Supplementary materials: Reproducibility of SNV-calling in multiple sequencing runs from single tumors

Dakota Z. Derryberry*[1], Matthew C. Cowperthwaite[2,3], and Claus O. Wilke[1,4]

[1]Cell & Molecular Biology, The University of Texas at Austin, Austin, TX USA
[2]St. David's NeuroTexas Institute Research Foundation, Austin, TX, USA
[3]Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin TX, USA
[4]Integrative Biology, The University of Texas at Austin, Austin, TX, USA

November 23, 2015

*Dakota Z. Derryberry: 2500 Speedway, Austin, TX, 78712; (512)232-2459; dakotaz@utexas.edu

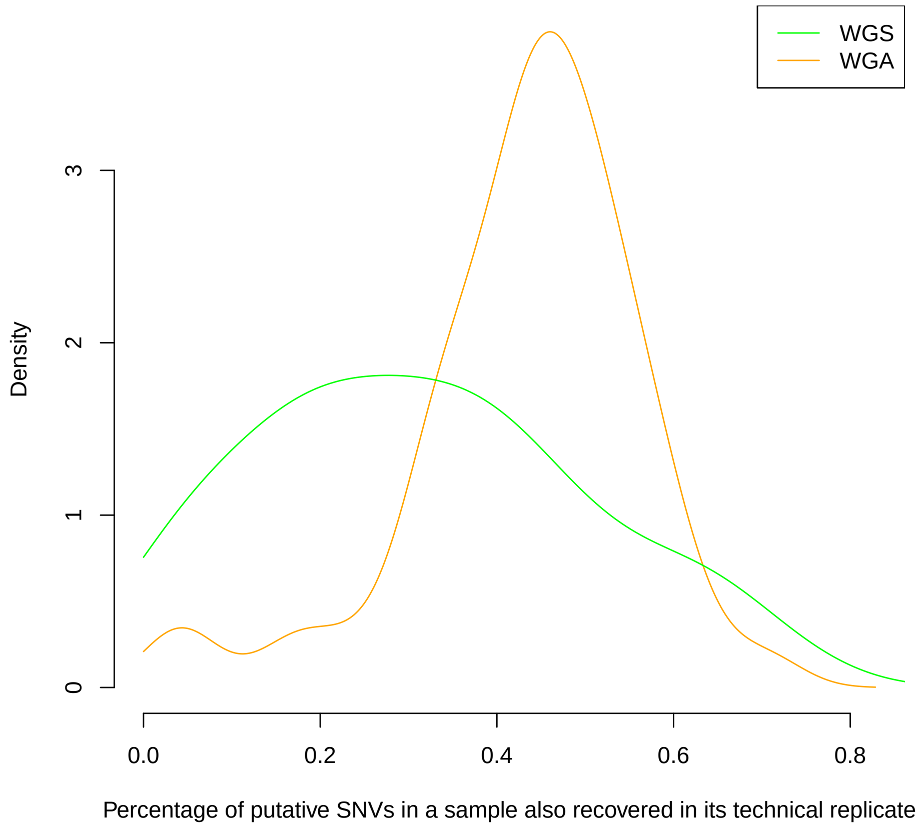

Figure 1: **One third (WGA) to one half (WGS) of putative SNVs were recovered in technical replicates.** Density of the percentage of each WGS (green) and WGA (orange) sample that is present in the overlap between replicates for each patient. The WGS distribution is higher and narrower, showing that the WGS samples overall have a higher percentage overlap than the WGA samples, and less range in this parameter.

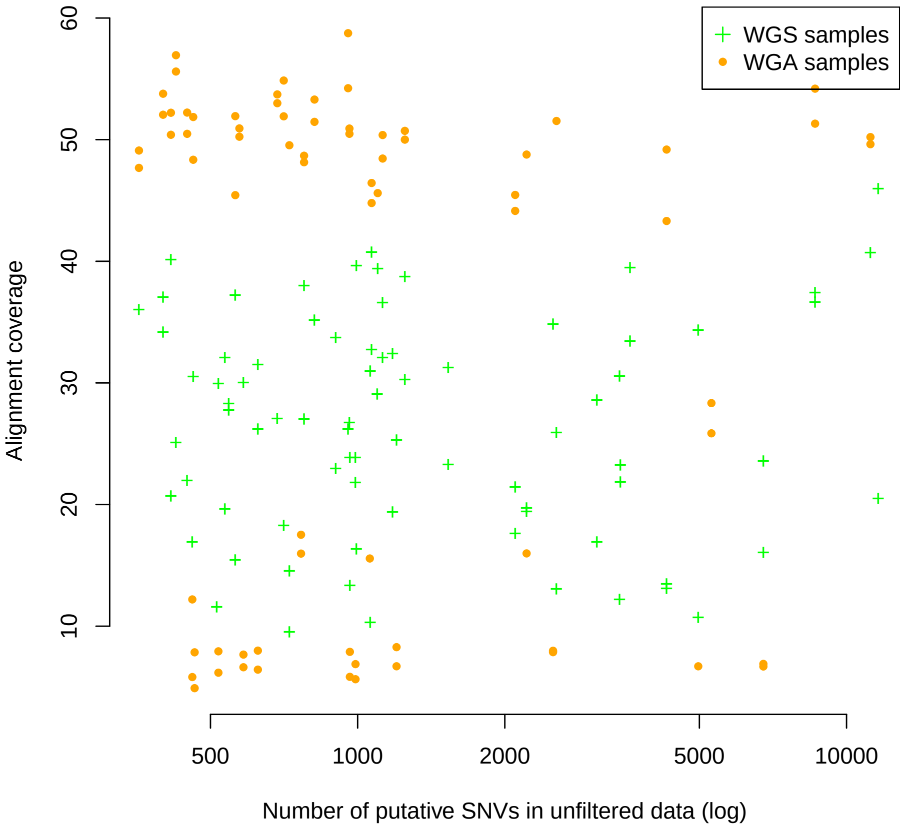

Figure 2: **Number of putative SNVs in a sample does not correlate with coverage.** The number of SNVs called in a sample does not correlate with the coverage of that sample (Spearman $\rho = -0.13$, $S = 671817$, $P = 0.12$). While at first glance, it appears that there may be significant separation between WGS and WGA samples along the y-axis (WGA samples lie above and below WGS samples), this is not the case because the y-axis value (coverage) is an experimental and not a test parameter. It does not impact our results.

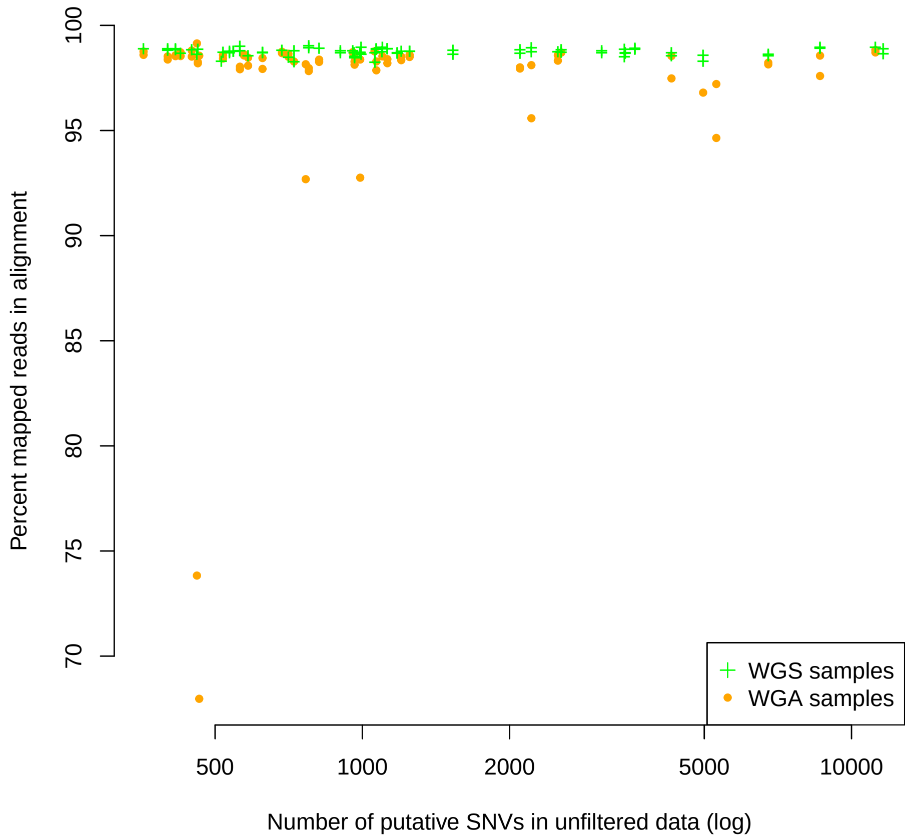

Figure 3: **Number of putative SNVs in a sample does not correlate with percentage of mapped reads.** The number of SNVs called in a sample does not correlate with the percentage of mapped reads in the alignment of that sample (Spearman $\rho = -0.068$, $S = 637326$, $P = 0.41$).

Table 1: **Back-end Processing.** This table shows the software packages we used in data processing, what we used each piece of software for, and the command associated with it. The rows are in order of use.

software	purpose	command
picard	regenerate fastq files from BAM file aligned to hg18	`java -d64 -Xmx4g -jar SamToFastq.jar I=$pfx.bam F=$pfx.1.fastq F2=$pfx.2.fastq 2>&1`
bwa	align fastq files to hg19	`bwa aln -q 30 -t 8 $hgReference $fastq > $fastq.aln.sai`
bwa, samtools	convert aligned fastq files into new BAM file	`bwa sampe -a 600 -P -r "$RG" $hgReference $fastq1.aln.sai $fastq2.aln.sai $fastq1 $fastq2 | samtools view -bSh -o $outprefix.bam -`
samtools	sort and index new BAM file	`samtools sort -@ 16 $outprefix.bam $outprefix.sorted 2, samtools index $outprefix.sorted.bam 2`
samtools	remove duplicate reads from BAM files	`samtools rmdup ../$tumorpfx/$tumorpfx.out.sorted.bam $tumorpfx.dedup.bam`
GATK	indel realignment	`java -d64 -jar $gatkJar -R $hgReference -T IndelRealigner -rf BadCigar -I $tumorpfx.dedup.bam -known $G1000.Mills -known $G1000.Phase1.Indels -targetIntervals $tumorpfx.intervals -o $tumorpfx.realn.bam`
GATK	base recalibration	`java -d64 -jar $gatkJar -nct 8 -T BaseRecalibrator -rf BadCigar -I $tumorpfx.realn.bam -R $hgReference -knownSites $dbSNP -o $tumorpfx.recal.grp`
samtools	index recalibrated BAM file	`samtools index $tumorpfx.realn.recal.bam`
SomaticSniper	call somatic mutations, generate VCF	`bam-somaticsniper -q 40 -Q 40 -J -s 0.001 -F vcf -f $hgReference $tumorbam $normalbam $tumorpfx.SS.vcf`