Geuvadis - sQTL discovery and transcript variation.

Jean Monlong

Centre for Genomic Regulation

October 25th 2012

Splicing QTL discovery using transcript quantifications

Method

- Samples are grouped according to their genotype.
- Non-parametrical MANOVA-like test on the transcript ratios.
- Permutations are used to compute the pvalue.
- False Discovery Rate control using Benjamini-Hochberg algorithm.

- Genes expressing at least two isoforms.
- SNPs close/within ($\pm$5kbp) to a gene.
- At least 5 individuals in at least 2 genotype groups.

Results

Using additional criterion 20% change in median

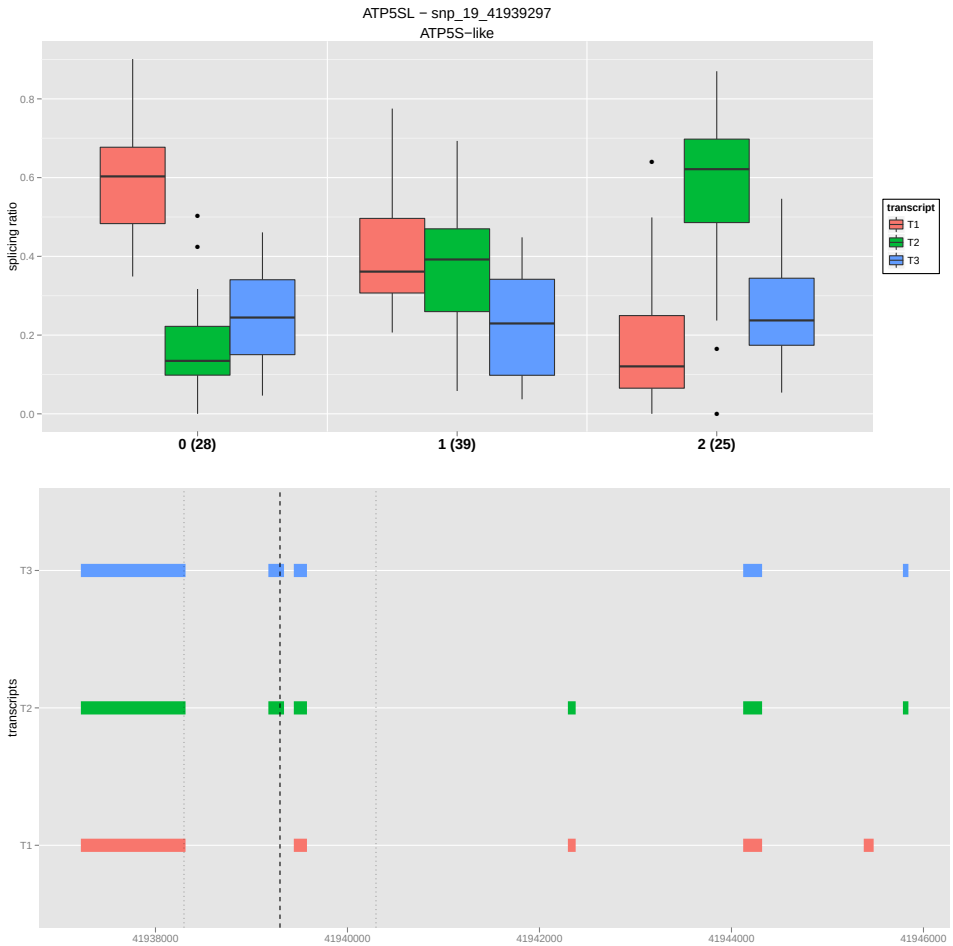
Population studied	nb studied association	nb studied genes	nb associations FDR 5% + 20% change	nb genes FDR 5% + 20% change	nb associations FDR 1% + 20% change	nb genes FDR 1% + 20% change
CEU	2178261	10015	2310	125	1579	61
FIN	2113898	9818	2209	152	1427	75
GBR	2070364	9731	1700	119	1316	60
TSI	2105315	9906	2136	145	1446	59
YRI	3199511	9814	1742	144	1365	79

Without additional criterion

Population studied	nb studied association	nb studied genes	nb associations FDR 5%	nb genes FDR 5%	nb associations FDR 1%	nb genes FDR 1%
CEU	2178261	10015	4242	268	2625	133
FIN	2113898	9818	4936	324	3411	158
GBR	2070364	9731	3457	283	2130	143
TSI	2105315	9906	4841	310	2971	156
YRI	3199511	9814	3898	364	2798	187

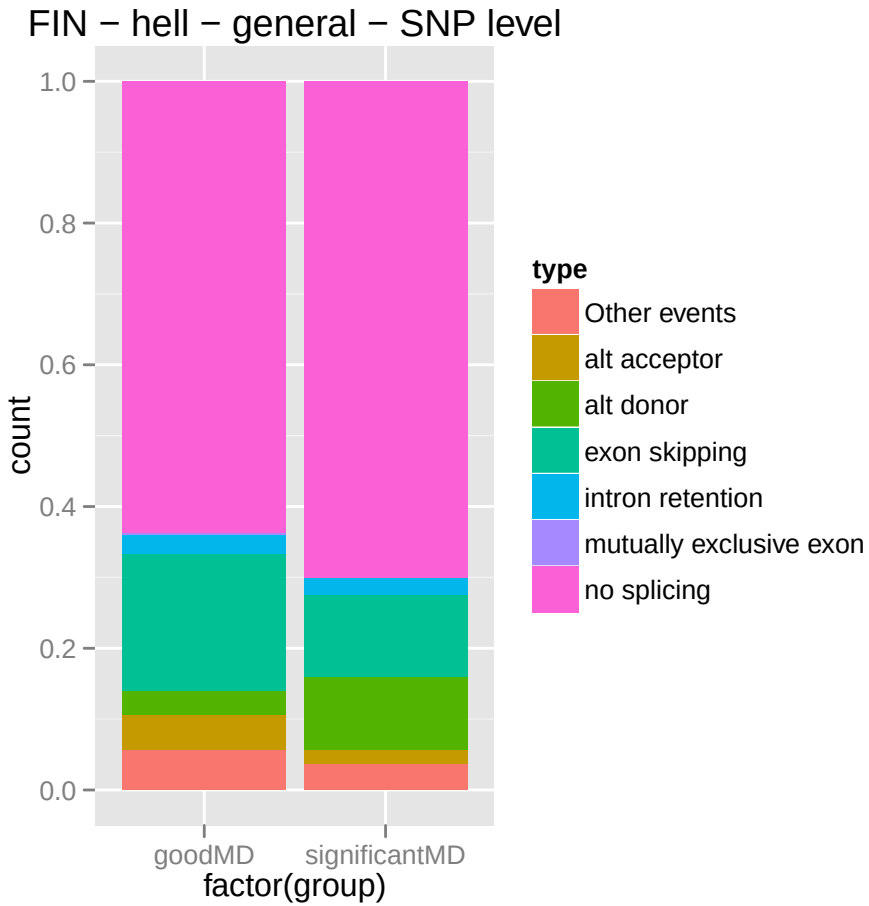
Classification events

Using AStalavista on the two transcripts with the biggest differences in the splicing ratio median, the event is inferred.



Classification events

- $\approx$ 70% of *no splicing*: changes in the first/last exons.
- Only 30% sQTL involve change in splicing.
- *other events*: changes too complex to classify.



sQTL position

At the SNP position

- Exonic/Intronic.
- Distance to the closest exon.
- Distance to the closest conserved region.
- Number of ESS/ESE motives at this position.

Window of 1kb upstream and downstream the SNP

- Intersection with exon.
- Number of GWAS hits.
- Intersection with conserved region.

A set of SNPs randomly chosen from the studied SNPs are used for comparison

sQTL position - FIN

FIN	Significant + 20% change	Good + 20% change + Splicing	Random
Exonic	31.5	11.2	7.4
Exon boundary w/ window	65.3	41.4	34.3
ESS/ESE motive	51.8	50.5	49.1
GWAS w/ window	17.4	2.3	0.7

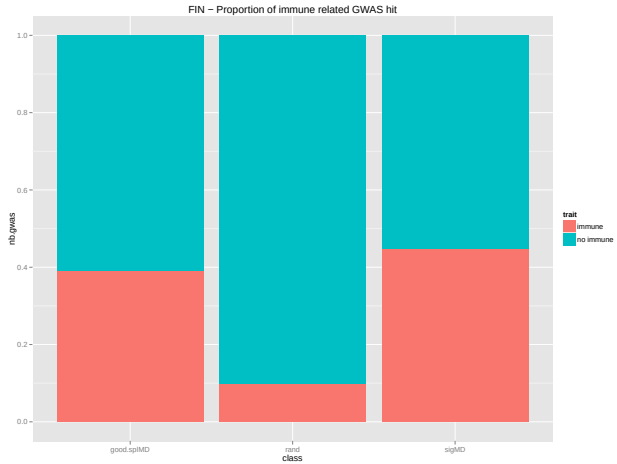
Distance to exon/conserved region

For intronic sQTLs, Mann-Whitney test to compare the distance to the closest exon or conserved region.

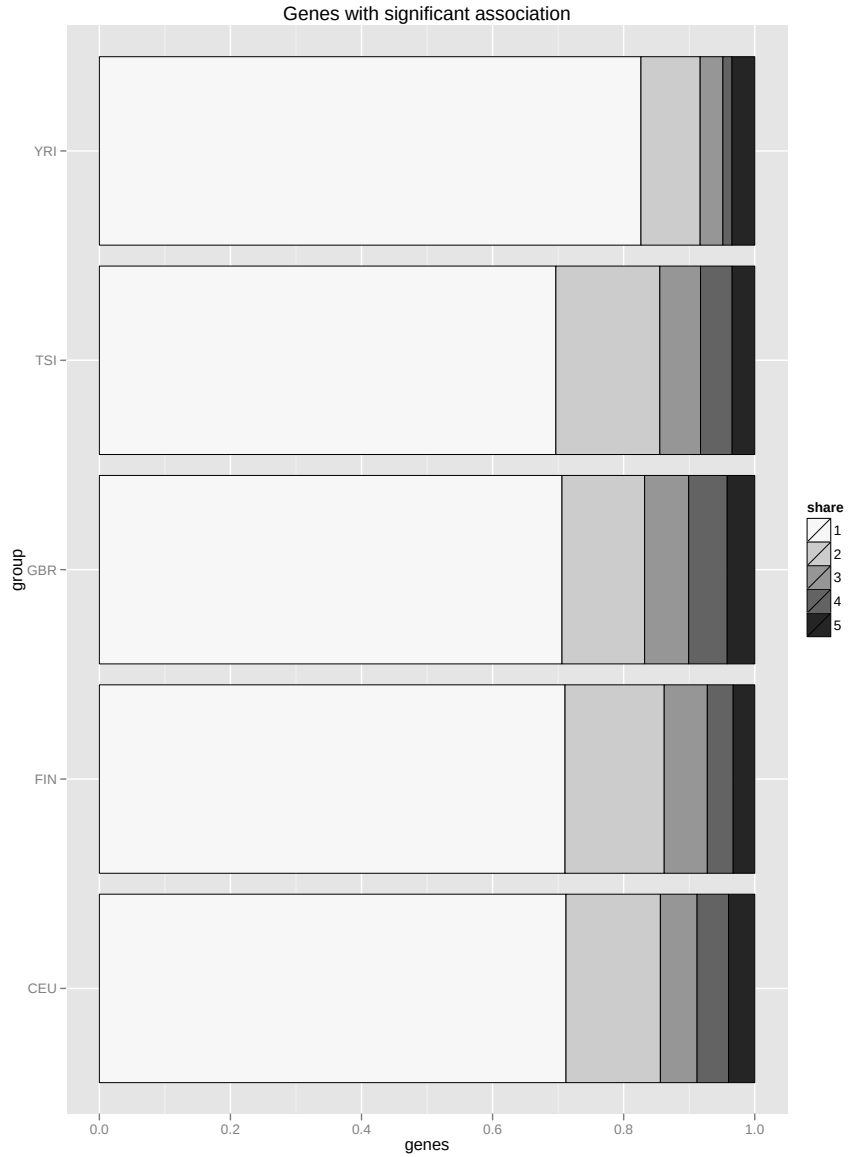
FIN	DistExon SigMD vs Rand	DistExon GoodMDSpl vs Rand	DistConsRegion SigMD vs Rand	DistConsRegion GoodMDSpl vs Rand
P-value	6.9e-23	0.000328	2.03e-05	0.0813
Direction	lower	lower	lower	lower

GWAS hits nearby

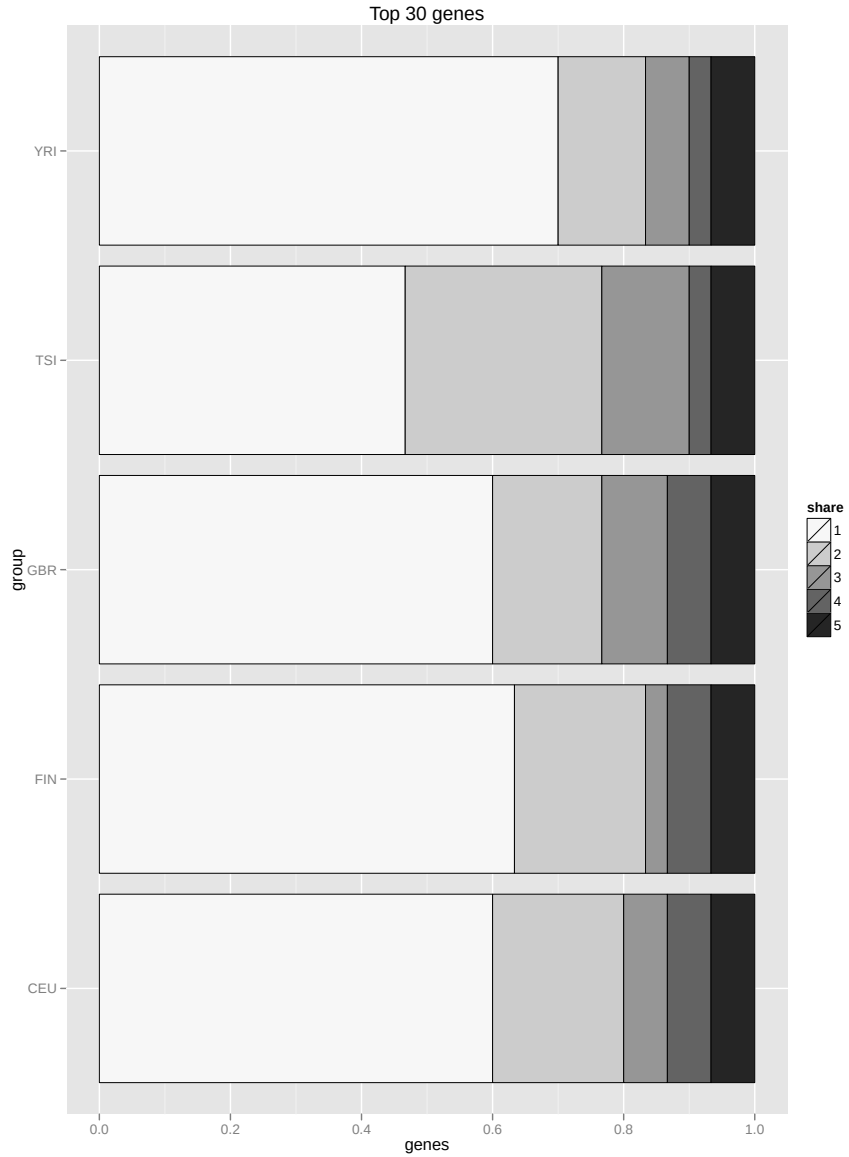
The GWAS hits seems to be related to immune related traits which would make sense as we use LCL samples. But a better test should be done...



sQTL shared between populations



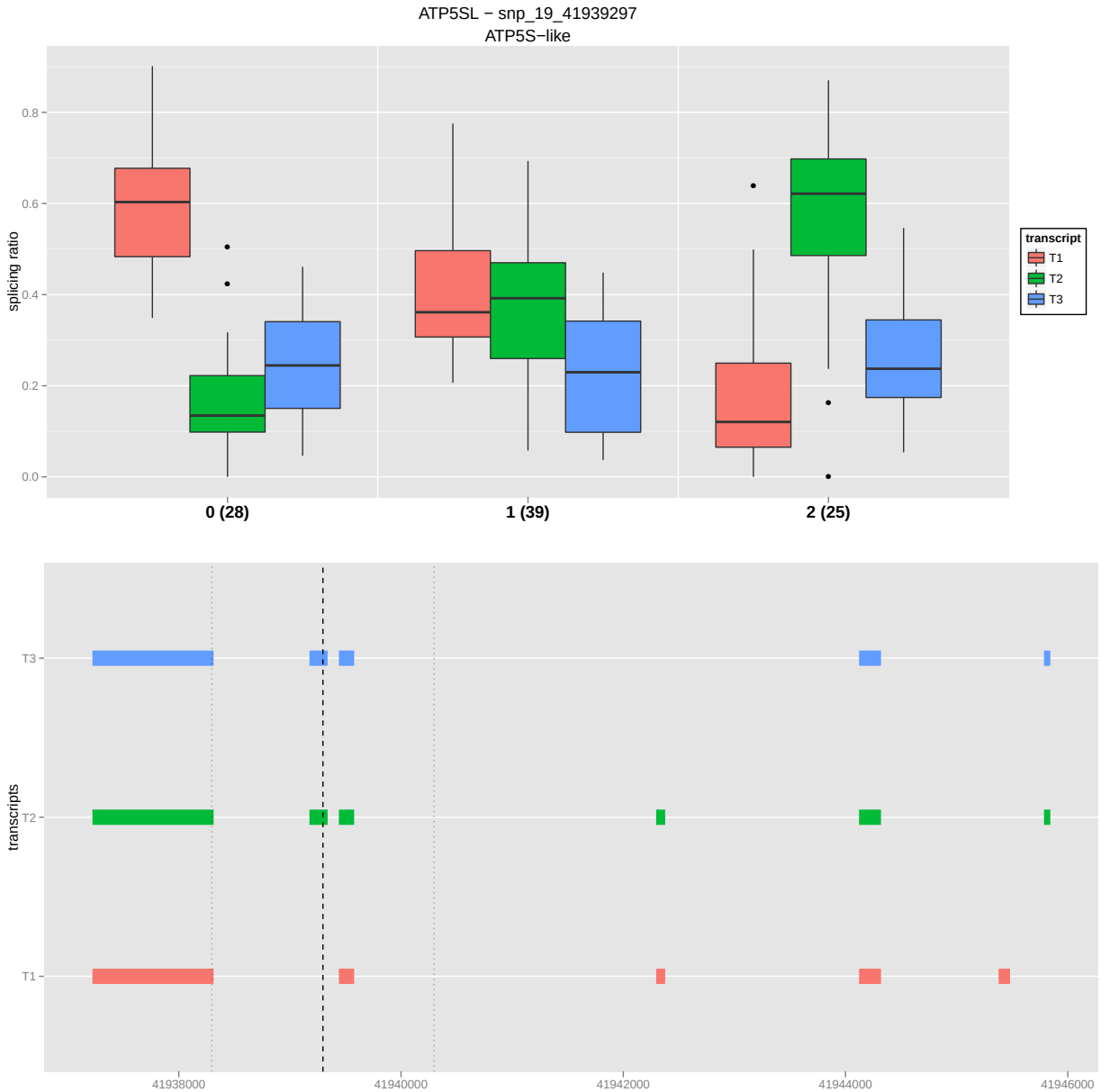
sQTL shared between populations



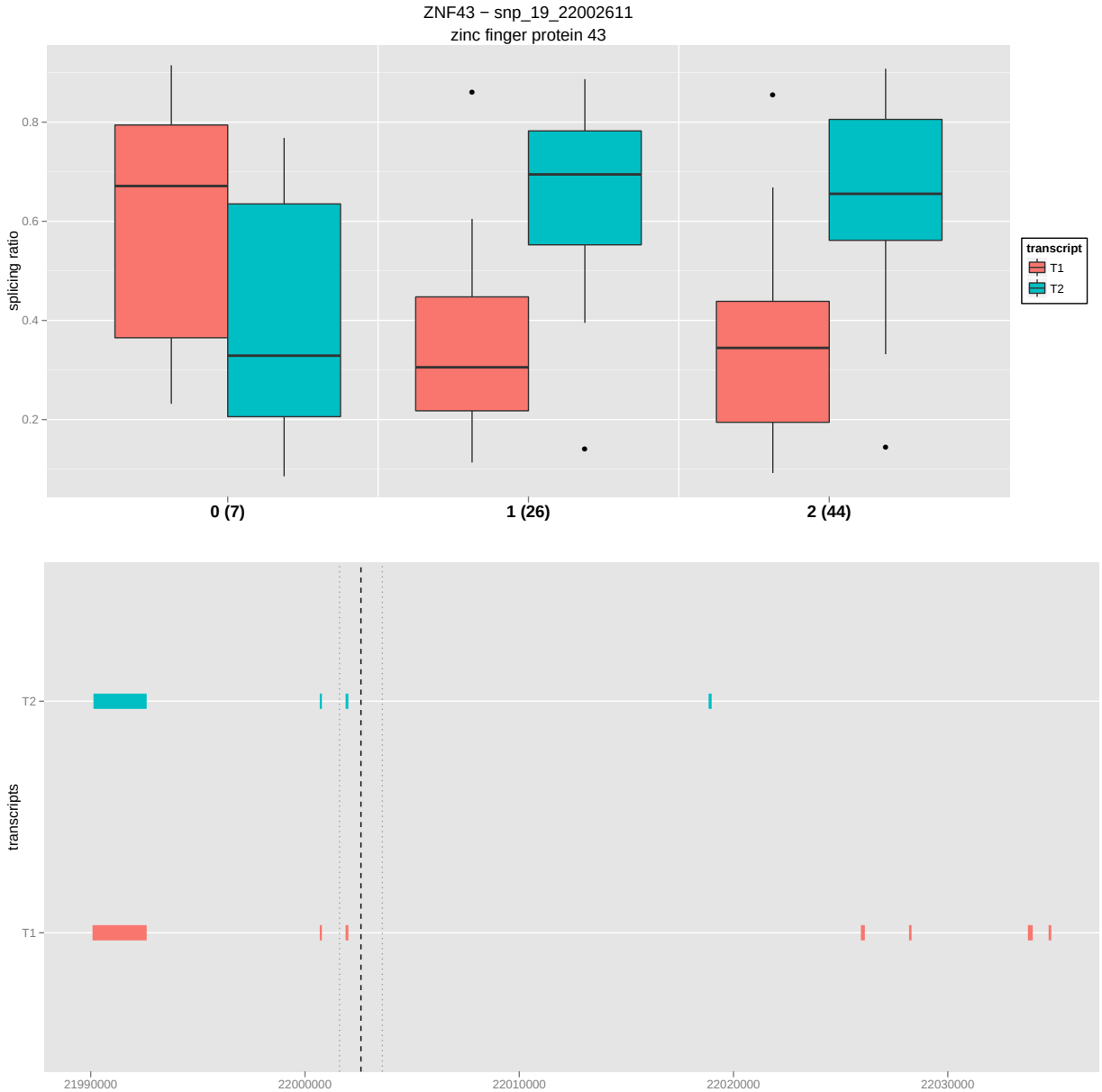
Now...

- sQTLs when pooling the European populations together.
- Compare with eQTL.
- Give a score to a sQTL to get the most interesting ones.
- Look at specific cases (good story hunting).
- ...

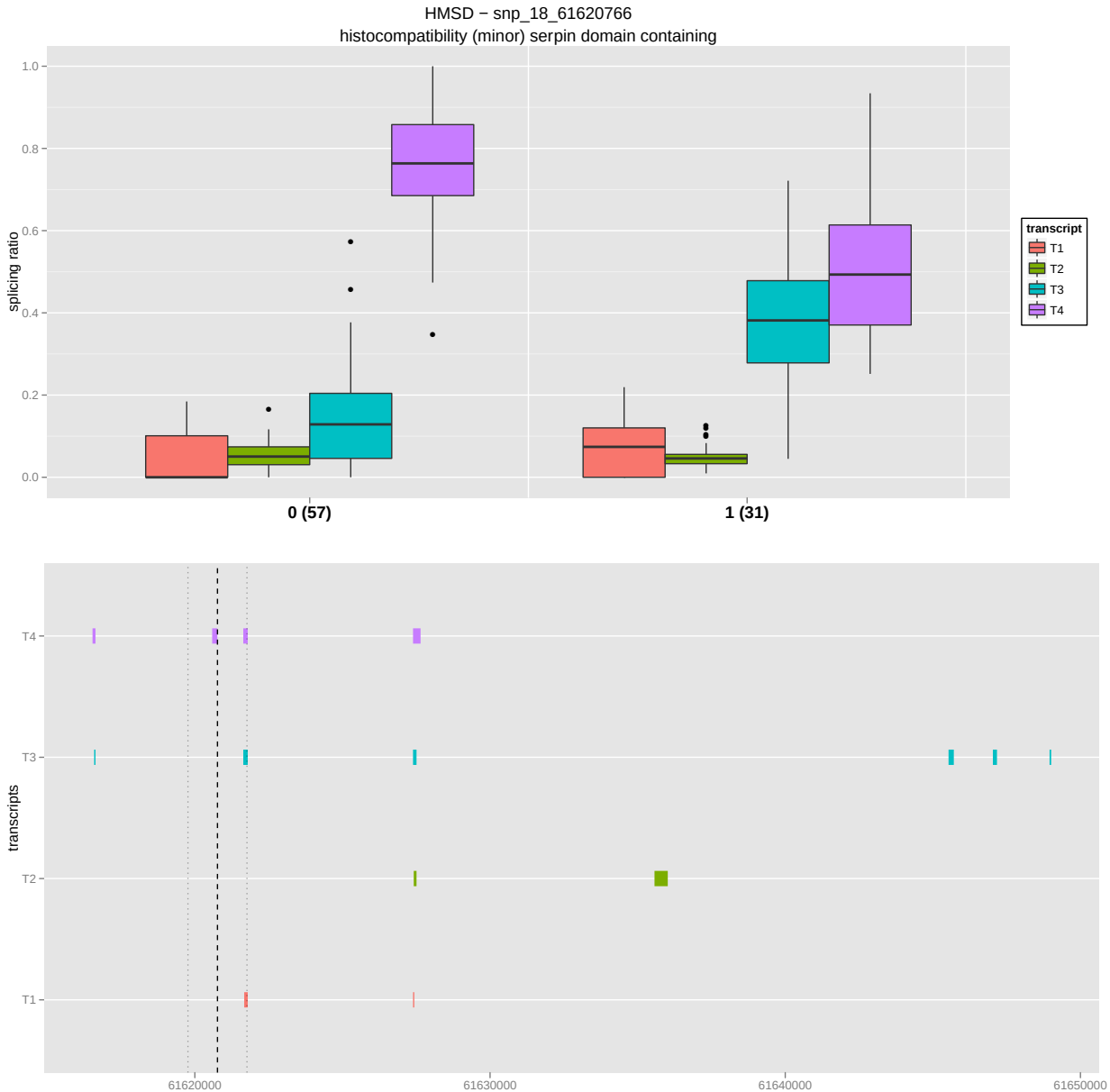
Examples - TSI



Examples - YRI



Examples - TSI



Transcript variation and alternative splicing

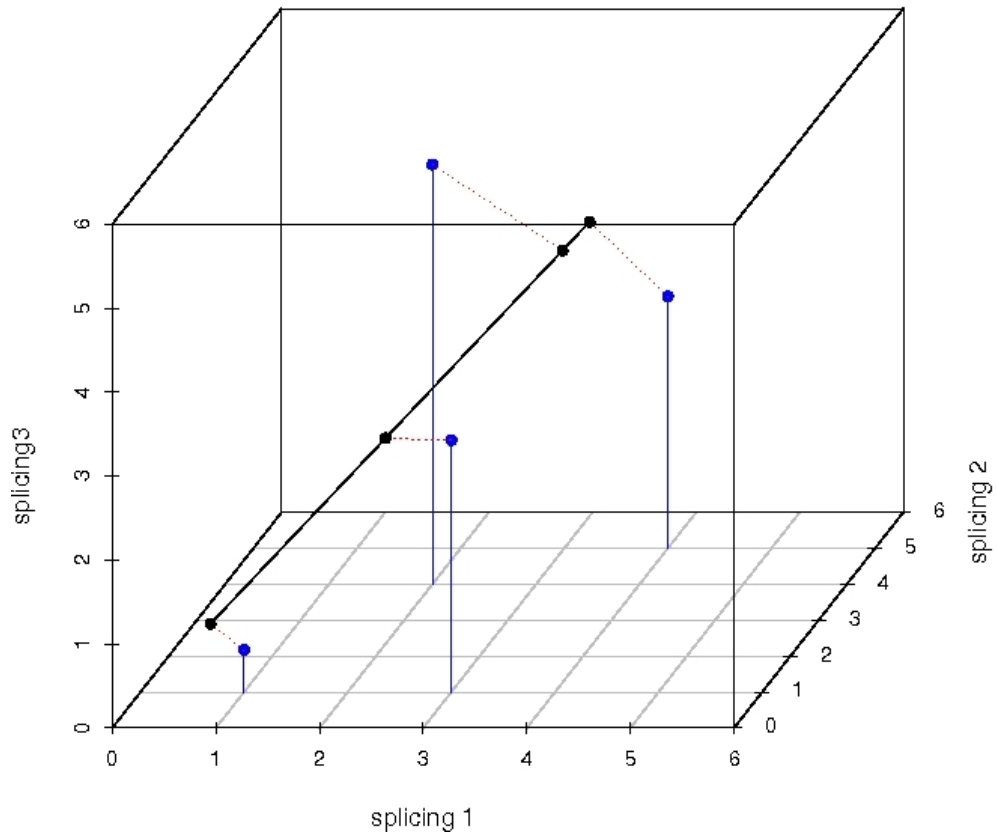
Questions

Transcript abundance variability

- What is the contribution of gene expression within a population ?
- What is the contribution of between-population variation ?
- What is the contribution of gene expression in the between-population variation contribution ?

Contribution of alternative splicing in the transcripts' abundance variability

Estimate the ratio $\frac{V_{ls}}{V_t}$ where V_{ls} is the variance when projecting the data to the model of constant splicing ratios and V_t the total variance.



Contribution of alternative splicing in the transcripts' abundance variability

- V_t: total observed transcript variation.
- V_{ls}: variation in the model of constant splicing.
- $V_{t,bw}$: between-population transcript variation.
- $V_{ls,bw}$: between-population transcript variation in the model of constant splicing.

- $\frac{V_{t,bw}}{V_t}$: contribution of between-population variation in the total variation.
- $\frac{V_{ls,bw}}{V_{t,bw}}$: contribution of gene expression in between-population variation.

Methods and summary of the results

Contribution of gene expression within a population

- Former method (*Vls/Vt*): ratio of the transcript variability under the model of "constant splicing" over the total observed variability.
- We see that around **50-55%** of the variation would be due to gene expression variation.

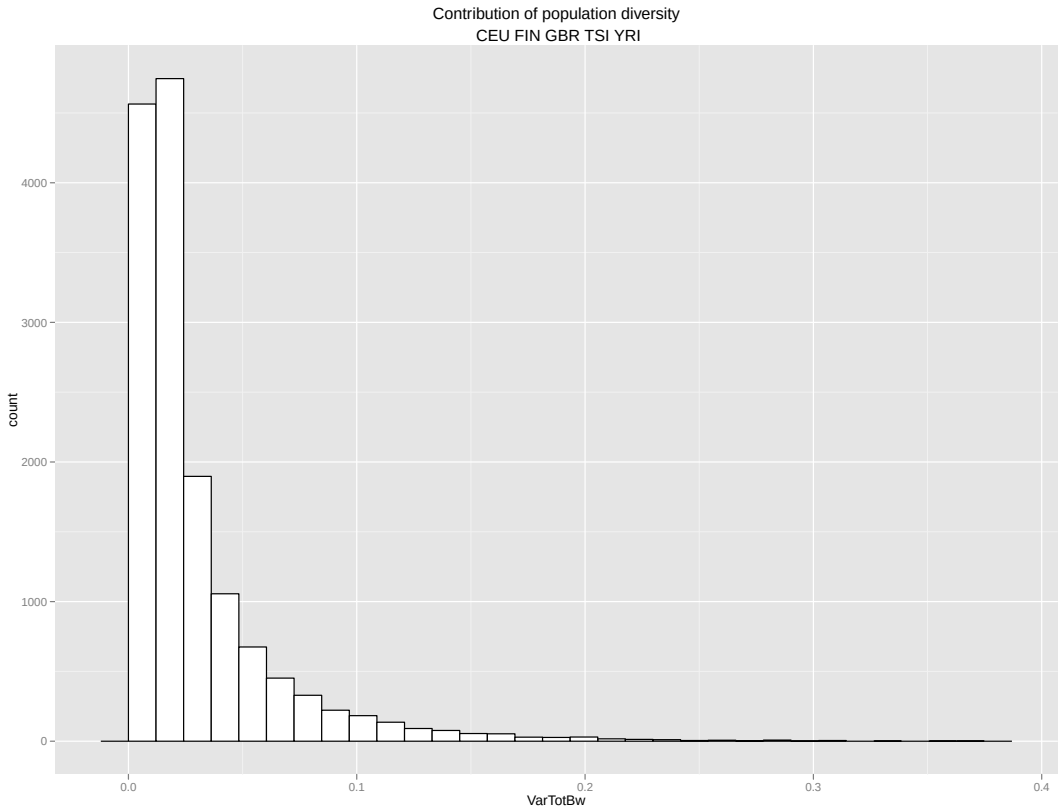
Contribution of between-population variation

- Ratio of the between-population variability over the total observed variability.
- We observe a very low population effect: **3%** of variation due to between-population variation.

Contribution of gene expression in population effect

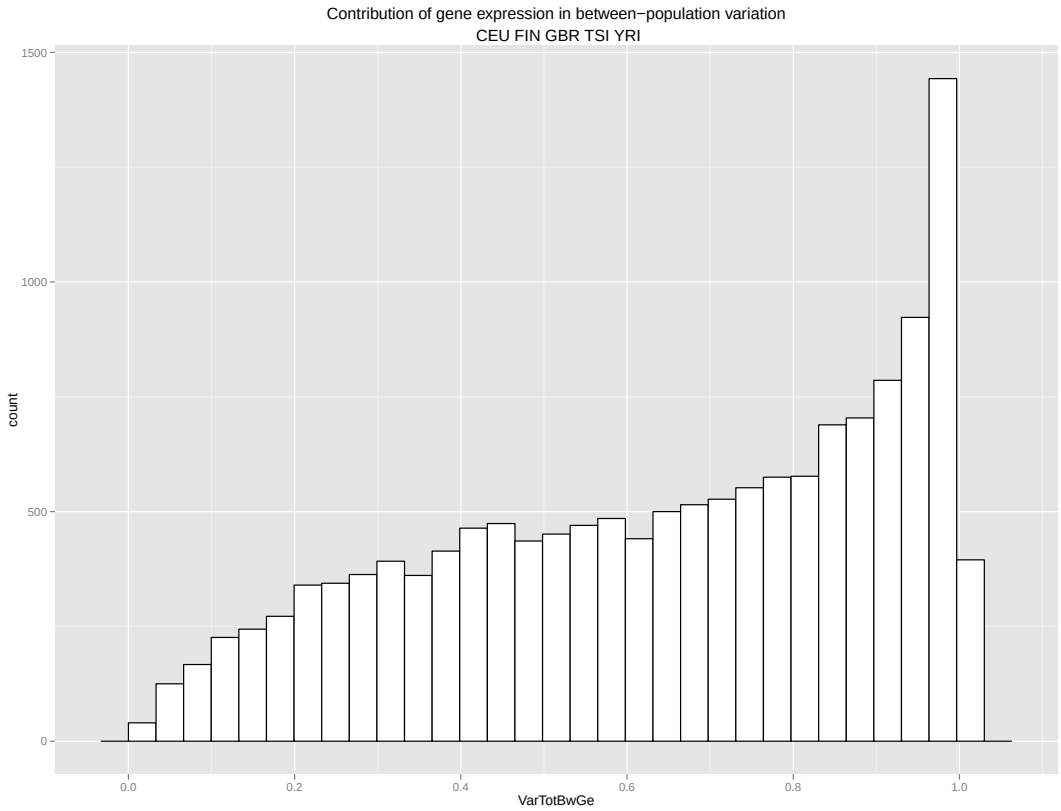
- Ratio of the between-population variability under the model of "constant splicing" over the between-population variability.
- We observe a higher contribution of gene expression: around **64%** of the between-population variation would be due to gene expression variation.

Pooling the five populations



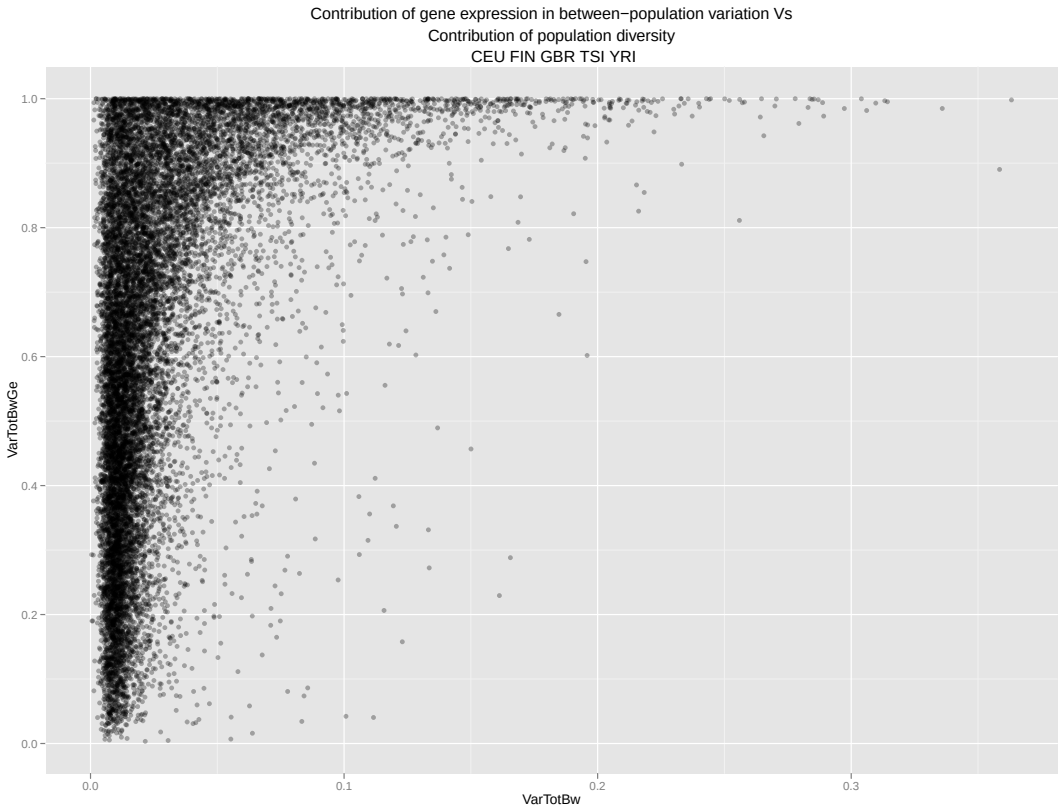
Not many genes seem to show a between-variation contributing at more than 10% to the total variation. *Each point is a gene. The x-axis represents the ratio of between-population variation over total variation.*

Pooling the five populations



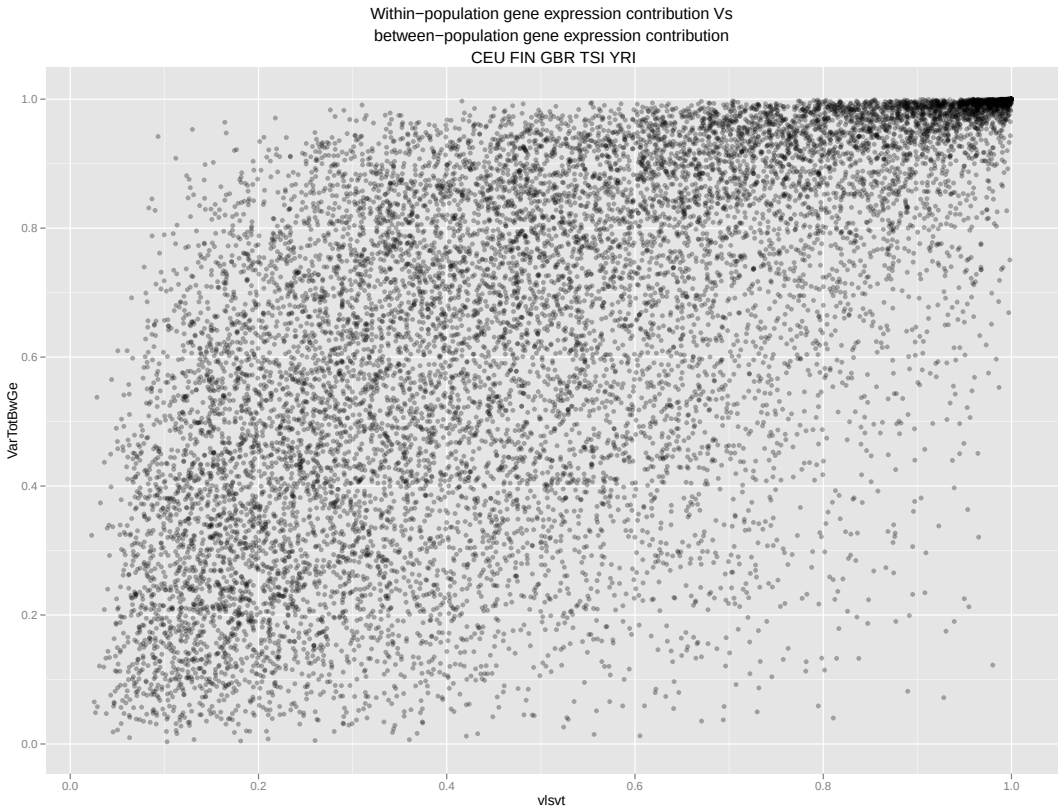
For many genes, gene expression seems to contribute almost completely to the between-population variation. *Each point is a gene. The x-axis represents the contribution of gene expression in the between-population variation.*

Pooling the five populations



Moreover when the splicing tends to contribute to btw-pop variation when it's itself lowly contributing to the total variation. *Each point is a gene. The x-axis represents the ratio of between-population variation over total variation. The y-axis represents the contribution of gene expression in the between-population variation.*

Pooling the five populations



The between-population variation seems to follow more or less the trend present in within-population variation. *Each point is a gene. The x-axis represents gene expression contribution in CEU samples total variation. The y-axis represents the contribution of gene expression in the between-population variation.*

Potential conclusion/interpretations

- The majority of the variation is present within each population.
- About half of this variation is due to gene expression.
- Among the small proportion of between population, gene expression mainly contributes.
- However, when splicing contributes it is in genes where it also have an important contribution in within population variation.