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 from 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.

RNA sequencing in 7 institutes with Illumina TruSeq protocol.
- Replicates: Samples in each lab = 168 samples in two labs.
- 32 billion total miRNA reads (median 48 M / sample)
- 700 M good-quality miRNA 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

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

We detect 1615 out of the 1821 mature miRNAs in the miRBase database, 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 that are detected in the dataset

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.

Quantitative trait loci for expression levels and splicing

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

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 (Ensemble 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.

Functional annotation of eQTLs points to causal variants and regulatory mechanisms

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

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

Allelic 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.
