Genetic variation and mapping bias in exon quantifications

- Reference allele mapping bias = reads carrying the reference allele of variant sites map better than reads with alternative alleles
- "SNP in a probe", so not just noise – a potential source of false genetic associations!

Analysis

- Two individuals (HG00355, NA06986), of the sandbox; mapping with bwa to
  1. Standard hg19 reference
  2. Hg19 with all European 1000g SNPs with MAF>5% masked (changed to a nucleotide that is not one of the SNP alleles)
  3. Personalized Hg19 built with AlleleSeq tool based on the 1000g SNPs of these individuals
     - built paternal & maternal genomes, map to both separately, and combine the result files taking the best quality mapping
- Analyzing exon counts, normalized by the total number of reads
- To be done: ASE analysis

Mapping statistics

COUNT	bwa		
HG00355	**hg19**	**hg19_mask**	**hg19_pers**
TOTAL READS	71,987,632	71,987,632	71,987,538
MAPPED READS	38,537,502	38,480,142	45,051,066
EXONIC READS	31,817,668	31,752,122	36,708,566
LINKED READS	579,678	584,696	622,680

TOTAL READS	100.0%	100.0%	100.0%
MAPPED READS	53.5%	53.5%	62.6%
EXONIC READS	44.2%	44.1%	51.0%
LINKED READS	0.8%	0.8%	0.9%

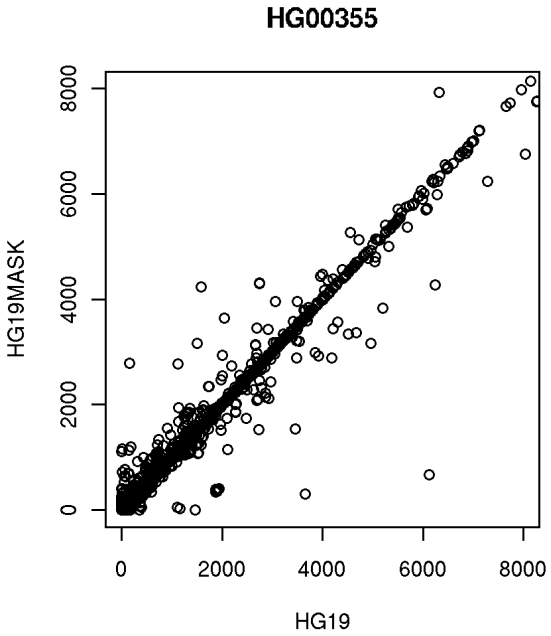
COUNT	bwa		
NA06986	**hg19**	**hg19_mask**	**hg19_pers**
TOTAL READS	61,003,566	60,814,748	61,003,472
MAPPED READS	31,580,394	31,482,522	38,206,888
EXONIC READS	26,762,074	26,670,238	31,794,952
LINKED READS	492,476	495,168	546,236

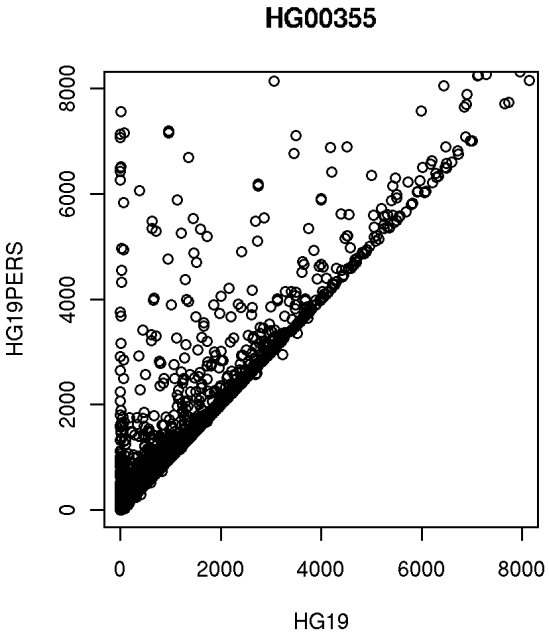
TOTAL READS	100.0%	100.0%	100.0%
MAPPED READS	51.8%	51.8%	62.6%
EXONIC READS	43.9%	43.9%	52.1%
LINKED READS	0.8%	0.8%	0.9%

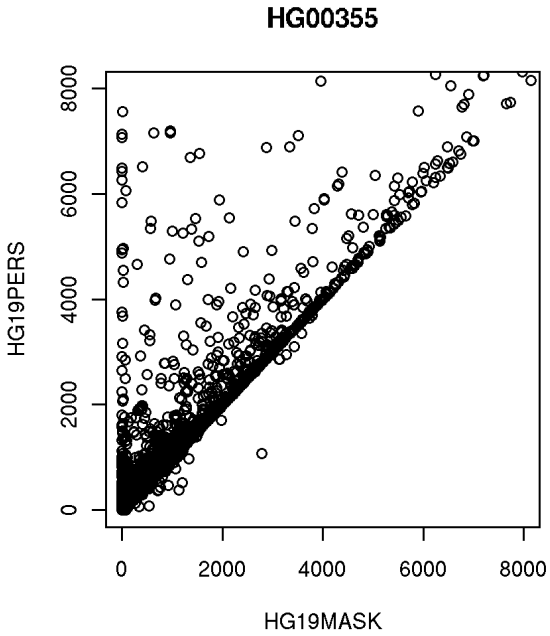
- Hardly any loss of reads in masking
  - need to try masking of all variant sites
- A lot more reads map to personalized genomes

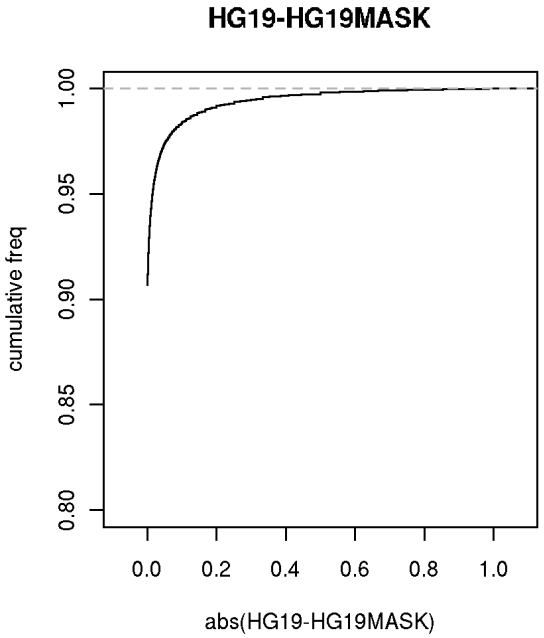
- 462,714 exons in the comparison

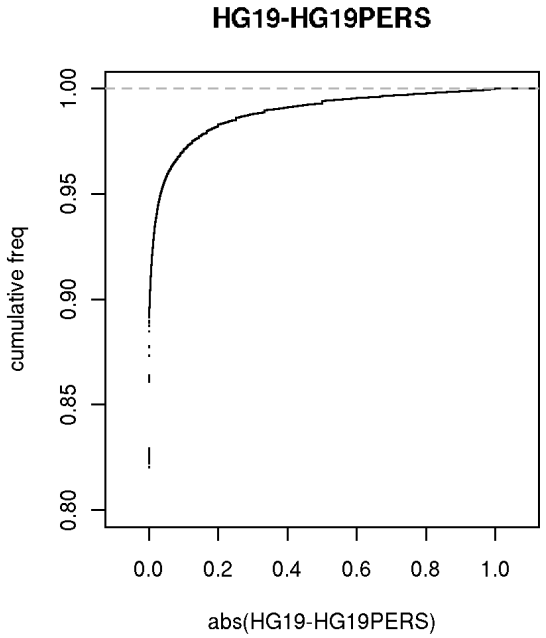
Differences between mapping approaches

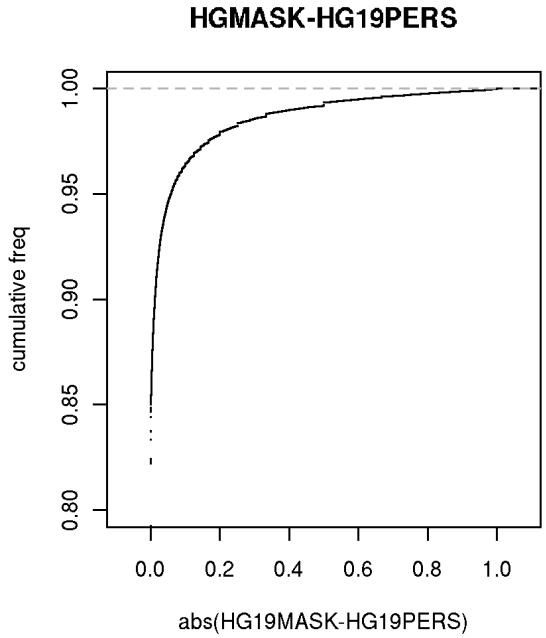




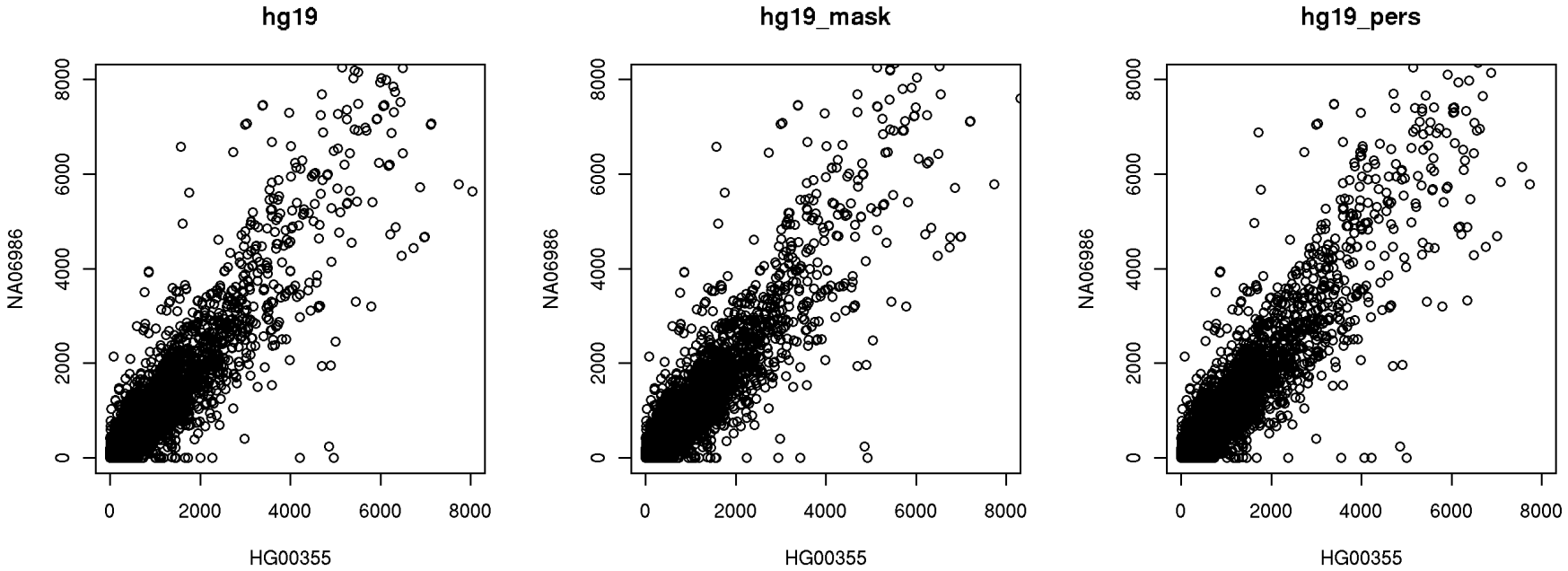








Mapping bias and differences between samples



rho=0.9066 rho=0.9068 rho=0.9076

…and some additional results suggesting that differences between individuals are smaller for personalized genome mapping.

Mapping biases are usually small relative to individual differences – but this is not necessarily true for those individual differences that associate to genetic variation

Conclusions

- Mapping bias has some effect on exon quantifications, and I suggest taking this into account in Geuvadis
- Personalized mapping would be ideal, but is more work and tricky for those 40 individuals that don't have a full genome
- Masking doesn't lead to a big loss of mapping – a straightforward option