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

T Lappalainen1, M Sammeth2,3, N Kurbatova4, J Monlong3, M Friedländer2, M Rivas5, T Wieland6, O Karlberg7, M Barann8, PAC ‘t Hoen9, T Griebel3, M Sultan10, D MacArthur11, E Lizano2, T Strom6, S Schreiber8, H Lehrach10, S Antonarakis1, GJ van Ommen9, R Sudbrak10, R Häsler8, A Brazma4, AC Syvanen7, P Rosenstiel8, T Meitinger6, R Guigo3, I Gut3, X Estivill2, ET Dermitzakis1, The Geuvadis Consortium1,2,3,4,6,7,8,9,10

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 Department of Medical Sciences, Uppsala University; 8 Institute of Clinical Molecular Biology, University of Kiel; 9 Center for Human and Clinical Genetics, Leiden University Medical Center; 10 Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin; 11 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston

Functional interpretation of genetic variants discovered in human genome sequencing is essential to understand human phenotypic variation. We sequenced mRNA and small RNA of LCLs from 465 individuals from CEU, TSI, GBR, FIN and YRI populations of the 1000 Genomes samples, with a median 40M\* mapped reads per sample for mRNA and 1.5M for miRNA. Our 203 replicate samples allowed us to measure technical variation in high precision. In our analysis, we first characterized human transcriptome variation in unprecedented depth. We show how the total number of expressed genes increased from a median of 400 miRNA and 14000\* protein-coding genes in a single individual to a total of 1651 and 20K in the entire sample, with nearly every individual expressing some unique genes. We discovered almost 200 novel miRNA genes, unannotated splice junctions and novel transcriptionally active regions, many of which were rare in the population. In most genes individual differences in splicing contributed more (mean 60-70%\*) to the total variation than differences in total gene expression levels. Furthermore, the combination of high-quality transcriptome and genome sequencing data gave us the opportunity to characterize both common and rare regulatory variants. We discovered over 5000 common eQTLs and hundreds of splicing QTLs (FDR < 10%), and show that indel variants are more likely to affect gene expression than SNPs (p < 10^-4). The enrichment of eQTL variants in regions with regulatory annotation (p < 10^-10) points to causal regulatory variants and sheds light on functional mechanisms underlying eQTL effects. We uncovered rare regulatory effects by allele-specific expression analysis, showing that they have higher effect sizes (p = 10^-15) and account for the majority of allelic ratio differences between individuals. Finally, we characterized transcriptome effects of hundreds of loss-of-function variants in our dataset. In addition to functionally validating variants predicted to cause e.g. nonsense-mediated decay or disrupt splice-sites, we developed a model to improve functionality predictions of previously unseen variants base on their properties. Altogether, this study takes us beyond cataloguing putative functional variants towards understanding and predicting the cellular effects of variants in the human genome.

\*get the exact numbers