Monday, November 23, 2015

Min Zhao

Academic Editor for PeerJ

Dear Dr. Zhao,

Thank you for inviting us to submit revisions for our manuscript, “Reproducibility of SNV-calling in multiple sequencing runs from tumors.” Here is our revised manuscript. A detailed rebuttal with our response to the reviewer’s comments follows.

Sincerely yours,

Dakota Derryberry, Matt Cowperthwaite, Claus Wilke

Thank you for your helpful comments on our manuscript. We have done our best to address all concerns.

**Reviewer 1 (Anonymous)**

1) The article is written in English using clear and unambiguous text.

2) The submission is within the scope of PeerJ.

3) The authors used public Glioblastoma multiforme (GBM) data set with technical replicates from TCGA.

Thank you for the compliment on our work. We’re glad you like the changes.

4) The manuscript looks much better now, and I think it is almost acceptable. Here are some little suggestions:  
  
Double-check the index of figures and tables. For example, on page 33 and page 35 of the 'peerj\_reveiwing\_\*.pdf' I believe the title should be Figure 5 and Figure 6 instead of Table 3 and Figure 5, which I am not sure if is due to PeerJ review system.

We have investigated this and believe this is due to the PeerJ Review system. We have also double checked our indexing, so there should be no problems in the final manuscript.

5) Table S1 could not be downloaded from PeerJ, which seemed to be in .tex format. It might be better provided in .xls or .pdf format.

PeerJ asked us to submit this in .tex format. Thus, we will leave the .tex as it is, but we can add a pdf as well.

6) The color (hue) for WGA and WGS is better to be the same or similar (accordant) in relevant figures, such as Figure 1, Figure 3 and Figure S1.

We have changes the relevant figures so that now, across figures, when there is a distinction to be made, green indicates WGS and orange indicates WGA.

7) To better answer previous Reviewer 2's comment (2), it might be better to distinguish WGA and WGS in Figures S2 and S3, such as in different color.

We have done so, consistent with comment (6) above. We have also made the green WGS points crosses, as in comment (12).

8) In Figure 5, as the top two subplot represent the same data, I think their y-axis should be in the same scale (0-80) and aligned. For better clarity, the y-axis labels might be needed for all four subplot, and it might be better added the formula of delta\_Jaccard to the figure legend.

We have changed the axes as requested. The formulas requested are highlighted in red here, but not in the paper. The figure legend now reads:

Figure 5: **Independent performance of individual filters of the difference between replicates.** For each sample, we compared the size of the difference between replicates (the number of SNVs found in only one of WGS or WGA) to the size of the overlap between replicates (the number of SNVs recovered from both the WGA and WGS replicates). *(top left)* A scatter plot of the ratio, per sample per filter, of putative SNVs removed from the difference versus the overlap of WGS and WGA (WGS△WGA / WGS∩WGA). *(top right)* Boxplot of the same data, divided by filter on the *x*-axis. *(below)* Plot of the percentage difference in the Jaccard(WGS, WGA), with Jaccard = | WGS ∩ WGA | /

| WGS ∪ WGA |, before and after the action of each filter (each filter was run on the whole data set independently).

9) In Methods and legend of Figures S2 and S3, the mean, stdev, spearman rho and p-value seem to have too many significant digits.

Thanks for pointing this out. It’s fixed!

10) The legend in Figure 3 is better moved to top left, so that it will not overlap a data point.

Thanks for pointing that out! It’s been done.

11) Adjust the x-axis label of Figure S1, so that it could be totally seen.

Fixed.

12) For Figure 3, the scatter points from WGA and WGS seemed to be closely distributed, so it might be better to use different symbols for them, such as circle and cross.

To improve the readability of the figure, we have changed the green WGS points to crosses, and left the orange WGA points as dots.

13) Double-check the figures and tables, including their legend, especially for supplementary materials. They are better to be self-contained.

In accordance with this suggestion, we changed the following figure legends so that they now read:

Figure 1: **Data processing pipeline.** For each of 55 patients, we began with a C484 tumor BAM file (WGS), a C282 tumor BAM file (WGA), and a normal BAM file, all aligned to hg18. For each BAM file, we used picard to regenerate fastq reads, bwa to realign the fastq files to hg19, and GATK to recalibrate bases and indels. We used SomaticSniper to call somatic mutations (differences between the tumor and normal sequences) for each replicate. When we had a VCF for each replicate, we calculated the overlap between the two lists as the number of individual mutations that appeared in both replicates.

Figure 2: **Number of putative SNVs in WGS versus WGA, as called by SomaticSniper before filtering.** For each patient (represented by one point), we compared the number of SNVs called in the WGS sample (*y*-axis) versus the number of SNVs called in the WGA sample (*x*-axis). As expected, the number of mutations found in one replicate correlates with the number of mutations found in the other replicate (Spearman ⍴=0.42, S=16142, P=0.002). The line shown is *y=x*, so points falling below the line agree with the hypothesis that an additional amplification step produces more sequencing errors in a sample.

Figure 3: **Number of putative SNVs per sample does not correlate with the number of putative SNVs recoverable in both replicates.** To test the hypothesis that samples with a greater number of SNVs also have a greater number of errors, and consequently a smaller percentage of SNVs recovered in the overlap between replicates, we compared the percentage of putative SNVs recovered in the overlap between replicates for a sample (*y*-axis) versus the number of mutations in that sample (*x*-axis). We found no correlation (Spearman ⍴=0.05, S=29268, P=0.68), suggesting that higher numbers of SNVs in a sample may not be only due to errors. We calculated the percent overlap in two ways: with reference to the total number of putative SNVs in the WGS sample (green) and with reference to the total number of putative SNVs in the WGA sample (orange). The correlation was calculated with respect to WGA.

Figure 4: **Independent performance of the individual filters of the overlap between replicates.** For each filter (names given on the *x*-axis, with detailed description in Table 1), we looked at the number of mutations removed by a given filter in a given sample on a log scale *(above)*, and the percentage of the WGS--WGA overlap removed by the filter *(below)*, per sample. The LOH and VAQ filters removed a large number of putative SNVs and portion of the overlap. The <10% filter removed very few putative SNVs and almost none of the overlap. The 10bp-SNV, 10bp-INDEL, and dbSNP filters removed an intermediate number of putative SNVs (~100), but only a very small portion of the overlap, making them the best performers on the overlap data.

Figure S1: **One third (WGA) to one half (WGS) of putative SNVs were recovered in technical replicates.** 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 S2: **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 ⍴=-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.