Geuvadis RNAseq project
bwa *vs* GEM

Slides by Tuuli Lappalainen

Mapping and bam files by:
Ismael Padioleau & Tuuli Lappalainen (bwa)
Thasso Griebel, Paolo Ribeca, Micha Sammeth (GEM)

Analysis

- bwa: bwa-0.5.9 was run with default parameters except for sample -a 500000. bwa doesn't split reads. reference: autosomes + X + Y + M + EBV virus genome.
- GEM: …need to ask Thasso for details of the mapping. reference: autosomes + X + Y + M. Paolo did the conversion from GEM format to bam, with mapping qualities as follows:
  1) Matches which are unique, and do not have any subdominant match:
     251 >= MAPQ >= 255, XT=U
  2) Matches which are unique, and have subdominant matches but a different score:
     175 >= MAPQ >= 181, XT=U
  3) Matches which are putatively unique (not unique, but distinguishable by score):
     119 >= MAPQ >= 127, XT=U
  4) Matches which are a perfect tie:
     78 >= MAPQ >= 90, XT=R.

- exon quantifications in Geneva, using read counts over exons and merging exons of the same gene with overlapping coordinates. Can handle split reads (adds 1/ number_of_read_fragments to each exon)
- ASE analysis in Geneva to look at reference allele mapping bias
- GEM reads used in these tests: 1+2+3 or 1+2 of the categories above

Mapping stats

		GEM1+2+3 (MAPQ > 150)				GEM1+2+3 (MAPQ > 100)			
SAMPLE	TOTAL	MAPPED WELL, % of TOTAL		EXONIC, % of TOTAL		MAPPED WELL, % of TOTAL		MAPPED QC OK & EXONIC, % of TOTAL	
HG00117.1.M_120209_1	60,719,248	53,981,282	89%	40,688,410	67%	55,983,156	92%	41,464,462	68%
HG00355.1.M_120209_1	53,652,850	48,009,726	89%	35,706,428	67%	49,644,494	93%	36,352,484	68%
NA06986.1.M_120209_1	52,243,706	46,131,732	88%	34,670,526	66%	47,928,466	92%	35,384,330	68%
NA19095.1.M_111124_8	63,781,962	58,012,308	91%	46,804,228	73%	59,862,164	94%	47,596,036	75%
NA20527.1.M_111124_6	72,760,118	65,469,434	90%	49,672,692	68%	67,815,900	93%	50,586,008	70%

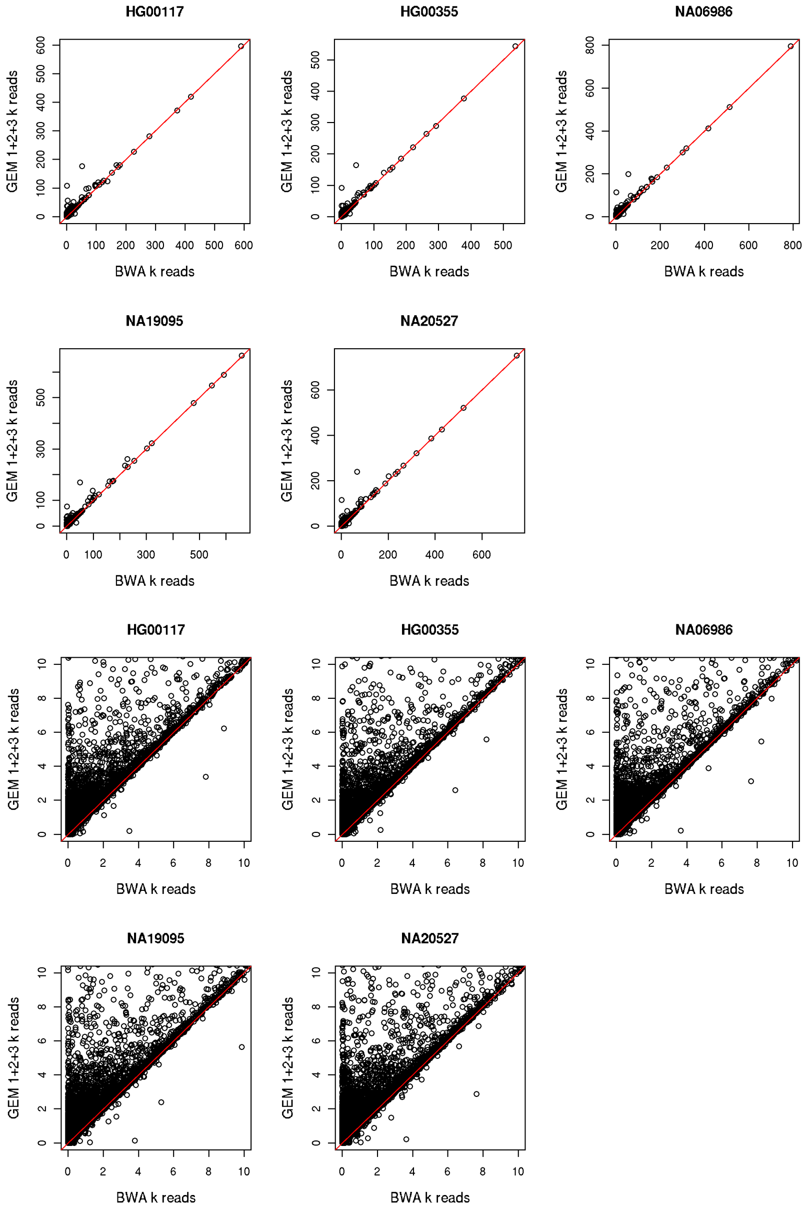
		BWA			
SAMPLE	TOTAL	MAPPED WELL, % of TOTAL		MAPPED QC OK & EXONIC, % of TOTAL	
HG00117.1.M_120209_1	60,719,248	41,366,247	68%	29,282,568	48%
HG00355.1.M_120209_1	53,652,850	36,076,837	67%	24,590,200	46%
NA06986.1.M_120209_1	52,243,706	35,227,333	67%	24,869,402	48%
NA19095.1.M_111124_8	63,781,962	43,328,203	68%	32,861,144	52%
NA20527.1.M_111124_6	72,760,118	49,122,531	68%	34,067,436	47%

Mapped well = properly paired and MAPQ >150 (GEM123), >100 (GEM12), and >10 (BWA)

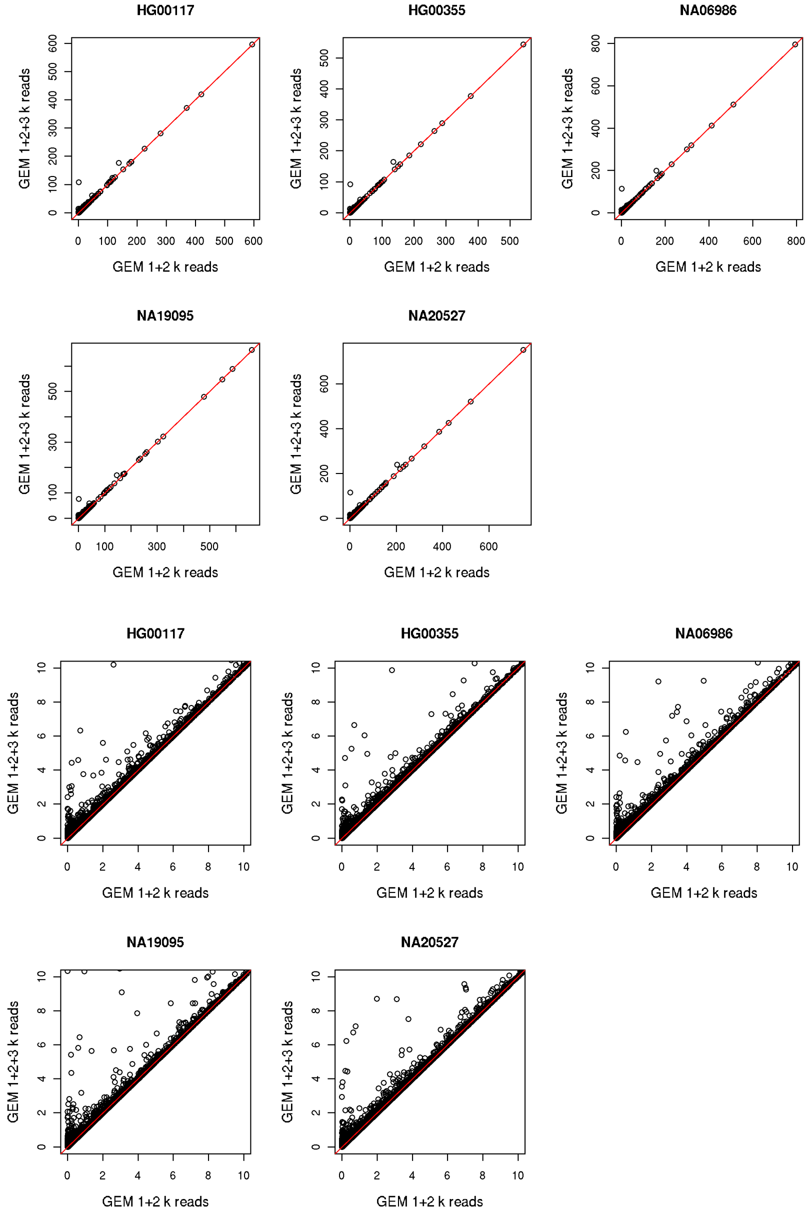
GEM maps a lot more reads.
Including or excluding the GEM category 3 reads doesn’t make a big difference

Exon quantification comparisons

bwa/GEM123



GEM12/GEM123



Exon quantification comparisons

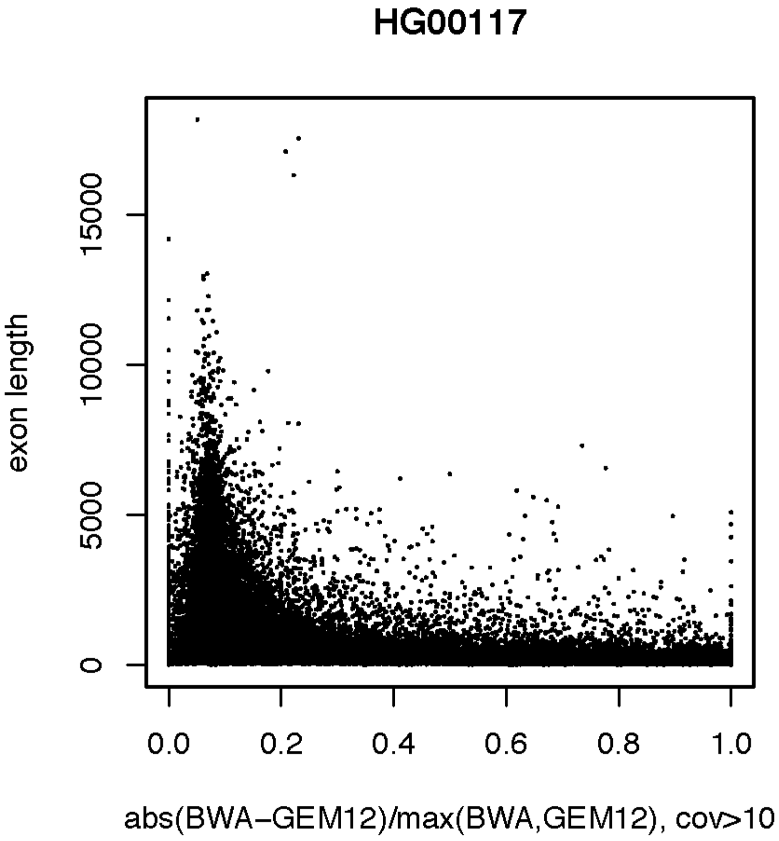
Spearman rho between mappers

	BWA-GEM123	BWA-GEM12	GEM123-GEM12
HG00117	0.8386	0.8466	0.9923
HG00355	0.8264	0.8338	0.9928
NA06986	0.8294	0.8368	0.9924
NA19095	0.8265	0.8335	0.9934
NA20527	0.8251	0.8328	0.9927

Spearman rho between samples for each mapper (each sample vs all others -> median)

	BWA	GEM12	GEM123
HG00117	0.9221	0.9541	0.9541
HG00355	0.9200	0.9558	0.9515
NA06986	0.9062	0.9409	0.9405
NA19095	0.9213	0.9523	0.9497
NA20527	0.9277	0.9634	0.9608

GEM-bwa differences are much bigger in short exons, as expected



Reference allele bias in allele specific expression analysis

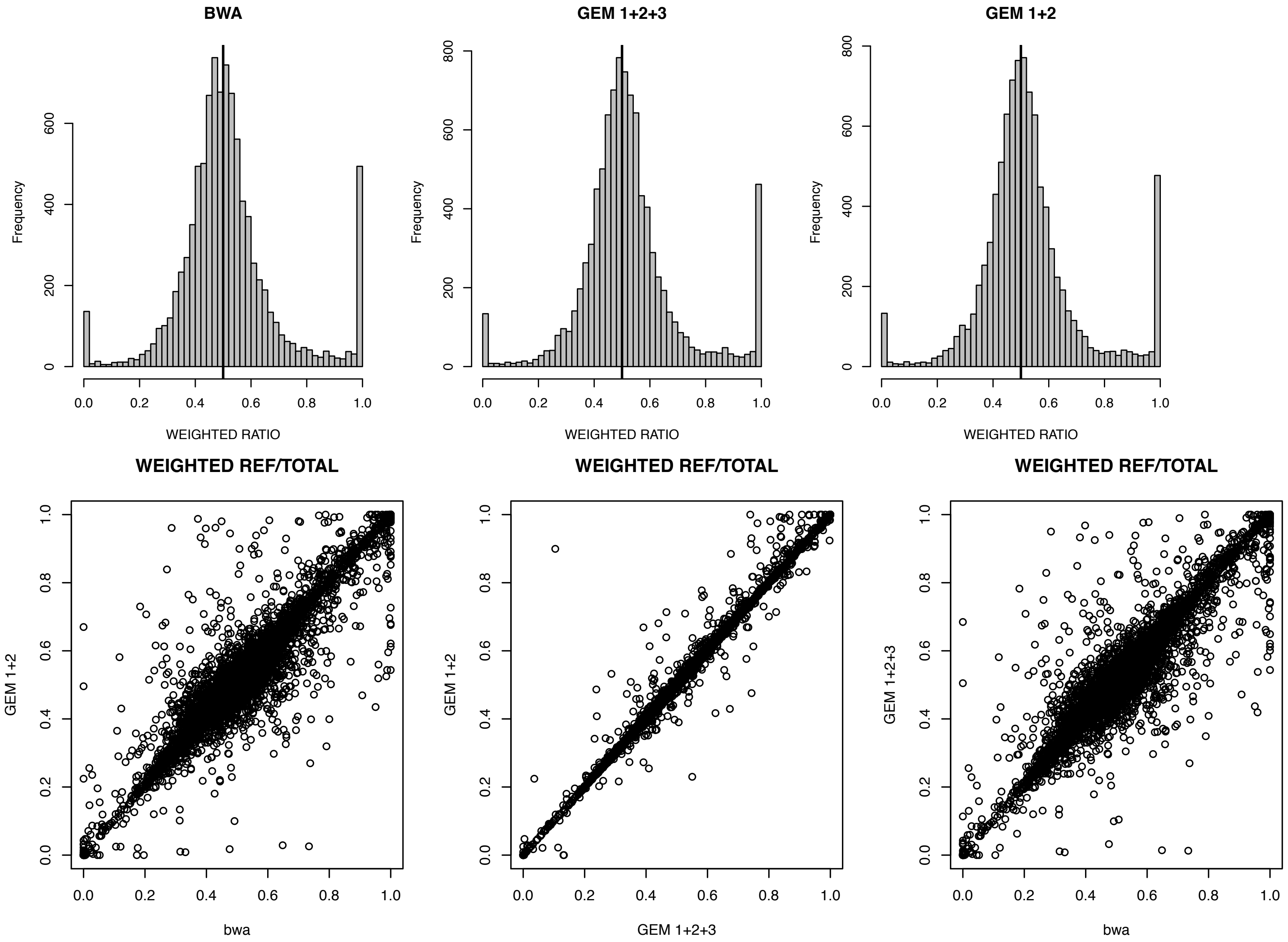
- Analysis: An individual’s RNAseq reference & nonreference read counts over heterozygote sites (from genotype data)

HG00117	BWA	GEM123	GEM12
N_HET	9811	11882	11661
N_HET_BAS	8988	10696	10592
N_ASE_01	1687	2542	2346
N_ASE_BAS_01	864	1356	1277
MEDIAN_COV	39	41	41
MEAN_COV	132.308	144.501	141.079
MEAN_REFRATIO	0.545	0.543	0.541
MEAN_WEIGHT_REFRATIO	0.522	0.528	0.526
MEAN_REFRATIO_BAS	0.525	0.524	0.524
MEAN_WEIGHT_REFRATIO_BAS	0.500	0.507	0.507

HET = heterozygous sites with coverage >15 (=OK for ASE analysis)
BAS = both alleles seen in RNAseq data (verifies the genotype and filters for some other problematic sites)
COV = coverage

GEM detects more sites due to higher coverage; otherwise the statistics look similar. Reference allele mapping bias is similar

Reference allele bias in allele specific expression analysis (HG00117)

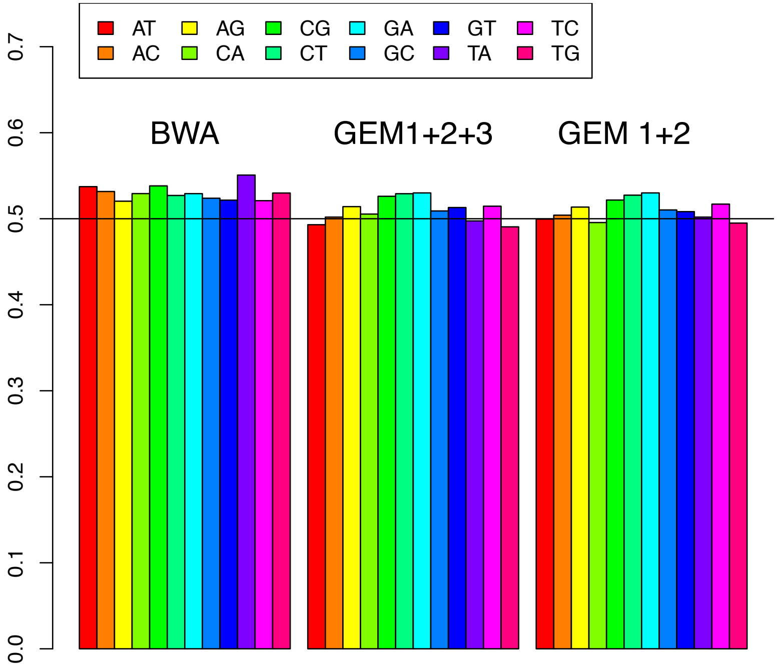


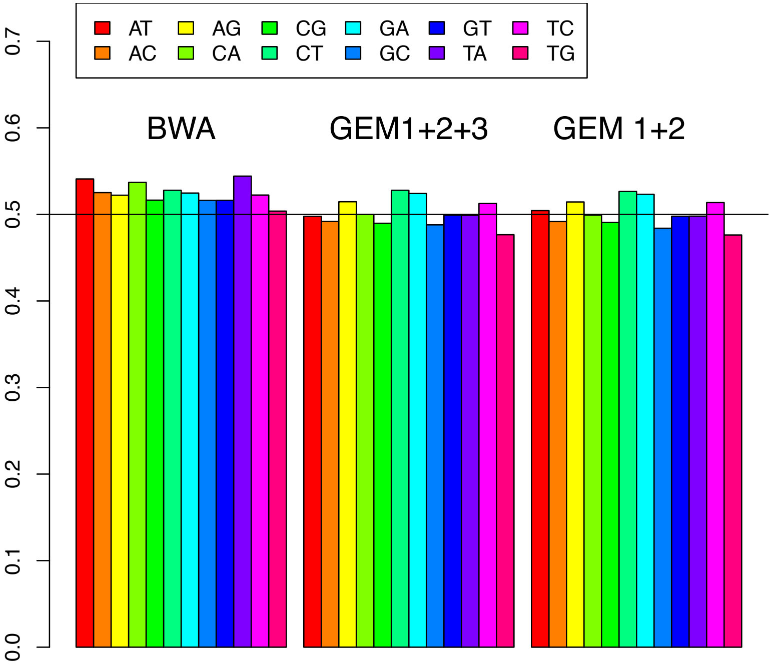
No systematic bias between methods. Deviations probably mostly random fluctuation

Reference allele bias in allele specific expression analysis

For each individual and each SNP base combination, we calculate the genome-wide REF/TOTAL ratio. This is used to correct for genome-wide average reference allele mapping bias, and can be used as a metric of the extent of deviation.

Below are the ratios for two individuals (HG00117 and HG00355). No systematic differences.


AT
AC
AG
CA
CG
CT
GA
GC
GT
TA
TC
TG
BWA
GEM1+2+3
GEM 1+2


AT
AC
AG
CA
CG
CT
GA
GC
GT
TA
TC
TG
BWA
GEM1+2+3
GEM 1+2