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

Tuuli Lappalainen[1], M Sammeth[2,3], M Friedländer[2], PAC 't Hoen[4], M Rivas[5], J Monlong[2], M Gonzales-Porta[6], N Kurbatova[6], T Griebel[2,3], Ferreira P[2], M Barann[7], T Wieland[8], L Greger[6], M van Iterson[4], J Almlof[9], P Ribeca, I Pulyakhina[4], D Esser[7], E Lizano[2], M Sultan[10], D MacArthur[11], I Padioleau[1], T Strom[8], McCarthy M[5], H Lehrach[10], S Schreiber[7], R Sudbrak[10], A Carracedo Alvarez[12], S Antonarakis[1], Robert Häsler[7], AC Syvanen[9], GJ van Ommen[4], A Brazma[6], T Meitinger[8], P Rosenstiel[7], R Guigo[2], I Gut[3], X Estivill[2], ET Dermitzakis[1], on behalf of the Geuvadis Consortium[1,2,3,4,6,7,8,9,10,12]

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 Center for Human and Clinical Genetics, Leiden University Medical Center; **5** Wellcome Trust Centre for Human Genetics, Oxford; **6** European Bioinformatics Institute, Hinxton; **7** Institute of Clinical Molecular Biology, University of Kiel; **7** Institute of Human Genetics, Helmholtz Zentrum München, Munich; 9 Department of Medical Sciences, Uppsala University; **10** Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin; 11 Massachusetts General Hospital; 12 University of Santiago de Compostela

mRNA and miRNA sequencing of 465 samples from the 1000 Genomes project

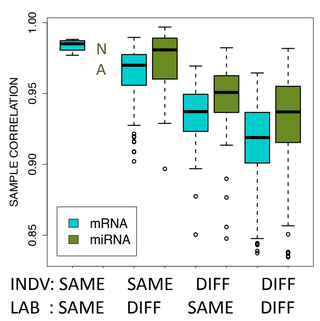
Aims of the study: (1) How to do distributed RNA sequencing? (2) What can we learn of transcriptome variation and its genetic component by integrating genome and transcriptome data from hundreds of individuals? (3) Create one of the biggest reference datasets for transcriptomics.

	mRNA	miRNA
TSI	93	89
GBR	94	94
FIN	95	93
CEU	91	87
YRI	89	89
TOT	462	452

RNA sequencing in 7 institutes with Illumina TruSeq protocol [1]
- Random distribution of samples
- Replicates: 5 samples in each lab + 168 samples in two labs.
- Genotypes from 1000 Genomes [2]: 27 M total variants. 90% of samples in Phase1, the rest imputed from Omni2.5 M SNP data

RNAseq reads	Individual median
mRNA total	58.4 M
mRNA QC pass	48.8 M
miRNA total	8.8 M
miRNA QC pass	1.2 M

Quantifications	In >50% individuals
Genes	14,779
Exons	141,951
Transcripts	74,533
Splice junctions	134,293
Fusion genes	5
Transcribed repeats	47,438
RNA edited sites	100
miRNA genes	706

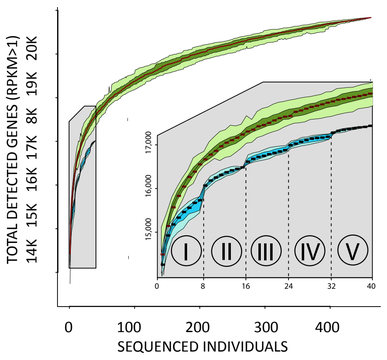


Distributed RNA-sequencing works well: technical variation due to laboratory effects is less than biological variation between samples

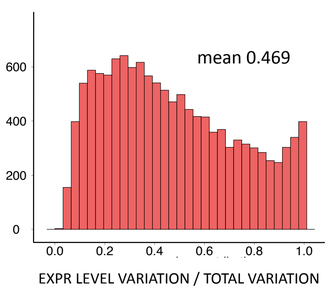
Data available : www.geuvadis.org ; ENA accessions E-GEUV-1, E-GEUV-2

Transcriptome variation within and between populations: mRNA, miRNA, and their interactions

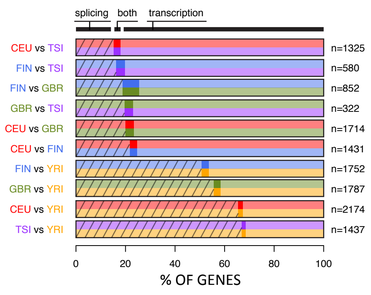
Population diversity adds 10% to gene detection



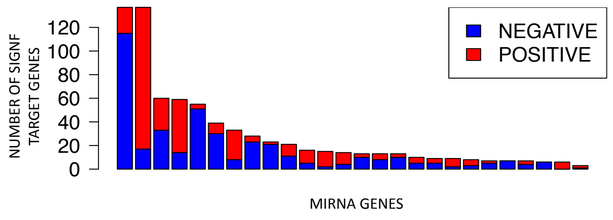
Gene expression levels and splicing contribute almost equally to transcription variation within populations. [3]



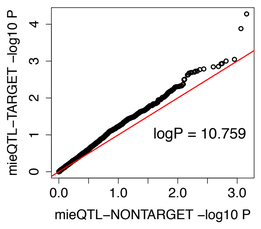
Differential splicing is more common between than within continents



The expression levels of 26 miRNAs correlate with predicted target[4] quantifications in the population



miRNA cis-eQTLs have increased trans-eQTL signal with miRNA target genes

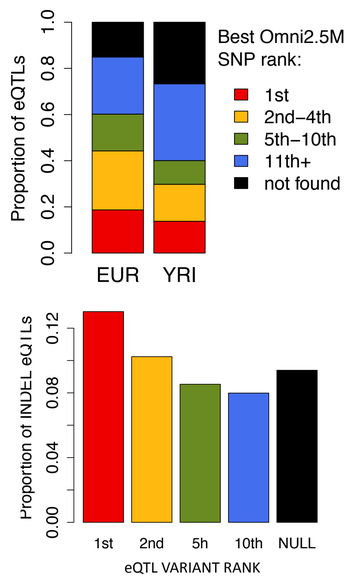


Thousands of expression and splicing *cis*-QTLs with increased discovery of causal variants

The majority of protein-coding genes and 10% of miRNA genes have a common regulatory variant.

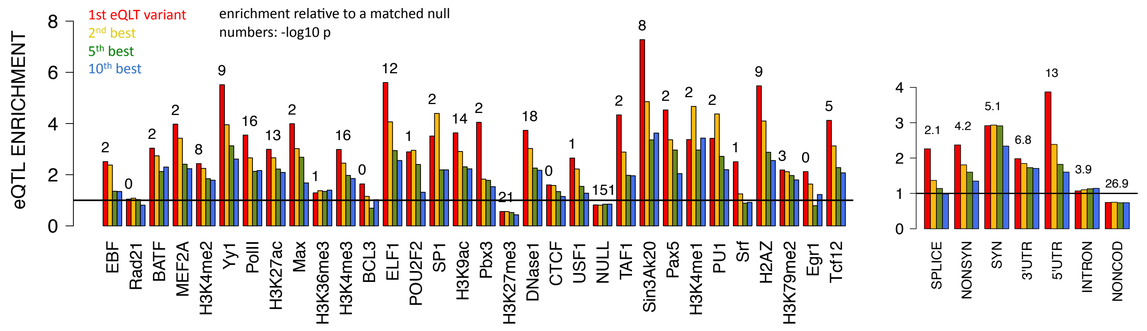
Genome sequencing data allows better discovery of the best-associating variants, and shows a significant enrichment of indels.

	exon eQTLS (12982)	exon link asQTLs (16172)	mi-eQTLs (644)
EUR (n=373)	7486	will have the numbers today	57
YRI (n=89)	2308		15
EUR & YRI union	7877		60

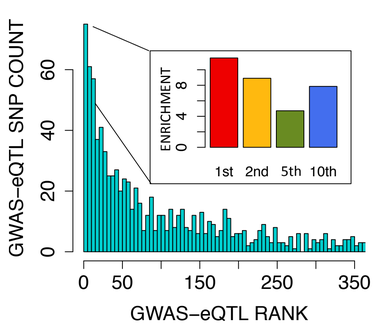


Enrichment of eQTLs in functional regions uncovers causes and effects of regulatory variation

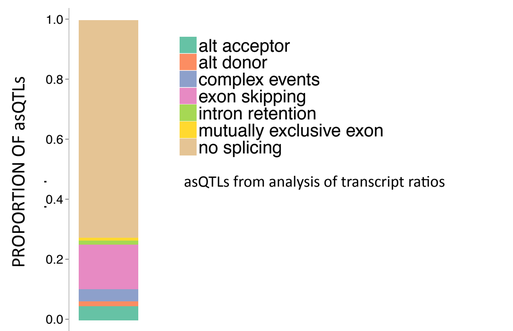
The best eQTL variants are significantly enriched in functionally annotated regulatory and coding regions (Ensembl Regulatory Build, Gencode v12), with an overrepresentation of especially promoter and enhancer annotations as well as splicing and nonsynonymous variants.



Variants from the NHGRI GWAS database are enriched among top eQTLs.

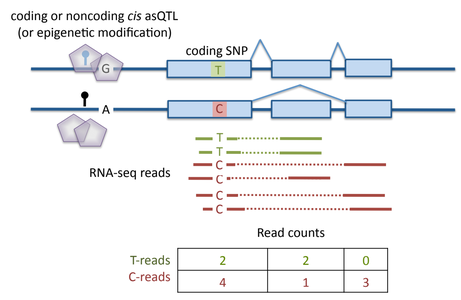


Most asQTLs affect UTR length rather than splicing.

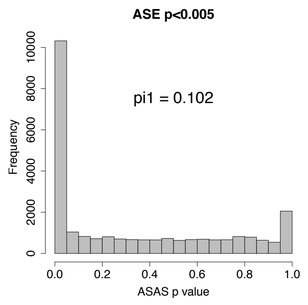


Variation in allelic expression is often driven by transcript structure variation and is dominated by rare effects

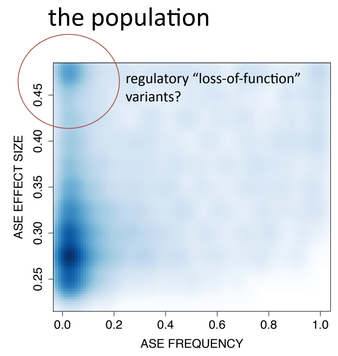
Detection of allele-specific expression (ASE) and transcript structure (ASTS)



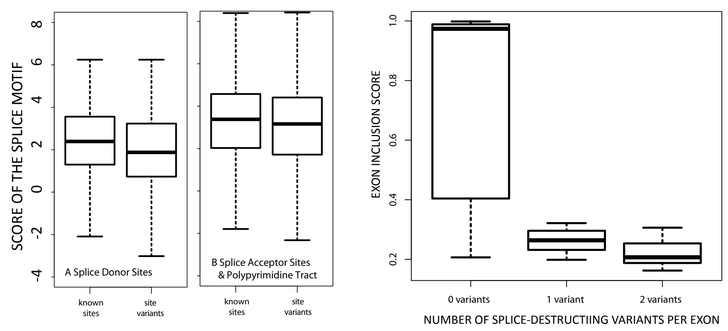
Much of ASE is driven by transcript structure changes



The majority of ASE is caused by rare events in the population

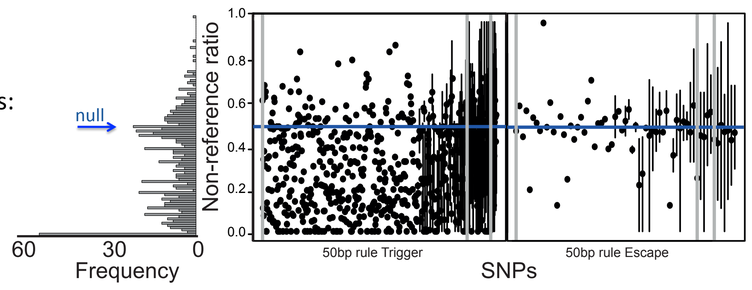


Functional characterization of loss-of-function variants



Genetic variants in splicing motifs significantly decrease both predicted and observed splicing efficiency.

Premature stop-codon variants lead to nonsense-mediated decay frequently but not always: better predictions needed



Acknowledgements: 1000 Genomes consortium; the Geuvadis project funded by the European Union. **References**: (1) Marco-Sola et al. (20012) Nature Methods in press (2) The 1000 Genomes Project Consortium (2012) Nature 491, 56–65 (3) Gonzalez-Porta et al (2012) Genome Res. 22:528-38. (4) Lewis et al. (2005) Cell 120(1):15-20