**Transcriptome and genome sequencing uncovers functional variation in human populations**

Geuvadis main paper outline

*Preliminary* pointers to main figures and tables in green.

Major question marks and pending analyses marked in blue

\* Meta-level messages of each section marked wit asterisk

1. **Data production and quality**

\* We have a great dataset that has been processed with most up-to-date methods

* Table: basic numbers of read counts and quantifications
  + Status: stats collection ongoing
    - Tuuli
* Distributed RNAseq works well
  + Figure: replicate correlations for mRNA & miRNA before and after normalization
  + Status: almost done
    - Peter & Tuuli

1. **Transcriptome variation in human populations**

\* Each individual has lots of rare transcriptome features that are only seen when RNA-sequencing populations

\* Annotations tend to lack rare (and African?) features

* We keep on finding more expressed genes with every new individual sequenced: the benefit of the N+1 transcriptome
  + Figure: Number of quantified genes as a function of sequenced samples
  + Status: done / almost done
    - Micha
* Individual and population variation in mRNA transcriptome is driven almost equally by both expression level and splicing variation
  + Mean across genes is about 50-50, with large variation between genes in a manner that is usually consistent between populations
    - Status: almost done
      * Jean
  + However, there are hundreds of genes with differential expression or differential transcript ratios between populations
    - Status: almost done
      * Pedro (expression levels), Mar (transcript ratios)
      * Ongoing analysis: biologically relevant vs cell line batch effects. Are transcript ratios/splicing less sensitive to batch effects?
  + Figure: Some characterization of splicing variation between populations, and inserts of example genes that are similar/different in terms of expression/splicing.
* Splicing variation; both major splicing events and soft splicing
  + Status: Analysis ongoing
    - Micha (“hard” splicing), Matthias (soft splicing)
* Transcription beyond the annotated boundaries of genes
  + Hundreds of fusion genes of which especially the population-specific and rare ones are novel
    - Status: Almost done
      * Liliana
  + N-TARs
    - Status: Analysis ongoing
      * Daniela
* RNA editing is variable between individuals and populations
  + Status: Analysis ongoing
    - Thomas W
* miRNA variation
  + Status: Analysis ongoing
    - Marc?
* **Figure(s):** Some representation of variation in transcriptome features (across all: splicing, fusion genes, n-TARs, editing, miRNA, or selected few with most exciting results/novelty value).
* miRNA variation contributes to mRNA variation in human populations
  + Status: Analysis ongoing
    - Marc, Peter’s group
  + Figure: miRNA-mRNA interaction

1. **Regulatory variation in the human genome**

\* Genome + RNA sequencing data gives us an unprecedented view to both rare and common regulatory variation and its functional mechanisms

* We find a lot of classical eQTLs, but we also go beyond that to discover a variety of transcriptome QTLs. This is a major analysis item that is likely to yield interesting and important findings that can be highlighted more
  + Independent regulatory variants for the same gene
  + Simultaneous effects on different transcriptome features (e.g. expression levels and splicing)
  + Status: Exon eQTLs done, others ongoing
    - Tuuli
  + Figure/Table: tQTL characteristics/statistics
* With genome sequencing data we can often identify likely causal variants and characterize the mechanisms how genetic variation affects gene expression
  + Figure: Enrichment of eQTL and sQTL putative causal variants in functional annotations, compared to a matched null.
    - Status: Analysis ongoing
      * Tuuli
* The vast majority of variation in allelic expression (and splicing?) is rare, and differences between individuals especially from the same population are predominantly driven by rare effects
  + Figure: Frequency distribution of ASE differences between individual pairs
  + Status:Done for ASE, ASAS running
* We can map rare regulatory variants that underlie some of the rare allelic effects. This is likely to be an extremely important class of functional variants.
  + Figure: Something to illustrate this
  + Status: analysis ongoing.
    - Tuuli

**4. Improved interpretation of loss-of-function variation**

\* Transcriptome sequencing gives us a wealth of information of loss-of-function variants: validation, improved prediction of functional effects, and frequent compensatory mechanisms that cancel out the predicted functional impact

\* Even the “easiest” class of functional variants is actually painfully complex

* Stop-gained variants lead to NMD in about ~60% (?) of cases, and we improve predictions when a variant is likely to cause/escape NMD
  + Figure: Visualization of NMD frequency & something about predictions?
  + Status: Half done
    - Manny & Tuuli
* Variants in the splice-site lead to disruption in splicing in X% of the cases, which can be partially predicted from the splice motif
  + Figure: Visualization of splice site variant effect on splice quantifications
  + Status: Half done
    - Manny & Micha
* There are various compensatory mechanisms to buffer the effects of LoF variants
  + In ~30% (?) of cases, heterozygote NMD doesn’t lead to decreased gene dosage (regulatory compensation?)
    - Analysis ongoing
      * Manny
  + Loss-of-function variants are enriched on lower expressed eQTL haplotypes and sQTL haplotypes that skip the exon (genomic compensation?)
    - Analysis ongoing
      * Tuuli
  + Figure: Something to visualize this

**5. Data sharing and visualization**

\* This is the one of the most important transcriptome variation reference datasets, and the data is there for everyone to access

Brief descriptions in the paper; most content online

* Annotations: the best functional annotation of 1000g Phase 1 variants
  + Analysis: done (unless we want updates)
    - Daniel
  + Where do we put the file?
* Data file sharing at EBI ENA/arrayexpress
  + Bam files
  + Quantificatons: exons, genes, transcripts, splice junctions?
* Visualization in Ensembl browser
  + Individual and population level results
  + At least expression level quantifications, eQTLs and ASE results