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

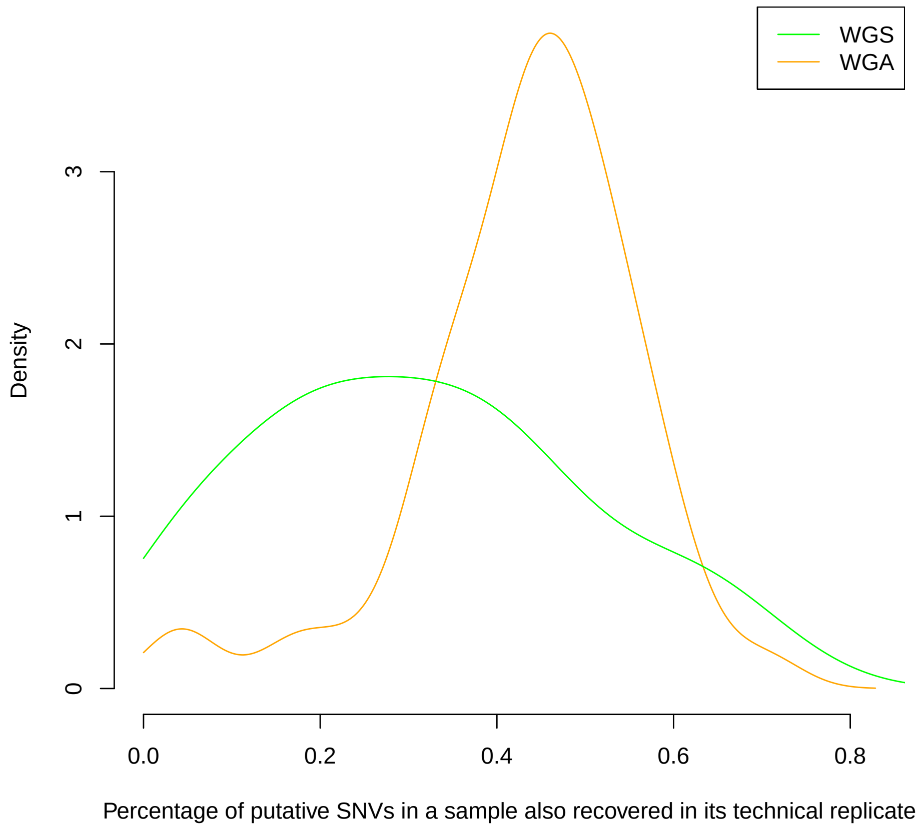

Figure 1: **One third (WGA) to one half (WGS) of putative SNVs were recovered in technical replicates.** For each pair of replicates (WGS and WGA), we looked at the percentage of WGS SNVs that were recovered in the WGA sample (about one half, in green), and the percentage of the WGA SNVs that were recovered in the WGS sample (about one third, in orange). 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.

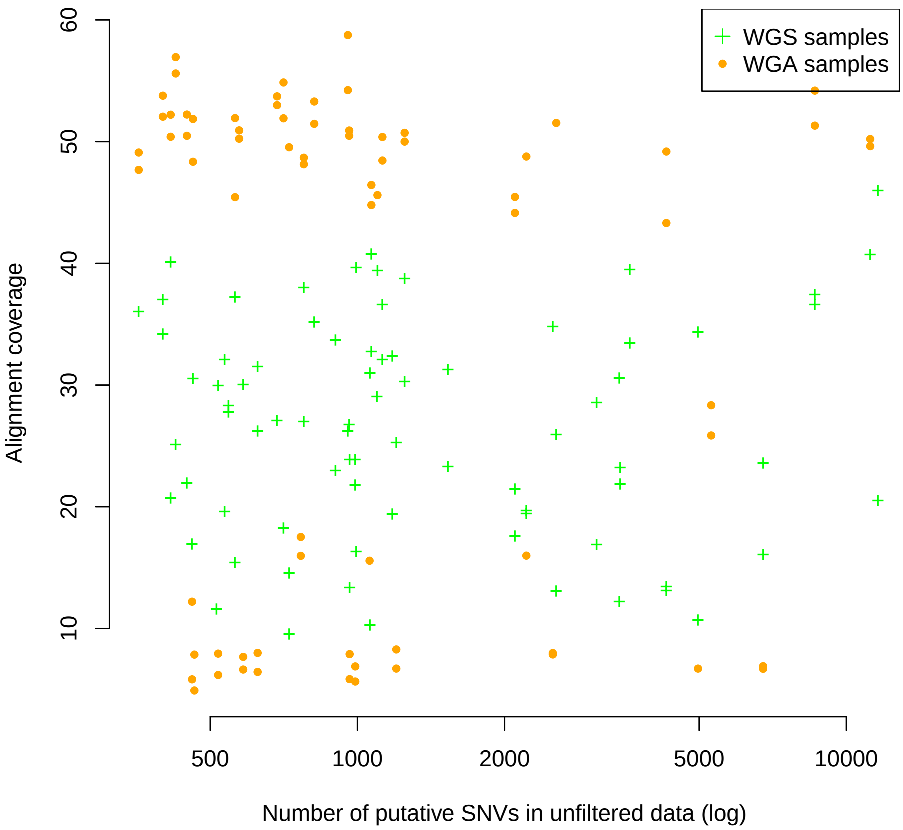

Figure 2: **Number of putative SNVs in a sample does not correlate with coverage.** The number of SNVs called in a sample does not correlate with the coverage of that sample (Spearman $\rho = -0.13$, $S = 671817$, $P = 0.12$). This is shown by the consistent variability along the x-axis at each level of overage (shown on the y-axis). The separation of the two experimental condition on the y-axis is not relevant to this measure.

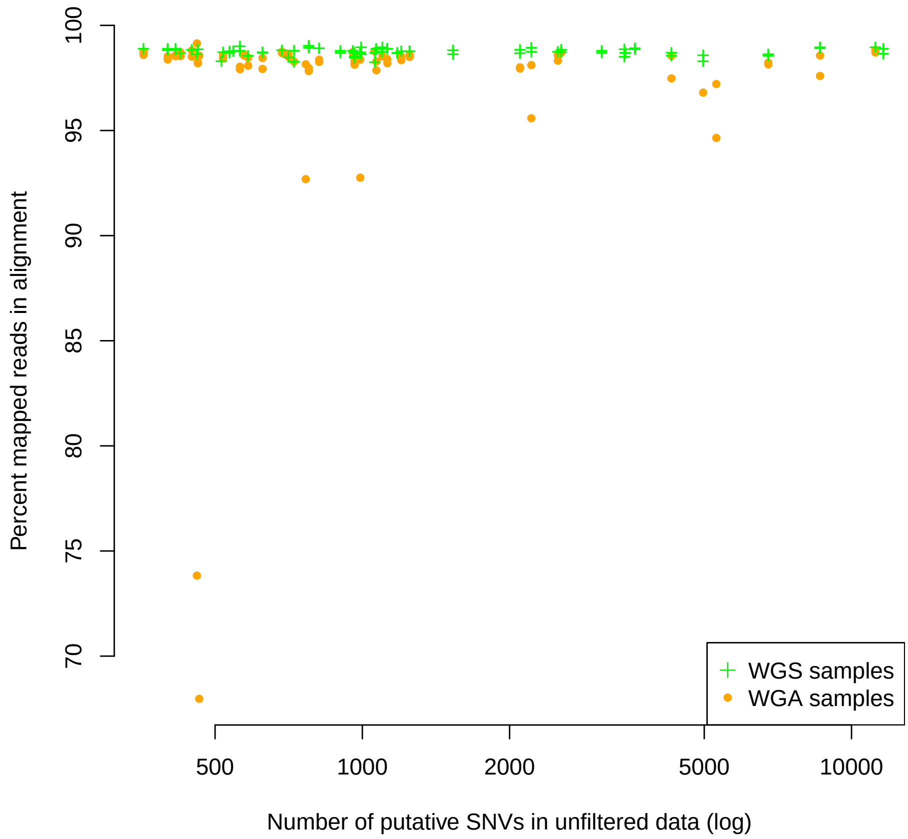

Figure 3: **Number of putative SNVs in a sample does not correlate with percentage of mapped reads.** The number of SNVs called in a sample does not correlate with the percentage of mapped reads in the alignment of that sample (Spearman $\rho = -0.068$, $S = 637326$, $P = 0.41$).

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 the use of it. The rows are in order of use.

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