A Supplementary figures and tables for *SNP calling using genotype model selection on high-throughput sequencing data*

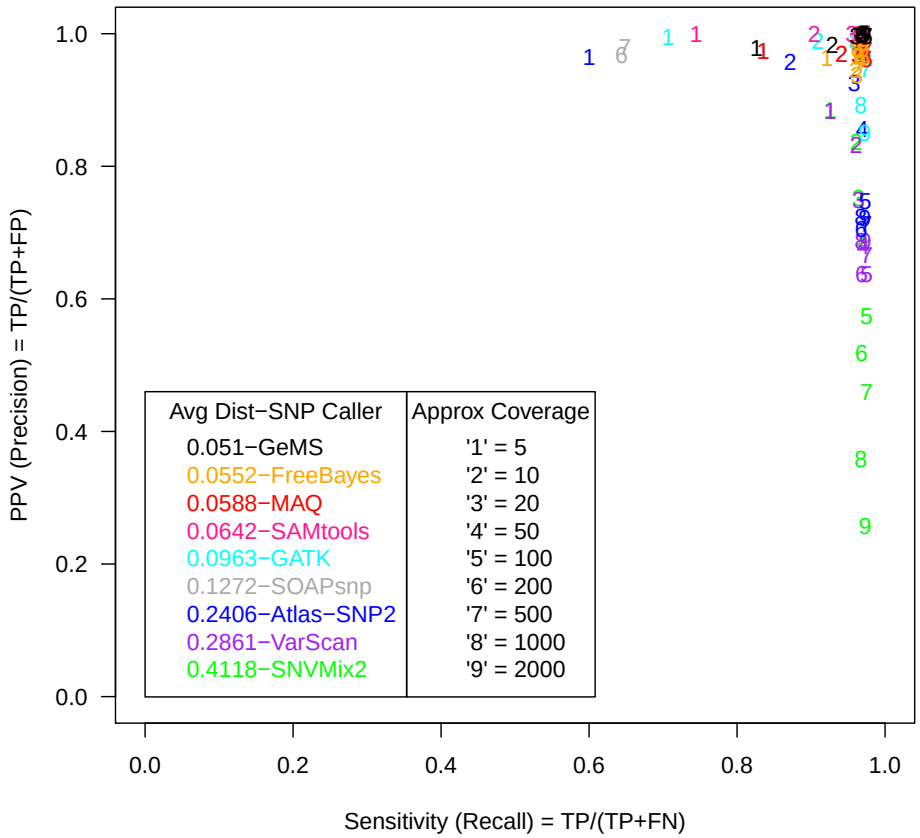

Figure 1: **Sensitivity and PPV plot of SNP caller performance.** SNP caller performance across coverage levels and SNP callers. See figure 2 for zoomed-in view.

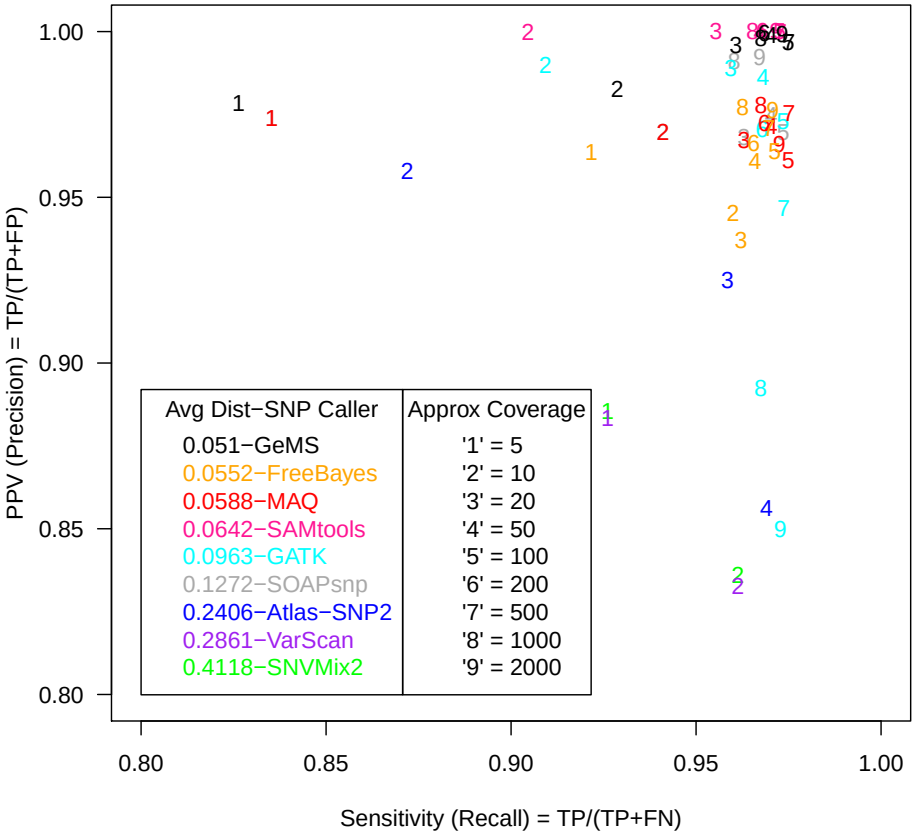

Figure 2: **Zoomed-in view of sensitivity and PPV plot of SNP caller performance.** SNP caller performance across coverage levels and SNP callers.

Simulation Study	Specifications and non-default options used[a]
Reference organism	Thermoanaerobacter sp. X514 (2.5Mbp)
Simulation package	MAQ-0.7.1 `simulate` and `simutrain` based on the X514 Xw2010 dataset in table 4 below
Simulation options	`-N` was set to simulate average coverage values of 5, 10, 20, 50, 100, 500, 1000 and 2000
Sample size	Single sample
Ploidy	X514 is haploid but MAQ `simulate` was implemented in its default diploid mode
Read Length	40bp
Mate pair	Paired-end
Quality encoding	Illumina 1.6+
Alignment options	BWA-0.5.9 `aln -I`, `sampe` and SAMtools-0.1.16 `view -q 20`[b]
Pileup options	SAMtools-0.1.16 `view -q 20 | pileup -Q 17 -B -s -`
GeMS options	$\alpha = 0.05$
MAQ options	SAMtools-0.1.9 `pileup -vc`
FreeBayes options	FreeBayes-0.9.0 `-4 -m 20 -q 17 -R 20,17`
SAMtools options	SAMtools-0.1.16 `mpileup -Q 17 -g` and `bcftools view -vcg`
GATK options	GenomeAnalysisTK-1.0.5336 `-T UnifiedGenotyper -bad_mates -mbq 17 -mmq 20`
SOAPsnp options	SAMtools-0.1.9 `pileup -avc`
Atlas-SNP2 options	Atlas-SNP2-1.1 `-s -f 1000000`
VarScan options	VarScan-v2.2.5 `-min-coverage 1 -min-reads2 1 -min-avg-qual 17`
SNVMix2 options	SNVMix2-0.11.8 `-m shah_lobular_snvmix_model.txt`[c] `-t MB -q 16 -Q 19`

Table 1: **Simulation data specification and alignment and SNP calling options.** To a reasonable extent, the options of the SNP callers were chosen to maximize the fairness of the comparison. Default options were preferred unless the default options were not applicable or gave an unfair advantage to any particular SNP caller.

[a] A note on post SNP calling filters: in theory, any post SNP calling filter, based on information within the alignment file (in this case BWA's SAM output), should be able to be applied to any SNP caller's variant results. Thus, since we were trying to assess the SNP callers and not the filters, all efforts were made to eliminate optional post SNP call filtering.

[b] Using SAMtools `view` is just one way of refining the alignment file in preparation for downstream analyses such as SNP calling. For more information on using SAMtools `view` to pre-filter alignment files, please see the PDF document "Pre-Filtering Alignment Files" available at `https://sites.google.com/a/bioinformatics.ucr.edu/xinping-cui/home/software/ngs-snp-calling`.

[c] File is available at `http://compbio.bccrc.ca/wp-content/uploads/2009/10/shah_lobular_snvmix_model.txt`.

SNP Caller	Sens	PPV	Joint Avg
GeMS	**0.9495**	**0.9940**	**0.9718**
SAMtools	0.9358	1.0000	0.9679
FreeBayes	0.9610	0.9625	0.9617
MAQ	0.9520	0.9706	0.9613
GATK	0.9333	0.9548	0.9441
SOAPsnp	0.8781	0.9763	0.9272
Atlas-SNP2	0.9173	0.8128	0.8650
VarScan	0.9646	0.7179	0.8412
SNVMix2	0.9646	0.5919	0.7783

Table 2: Average of sensitivity, PPV and joint average of sensitivity and PPV over all coverage levels, listed in descending order of joint average.

Sensitivity	5	10	20	50	100	200	500	1000	2000
10 Model GeMS	0.8264	0.9287	0.9607	0.9704	0.9749	0.9684	0.9751	0.9676	0.9733
3 Model GeMS	0.3767	0.7536	0.9438	0.9668	0.9735	0.9679	0.9737	0.9671	0.9725
SAMtools	0.7449	0.9045	0.9554	0.9672	0.9726	0.9679	0.9728	0.9653	0.9716

PPV	5	10	20	50	100	200	500	1000	2000
10 Model GeMS	0.9784	0.9828	0.9958	0.9991	0.9967	0.9995	0.9968	0.9981	0.9991
3 Model GeMS	0.9976	0.9948	0.9981	0.9995	0.9967	0.9995	0.9991	0.9991	0.9991
SAMtools	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3: **Comparing 3 genotype models to 10 in the GeMS procedure.** For each table, the first row indicates the average simulated coverage values. These results indicate that given the otherwise unchanged GeMS procedure, considering only 3 models reduces the sensitivity at every coverage level when compared to considering the 10 possible genotype models.

Arabidopsis sup1ros1	Specifications and non-default options used
Organism Reference	Arabidopsis thaliana Columbia TAIR9 (121Mbp)
Sample size	Single sample
Ploidy	Diploid
Read Length	110bp
Mate pair	Single-end
Quality encoding	Illumina 1.5+
Alignment options	BWA-0.5.9 `aln -I`, `samse` and SAMtools-0.1.16 `view -q 20`
Pileup options	SAMtools-0.1.16 `view -q 20 | pileup -Q 17 -B -s -`
GeMS options	GATK priors, $\alpha = 0.005$
GATK options	GenomeAnalysisTK-1.0.5336 `-T UnifiedGenotyper -bad_mates -mbq 17 -mmq 20`
SAMtools options	SAMtools-0.1.16 `mpileup -Q 17 -g` and `bcftools view -vcg`

X514 Xw2010	Specifications and non-default options used
Organism Reference	Thermoanaerobacter sp. X514 (2.5Mbp)
Sample size	Single sample
Ploidy	Haploid
Read Length	40bp
Mate pair	Single-end
Quality encoding	Illumina 1.6+
Alignment options	BWA-0.5.9 `aln -I`, `samse` and SAMtools-0.1.13 `view -q 20`
Pileup options	SAMtools-0.1.16 `view -q 20 | pileup -Q 17 -B -s -`
GeMS options	$\alpha = 0.05$
FreeBayes options	FreeBayes-0.9.0 `-4 -p 1 -m 20 -q 17 -R 20,17`

Table 4: **Real dataset specifications and alignment and SNP calling options.** To a reasonable extent, the options of the SNP callers were chosen to maximize the fairness of the comparison. Default options were preferred unless the default options were not applicable or gave an unfair advantage to any particular SNP caller.

Time (Sec)	5	10	20	50	100	200	500	1000	2000	Caller Avg
SNVMix2	2	3	4	9	15	29	77	176	346	74
SOAPsnp	2	3	6	13	25	52	131	265	548	116
MAQ	4	5	7	14	26	54	137	268	562	120
GeMS	**12**	**17**	**28**	**60**	**112**	**212**	**339**	**512**	**854**	**239**
GATK	33	41	53	99	154	264	495	727	1208	342
FreeBayes	16	26	45	99	196	381	916	1828	3639	794
SAMtools	24	43	81	192	383	756	1872	3729	7444	1614
VarScan	572	687	929	1597	2728	5003	NA	NA	NA	1919
Atlas-SNP2	236	336	510	1029	1954	4041	13669	55554	267224	38284
Coverage Avg	100	129	185	346	621	1199	2204	7882	35228	4833

Max Memory (MB)	5	10	20	50	100	200	500	1000	2000	Caller Avg
SOAPsnp	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*
SNVMix2	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*
FreeBayes	0*	0*	0*	0*	0*	0*	16	16	16	5
MAQ	32	32	32	32	32	32	32	32	32	32
SAMtools	32	32	32	32	32	32	32	32	32	32
VarScan	48	48	48	48	48	48	NA	NA	NA	48
GeMS	**336**	**384**	**368**	**384**	**384**	**400**	**400**	**400**	**400**	**384**
GATK	304	384	256	400	1392	304	480	320	528	485
Atlas-SNP2	176	192	208	256	320	464	848	1504	3136	789
Coverage Avg	103	119	105	128	245	142	226	288	518	197

Table 5: **SNP caller time to completion in seconds and maximum memory usage in megabytes (MB)**. The first row indicates the simulated coverage values. The bottom row gives the column averages at each coverage level. The remaining rows are sorted by the "Caller Avg" column, which gives the average of the rows over all the coverage levels. Given the stated computing setup, the VarScan procedure was not able to be completed for coverage levels 500, 1000 and 2000. Procedures notated with "0*" used less than 16MB of memory.

Avg Memory (MB)	5	10	20	50	100	200	500	1000	2000	Caller Avg
SOAPsnp	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*
SNVMix2	0*	0*	0*	0*	0*	0*	0*	0*	0*	0*
FreeBayes	0*	0*	0*	0*	0*	0*	5	13	16	4
MAQ	32	32	32	32	32	32	32	32	32	32
SAMtools	32	32	32	32	32	32	32	32	32	32
VarScan	48	48	48	48	48	48	NA	NA	NA	48
GeMS	**163**	**232**	**210**	**205**	**201**	**200**	**200**	**203**	**203**	**202**
GATK	262	338	176	368	1209	220	258	234	321	376
Atlas-SNP2	104	99	105	145	182	275	593	1098	2444	561
Coverage Avg	71	87	67	92	189	90	140	202	381	139

Table 6: **Average memory usage in megabytes (MB).** The first row indicates the simulated coverage values. The bottom row gives the column averages at each coverage level. The remaining rows are sorted by the "Caller Avg" column, which gives the average of the rows over all the coverage levels. Given the stated computing setup, the VarScan procedure was not able to be completed for coverage levels 500, 1000 and 2000. Procedures notated with "0*" used less than 16MB of memory.