GEUVADIS project

Differential expression analysis

Data

- 464 samples : 5 populations (CEU, FIN, GBR, TSI   YRI)

- Expression of 24 800 genes provided by read counts

- TweeDEseq (R):   uses a Poisson-Tweedie family of distributions. It improves the negative binomial distribution using a three parameter distribution. Well suited for datasets with more than 15 samples!

Processing

- After normalization (TMM) and filtering step :
  16 583 genes left (remove genes with less than 5 counts per million in all samples but one);

- Analysis performed in two modes: pair-wise comparison between the 5 samples; comparison of one sample against the others;

- Genes are considered to be differentially expressed when P < 0.05 and logFC > 2 or 3

Number of genes differentially expressed

Pairwise Comparisons:

Fold change > 2

	CEU	FIN	GBR	TSI	YRI
CEU	-	1 176	1 456	1 189	751
FIN	-	-	744	518	866
GBR	-	-	-	273	801
TSI	-	-	-	-	478
YRI	-	-	-	-	-

Fold change > 3

	CEU	FIN	GBR	TSI	YRI
CEU	-	570	723	551	236
FIN	-	-	228	116	350
GBR	-	-	-	65	344
TSI	-	-	-	-	167
YRI	-	-	-	-	-

Number of genes differentially expressed

One group against all the other samples :

Fold change > 2

	OTHERS
CEU	861
FIN	773
GBR	890
TSI	531
YRI	405

Fold change > 3

	OTHERS
CEU	334
FIN	332
GBR	417
TSI	248
YRI	116

Population Specific Up and Down regulated Genes
(log2 Fold Change > 2)
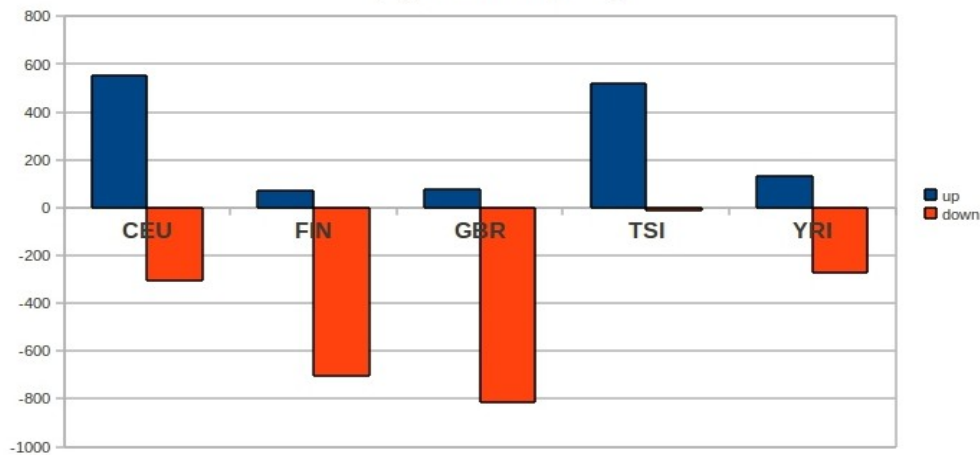
Population specific Up and Down regulated Genes
(log2 Fold Change > 3)
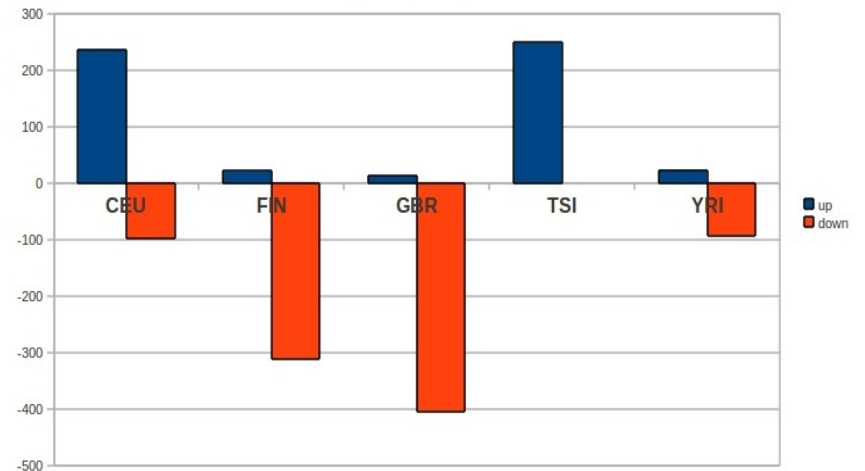
Functional Enrichment

- For functional enrichment analysis we considered the genes that are population specific (from the one vs all comparison) and up-regulated at the stringent threshold of logFC > 3.

- Only TSI and CEU show significant enrichment (FDR<0.05).

- See a more complete table at the wiki, but in both cases the GO terms related to cell adhesion appear as significantly enriched.

Discussion points

- CEU is the population that departs more considerably from the other populations in terms of diff. expressed genes;
- Comparison between CEU with FIN, GBR and TSI (logFC>2) we get respectively 79%, 79% and 80% of up-regulated genes in CEU.
- In the same comparison YRI shows 78%, 77% and 85% of genes down-regulated when compared with these three populations.

GEUVADIS project

Exon Splicing Inclusion Levels

Data

- Based on the exon counts calculate the percentage inclusion level – **PSI –** a measure that reflects the inclusion levels of the exon within the transcript;

- This measure was used and refined by Wang et al and Shapiro et al. Other programs like MISO implement this measure (Wang, E.T. et al "Alternative isoform regulation in human tissue", Nature, 2008. and Shapiro, I.M. et al , Plos Genetics 7, 2011.);

PSI calculation

- PSI = # inclusion_reads / (# inclusion_reads + # exclusion_reads) or PSI = A + B / (A +B + C)
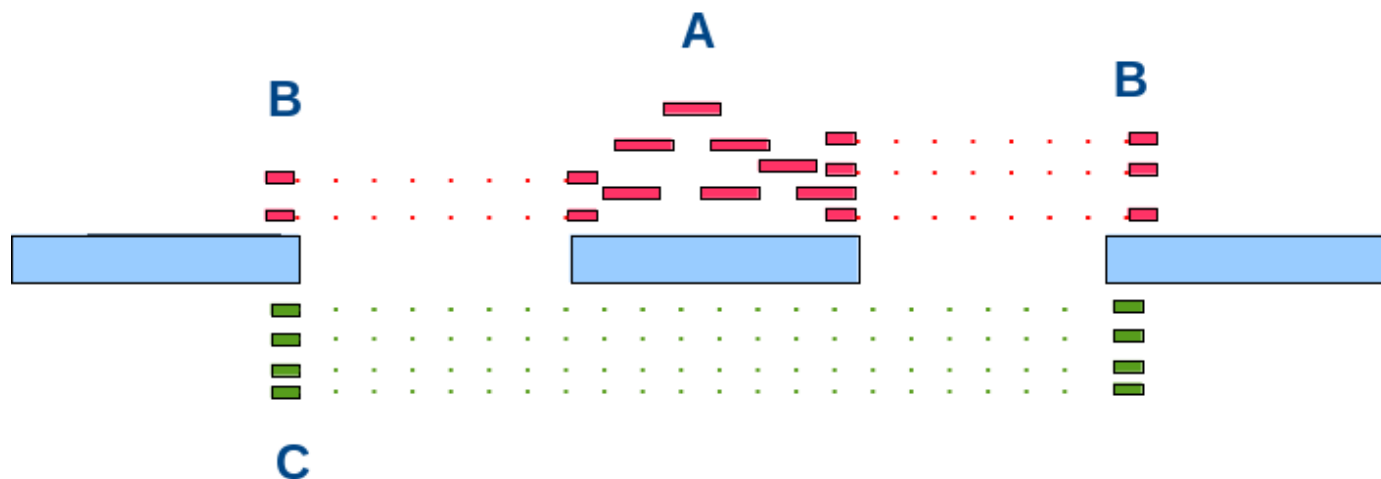
- **a** number of reads that map in the exon body (GD667.ExonQuantCount.txt) and **b** and **c** from flux files.

- A PSI value of **1** means that the exons is fully included and the other extreme a value of **0** means that the exon is not included.

Differentially Included Exons

- Differential exon inclusion between the different populations. For this we have done population pair-wise comparisons of the PSI values;

- Mann-Whitney test between the two sets followed by BH correction. To consider significantly differentially included exons:

  1) exon length > 150bp (with this we guarantee a minimum number of reads within the exon body);

  2) adjusted p-value < 0.05 (after BH correction);

 3) median diff ≥ 0.1 (absolute difference in the median of the two sets is at least 0.1);

Number of exons differentially included

- From a total of 199 798 internal exons, 175 210 exons have at least one PSI value and 64 120 exons have a PSI value for all samples (464 samples);

	CEU	FIN	GBR	TSI	YRI
CEU	-	6	13	7	68
FIN	-	-	10	3	59
GBR	-	-	-	4	57
TSI	-	-	-	-	57
YRI	-	-	-	-	-

- Selected exons with a minimum variabity (314 exons with stdv > 0.15) and performed hierarchical clustering to investigate if we find any particular inclusion/exclusion pattern on the samples
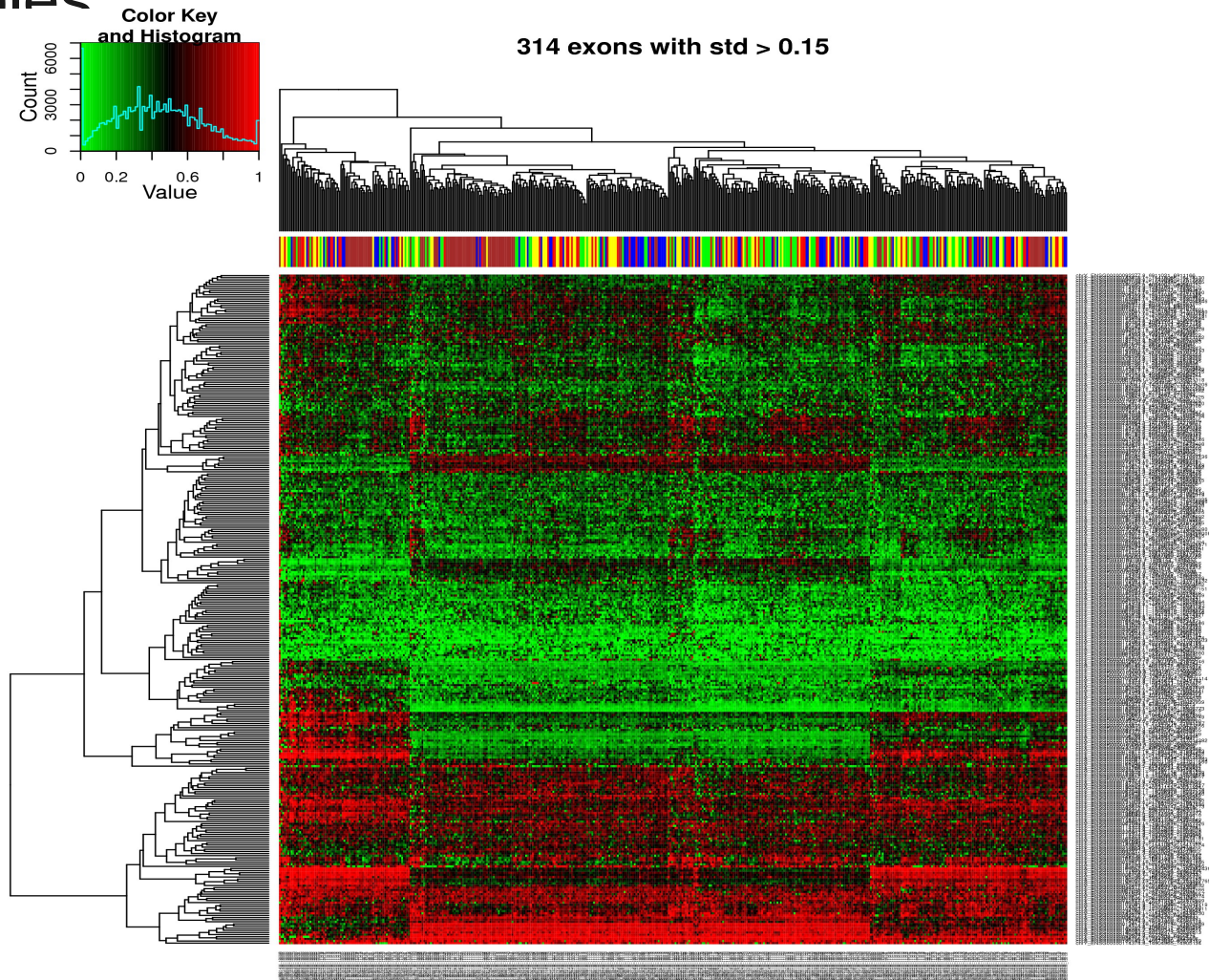
Discussion points

- Pair-wise population comparisons shows a relatively small number of significantly different included exons. In principle one would expect this since all the samples come from the same tissue. On the other hand the high number of tests my result in a stringent multiple test correction.

- Exons with a certain variability seem to have different patterns of inclusion although none correlates at population level.