUGBR	1	0.700	0.688	0.374
FIN	0.654	1	0.657	0.401
TSI	0.615	0.632	1	0.576
YRI	0.386	0.382	0.472	1

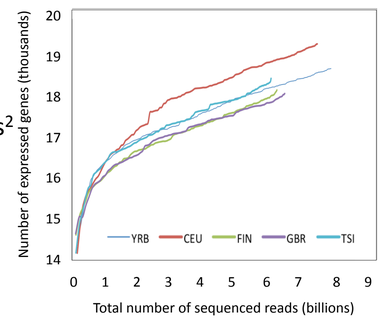
sQTL pi1 (low p value enrichment)

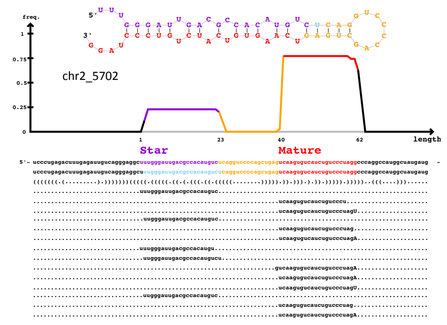
sQTL in	CEUGBR	FIN	TSI	YRI
CEUGBR	1	0.425	0.381	0.753
FIN	0.078	1	0.227	0.445
TSI	0.332	0.580	1	0.726
YRI	0.229	0.579	0.491	1

Population-scale deep sequencing improves gene discovery of miRNAs and poly-A transcripts

We detect 1615 out of the 1921 mature miRNAs in the miRBase database[1], 394 in >90% of the samples. Additionally, we discover 250 novel miRNAs with an estimated 30% FDR (see example below)

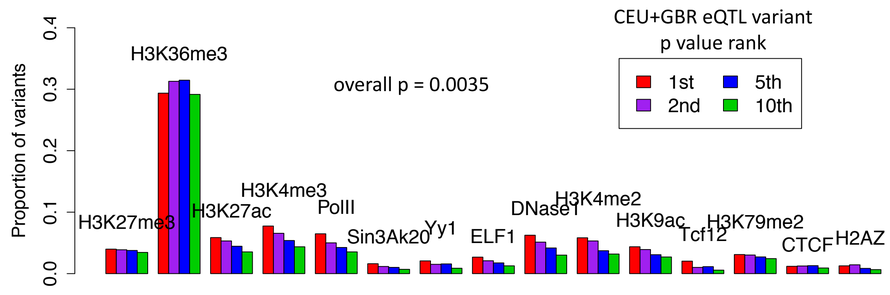
Population diversity and increasing total read count add significantly to the number of annotated genes[2] that are detected in the dataset[3]





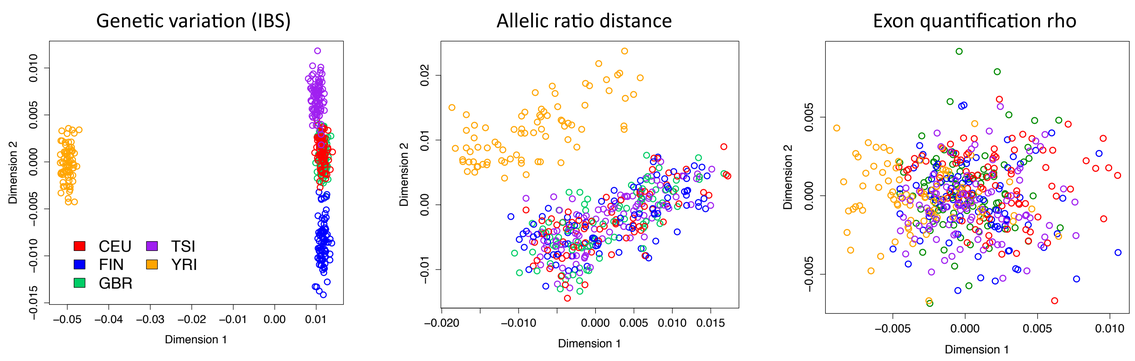
Functional annotation of eQTLs points to causal variants and regulatory mechanisms

Distinguishing the causal variant underlying a cis-eQTL signal has been a challenge. We find that the best eQTL variants overlap functionally annotated regions more often (Ensembl Regulatory Build, Annotated Features in GM12878), which suggests that we are discovering causal regulatory variants. Yet, in 25% of eQTLs none of the significant variants have an overlap with these functional elements.

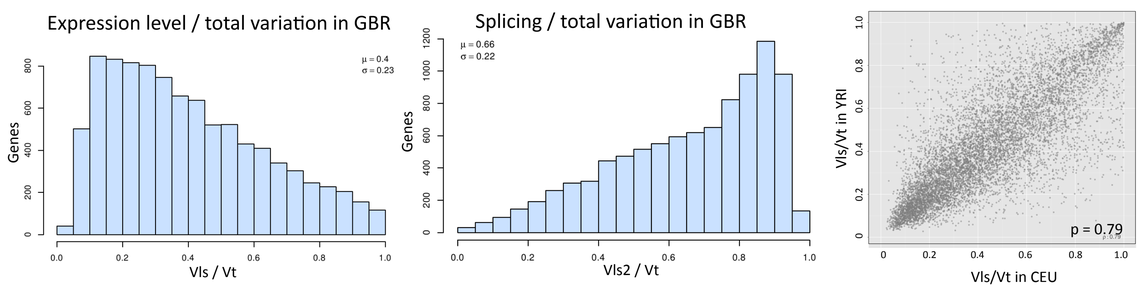


Genome-wide trends of transcriptome variation

Transcriptome variation shows clustering by continental groups, allelic ratios more than expression levels. This suggests a strong genetic component in allelic expression.

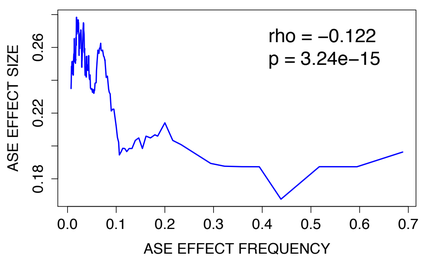


We estimate that only 40-45% of transcript level variation between individuals is due to variation in gene expression levels, and the rest is mainly from variation in splicing. The amount of splicing variation per gene is well correlated between populations. [4]



Effects of rare and loss-of-function variants in the transcriptome

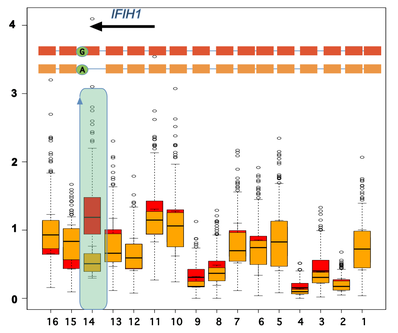
Allele specific expression analysis allows us to estimate the effect size of cis-regulatory events in a manner that is unbiased with respect to frequency. We observe that rare allelic effects have significantly bigger effect sizes, highlighting the importance of characterizing rare regulatory variants.

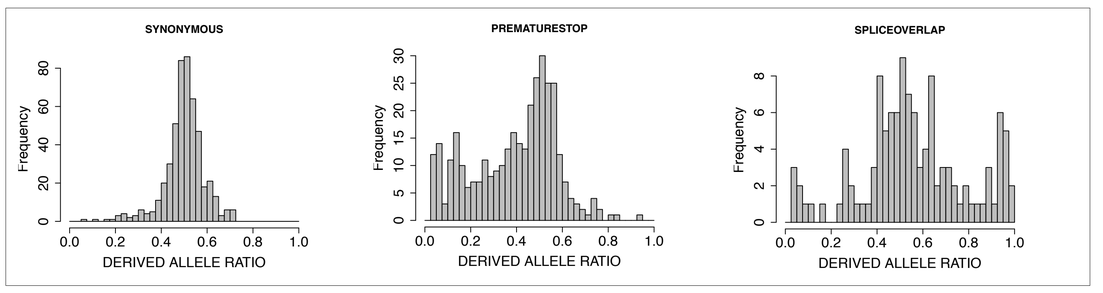


We can also functionally validate transcriptome effects of variants that are predicted to affect transcript structure – for example:

- splice site variants leading to exon skipping (right)

- premature stop variants leading to nonsense mediated decay (below) and splice site variants changing the allelic balance in complex ways





Acknowledgements: 1000 Genomes consortium; the Geuvadis project funded by the European Union.

References: (1) Kozomara A, Griffiths-Jones S (2011) NAR 39:D152-D157 (2) Harrow et al. (2006) Genome Biol 7:S4 (3) Montgomery et al. (2010) Nature 464:773-7 (4)Gonzalez-Porta et al (2012) Genome Res. 22:528-38.