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 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.
- Random distribution of samples
- Replicates: 5 samples in each lab + 168 samples in two labs.
- Genotypes from 1000 Genomes: 27 M total variants. 90% of samples in Phase1, the rest imputed from Omni2.5 M SNP data

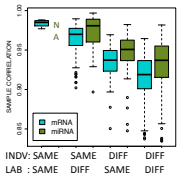
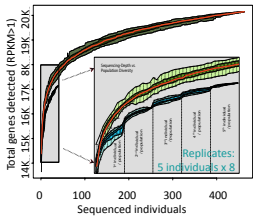


Table of stats

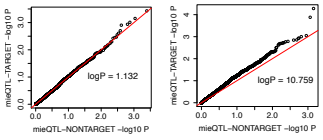
Transcriptome variation within and between populations: mRNA, miRNA, and their interactions

Population diversity adds 10% to gene detection



Both gene expression levels and splicing contribute to variation within and between populations

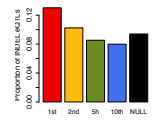
Expr of total
wi & bw pop

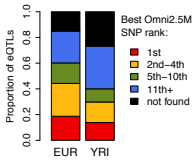


miRNA

Thousands of expression and splicing QTLs with increased discovery of causal variants

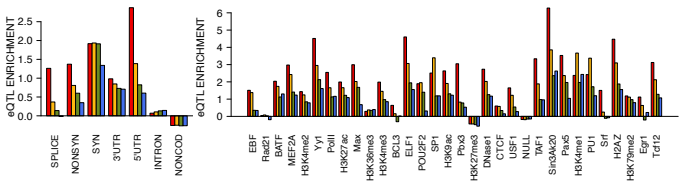
	exon eQTLS	exon inclusion asQTLs	gene eQTLs	transcr ratio asQTLs	mi-eQTLs
EUR (n=373)					
YRI (n=89)					
EUR & YRI					



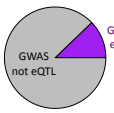


Enrichment of eQTLs in functional regions uncovers causes and effects of regulatory variation

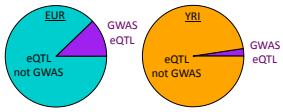
The best eQTL variants are enriched in functionally annotated regions more often than a matched null (Ensembl Regulatory Build, Annotated Features in GM12878; coding annotations from Gencode v12).



16% of the NHGRI GWAS database variants are significant eQTLs in our dataset

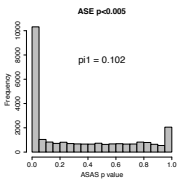


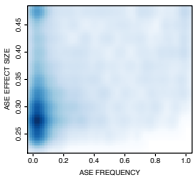
11% EUR and 2% of YRI eQTL genes have an eQTL in the GWAS database



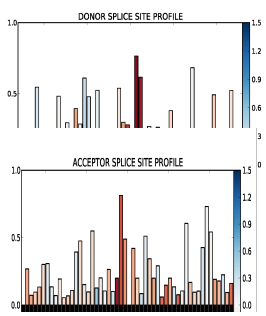
sQTL
other QTLs

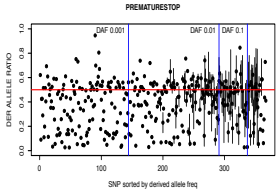
Variation in allelic expression: often driven by allele-specific splicing, and dominated by rare effects





Loss-of-function effects can be characterized by RNAseq





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.