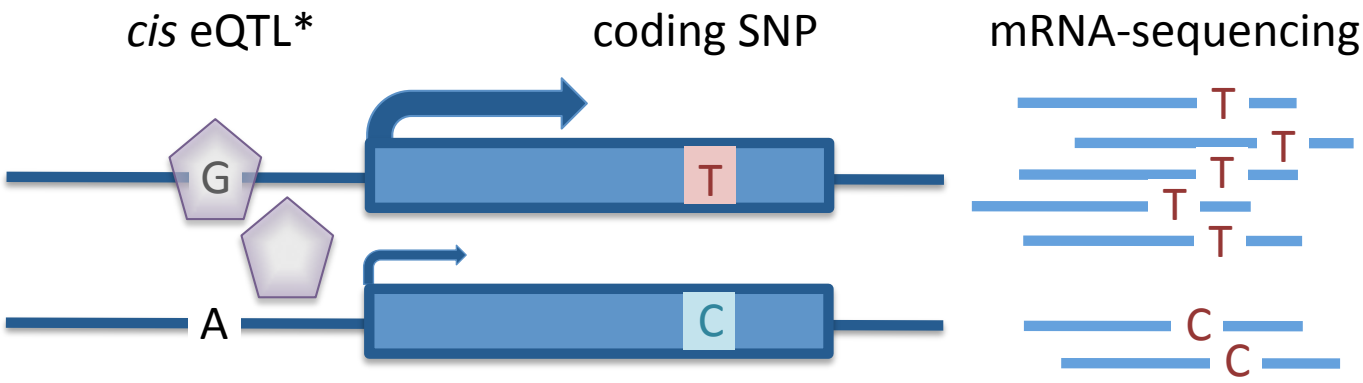
ASE analysis in Geuvadis

Tuuli Lappalainen
University of Geneva
May 24, 2012

The allele specific expression (ASE) approach

- Read counts over all heterozygous sites of an individual partitioned by the allele
    - Binomial test to detect deviation from the expected ~50/50 ratio
        - the expected is calculated from the overall reference/total ratio across the genome, partitioned by mapping quality and SNP alleles
        - filtering of sites with poor mapability and simulated evidence of allele-specific mapping bias
        - robust to confounding factors between individuals
- Regulatory variation in cis (or an epigenetic effect shared by all the cells)



ASE pipeline @ UNIGE

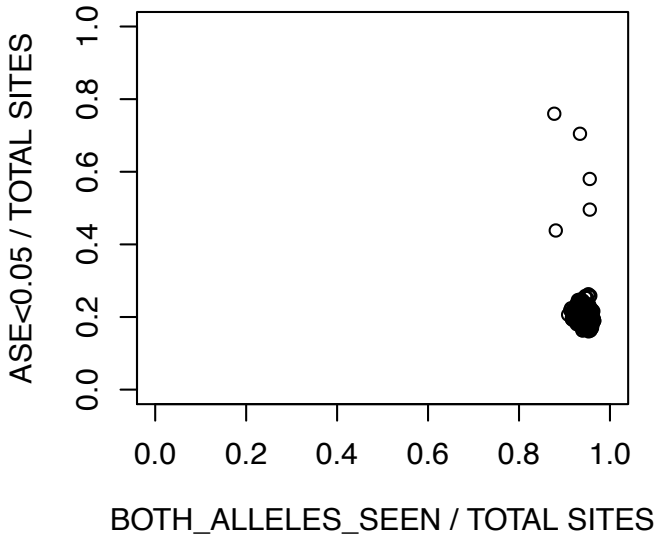
- Get the positions of all the variant sites in the entire study sample
  - Filter sites that are likely to have mapping error based on general mapability (UCSC track) and allelic mapping bias simulations
- For each individual, build pileups over these sites
- Parse the pileups to include only sites that are heterozygous in each individual and covered by >=16 reads, and count the alleles.
- Calculate a factor to correct for systematic bias in allelic ratios
  - calculate overall reference/total allele ratio for each individual for each SNP base combination
  - these ratios (generally only max 3% away from 0.5) are then used as the expected ratios.
- Calculate binomial probability for each covered site in each individual separately
- Build a results master file with annotations of the variants.

Caveats and concerns

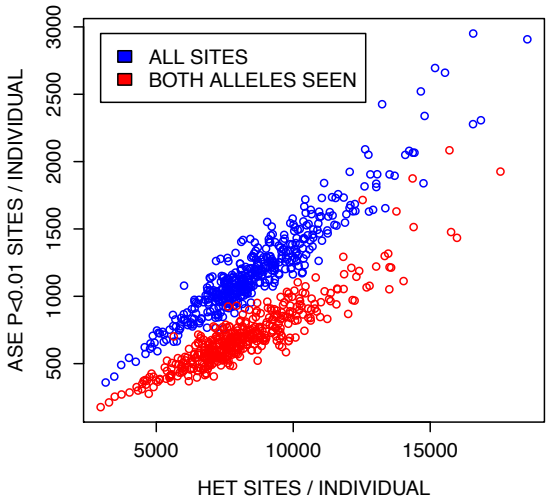
- Mapping bias = if your reference and alternative alleles don't map equally well (due to the variant itself or other flanking, linked variants)
  - I have developed our mapability filtering to relatively high sophistication
  - there isn't a filter that would get rid of all the bias
- Genotyping error
  - If you think that you're calculating ASE over a heterozygous site but the individual is actually homozygous, you'll have an extreme ASE signal that is completely false
  - We filter this by requiring to observe both alleles in RNAseq data (minor allele ratio >2%) by default. In some analysis this can be relaxed, but only with *a lot of caution*.
- P-value limit
  - <0.01 nominal p-value limit seems to work quite well, but it's not very stringent.
  - highly dependant on covegare and thus not really comparable between individuals or between sites
  - in many analyses I try to use continuous data of allelic ratios rather than the p-value.

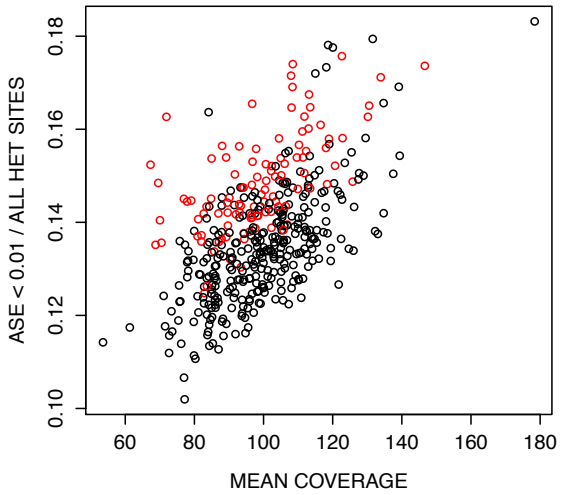
ASE as a QC tool

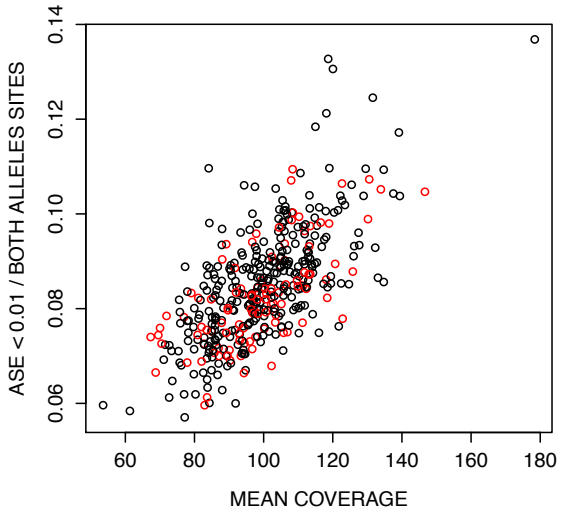
Cross-contamination in 5 samples



Differences in genotyping quality between individuals



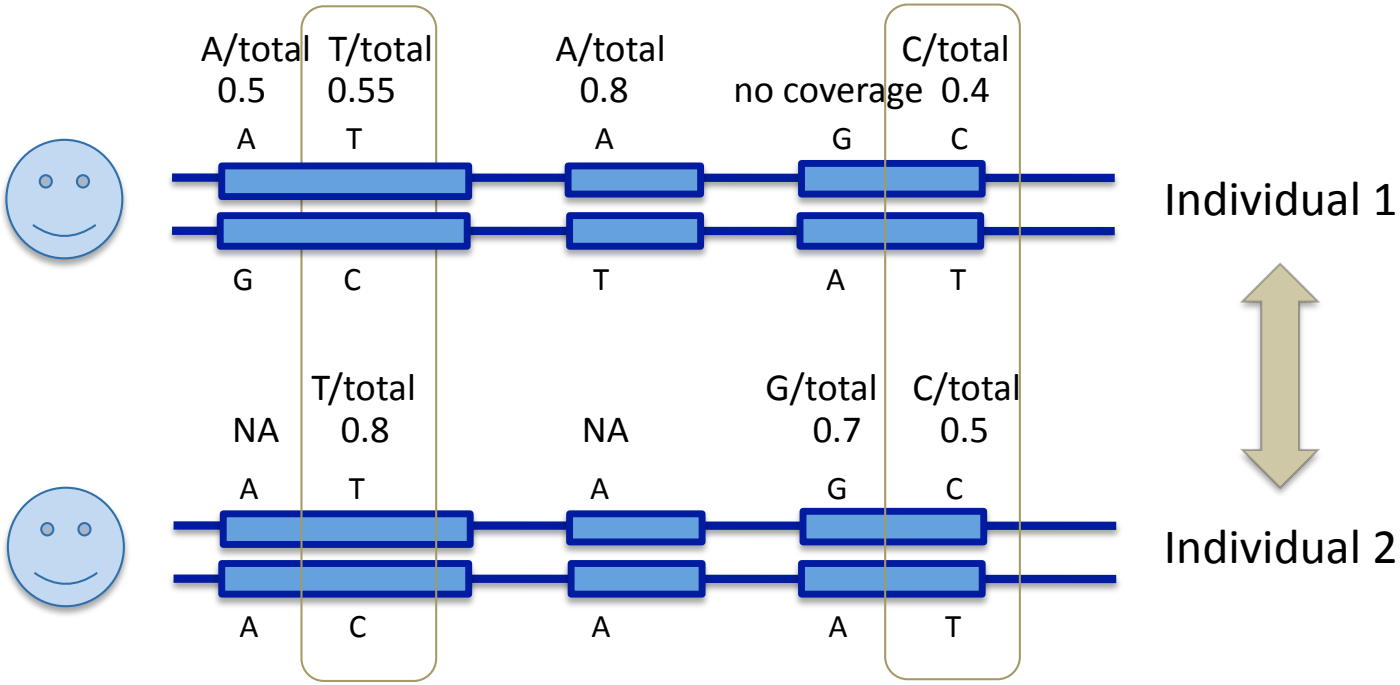




Geuvadis data: 462 full-coverage individuals

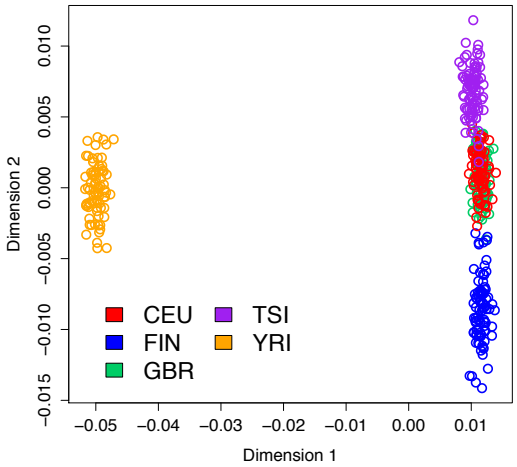
- Total number of heterozygous sites with enough coverage across individuals: 4,523,183
  - Unique SNPs: 200,133
  - Median per sample: 7,435
- Total number sites with ASE $p < 0.01$ across individuals 372,710
  - Unique SNPs: 59,022
  - Median per sample: 600 (7.79%)

Allelic expression distance between individuals

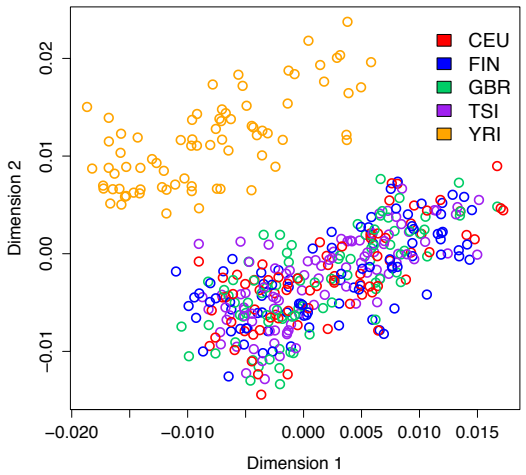


dist = median(c(abs(0.55-0.8), abs(0.4-0.5), …)

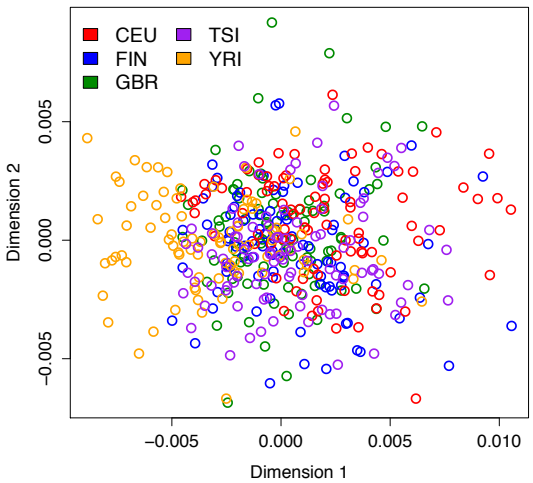
Genetic distance



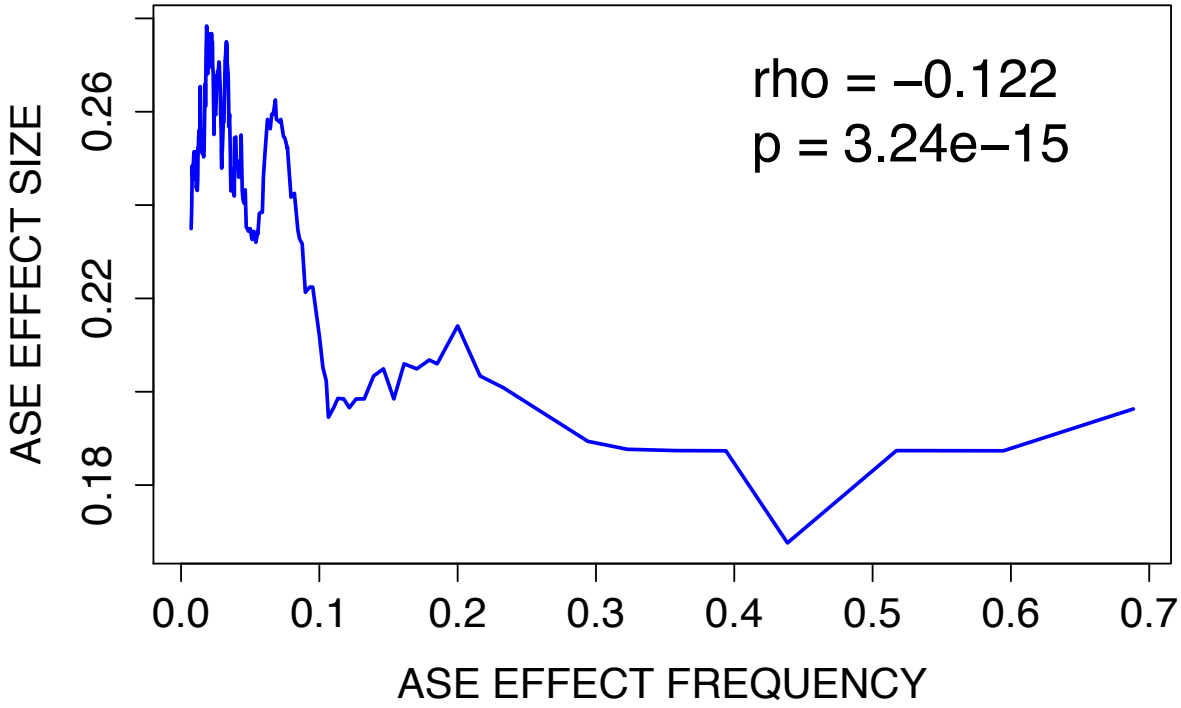
Allelic expression distance



Exon quantifications



Effect sites of regulatory events



- for each SNP, what's the proportion of individuals having significant ASE in this SNP? -> x-axis
- for each individual with ASE, calculate how far the allelic ratio is it from 0.5 = magnitude of the cis regulatory effect -> y-axis
- rare effects have bigger effect sizes

Partitioning the allelic expression differences between individuals