**Geuvadis RNAseq analysis plan for the main paper**

Based on presentations and discussion at the analysis group meeting in Geneva in April 16-17 2012. Summarized by Tuuli Lappalainen (UNIGE) in April 20, 2012. General points:

* Focus on biology: how does genomic variation contribute to variation to multiple levels of the transcriptome
* We should aim at a big main paper in a high ranking journal. There can be additional companion papers on:
  + methologody
  + separate biological analyses that don’t fit well into the scope of the main paper
  + issues that are addressed in the main paper but that need and deserve more in-depth analysis

Analysis topics below are ordered so that they move progressively to more interpretational analyses that require input from the previous steps – see the last page for a workflow. The order of the analyses does *not* reflect the importance that different types of analyses will have in the paper.

The names denote the person responsible for coordinating each analysis module and seeing that it gets done, as well as other people who want to be involved in the analysis.

1. **Data quality in a multicenter study**

Multicenter study design, biological versus technical variation.

* 1. mRNA

**Olof Karlberg (Uppsala)**, Jonas, Tuuli, Natalja, Micha

* 1. miRNA

**Marc Friedlaender (Barcelona),** Natalja, Esther

1. **DNA-RNA differences**

Variation in RNA editing of some very very well validated RNA editing sites. Validation of genotype calls can sometimes be important in LoF analyses.

**Thomas Weiland (Munich)**, Thomas S, Tim, Micha

1. **mRNA variation**

How much individual variation comes from differences in expression levels versus splicing. Exon/transcript quantifications and transcript ratios as quantitative traits for genetic association analysis.

Transcript ratios: **Jean Monlong (Barcelona)**, Pedro

Expression levels: **Tuuli Lappalainen (Geneva)**, Micha

1. **Variation of mRNA transcript features:**
   1. **Splice junctions**

**Micha Sammeth (Barcelona)**, Rob, Matthias, Pedro, Peter, Irina,

* 1. **UTR length and novel transcriptional active regions**

**Robert Haesler (Kiel)**, Matthias, Micha, Peter

Quantification of these features and their variation between individuals and populations. “Every individual has X number of unique splice junctions”. “We detect unannotated elongation in Y% of UTRs”. How much does knowledge of population variation add to the existing annotation? Do we see the biases of common European variation being better represented in the annotation? Quantifications as input for genetic association analysis.

**5. miRNA variation, quantitative and qualitative, effects of variants in miRNAs**

Quantitative variation in the levels of miRNA expression between individuals – both known miRNAs and unannotated ones. Qualitative variation due to genetic variants in the miRNA sequence. miRNA quantifications as input for genetic association analysis

**Marc Friedlaender (Barcelona),** Peter, Natalja, Esther

**6. miRNA – target interactions**

Do miRNA levels correlate with expression levels or variation in expression levels of their predicted target genes? How many targets can we validate? Can se see genetic variants leading to loss of target sites or birth of new ones?

**Marc Friedlaender (Barcelona)**, Peter & others from Leiden, Rob, Natalja, Esther

**7. Transcriptome QTLs in cis (cis-tQTLs)**

Characterization of genetic variants (snps and small indels) that associate to different transcriptome traits: mRNA expression levels, stochastic variation of mRNA expression, splicing, UTR length, miRNA expression levels. Several follow-up analyses to get beyond cataloguing the variants:

* Integration of different types of variants for the same gene into a single model
* Independent eQTLs for the same gene
* How much much these variants explain of allele specific expression
* How much of transcriptome trait variation do these variants explain, partitioned by gene type (ontology)
* Mapping the causal regulatory variants

**Tuuli Lappalainen (Geneva)**, Thomas W & S, Micha, contribution of quantitative phenotypes from many of the analyses listed above

**8. Functional annotation of tQTLs**

What can we learn from functional mechanisms underlying different types of tQTLs from the overlap of tQTLs and existing regulatory annotation of the genome?

**Tuuli Lappalainen (Geneva)**,

**9. Structural variant effects on the transcriptome**

Characterization of the effects that large CNVs have on transcript structure, gene fusions etc. eQTL analysis both in cis and trans – do these variants change expression of large blocks of genes?

**Tuuli Lappalainen (Geneva)**, (eQTLs), Peter, Kai (transcript structure)

**10. Mapping rare regulatory variants**

Allele specific expression - haplotype sharing approaches for mapping rare regulatory variants (see Montgomery et al. Plos Genetics 2011)

**Tuuli Lappalainen (Geneva)**,

**11. Loss-of-function variants.**

Functional effects and mechanisms of LoF variants annotated by the 1000g. Characterization of new putative LoFs variants especially at splice sites. NMD efficiency. Epistasis between common variation and LoFs, especially regarding splicing.

**Manuel Rivas (Oxford)**, Micha, Tuuli, Daniel, Thomas W & S

**12. Meta-level variation of transcriptome and genome**

Correlation of regulatory phenotypes and genetic similarity between individuals genome-wide and locally: distances between individuals genome-ASE-splicing-expression levels. Does high evolutionary conservation and/or lack of population variation correlate with level of transcriptome variation in different genes? Do genes in different pathways/ontologies differ in the amount of quantitative and qualitative variaton?

**Tuuli Lappalainen (Geneva)**,

**12. Visualization of the data**

Tools to visualize the data – both for the members of the consortium and for everyone after the publication of our paper. Should link with 1000g data, possibly also with Genevar from Geneva.

**Natalja Kurbatona (EBI)**

