November 13, 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 single tumors.” Here is our revised manuscript. A detailed rebuttal with our response to the reviewers’ and editor’s comments follows.

Sincerely yours,

Dakota Derryberry, Matt Cowperthwaite, Claus Wilke

Response to the reviewer comments

Reviewer 1

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

The article include sufficient introduction and background.

2) The structure of the article contains Title, Abstract, Introduction, Results, Discussion, and Methods, which almost conform to PeerJ standard sections (https://peerj.com/about/author-instructions/). The affiliations of authors need to be rearranged, as department, university, location, country, and contact information of corresponding author.

This has been fixed to comply with PeerJ requirements.

3) Figures are relevant to the content of the article. The font size in Fig. 1 is small. And some axis labels are confusing. 'No.' in Fig. 2 and Fig. 4 is better to be written as 'Number' or '#'. The x-axis label '% of sample obtained in overlap' of Fig. 3 is confusing. The x-axis label 'Length of overlap' in Fig. 7 and Fig. 8 is better to be 'Number of overlap'. And accordance between figures should be double-checked, including whether the first letter should be capitalized, usage of 'number' or '#', and 'percent' or '%'.

Thank you for all the suggestions towards making our figures more readable:

- In Figure 1, we have increased the font size from 12pt to 16-22pt. We have also heavily revised this figure, as detailed in response to Reviewer 1's comment (8).
- In Figures 2 and 4, we changed "No." to "Number." We have done further analysis of these figures, detailed in our response to Reviewer 1's comment (11).
- In Figure 3, we agree that the x-axis title is confusing. We changed it to "Percentage of putative SNVs in a sample also recovered in the samples's technical replicate." We discuss Reviewer 1's further comments on this figure in (13).
- In Figures 7 and 8, we changed both x-axes to be "Number of SNVs in both replicates," and correspondingly changed the y-axes to be "Percentage of SNVs in both replicates filtered out by XXX." We discuss these figures further in response to Reviewer 1's comments (15) and (17).
- In all figures, we have double-checked and fixed concordance: we have used "percentage," "versus," "number," and first-letter capitalization on figure titles and axes.

4) The work is 'self-contained', although some terminology looks not well-defined, and confusing, such as 'polymorphic mutation' and 'mutation frequency'.

Thank you for pointing this out. In order to better define our terminology, we have altered the text in the following ways:

First, we revised the sentence in which the phrase 'polymorphic mutation' first appears to read: "A polymorphic mutation is defined as a mutation that is not fixed and therefore is not present in all tumor cells. Such mutations may or may not be represented at high enough frequency in a particular tumor specimen to be identified with next-gen sequencing. Alternatively, a polymorphic mutation may appear to be fixed when present at at very high frequency in a tumor specimen."

Second, we revised the sentence in which the phrase 'mutation frequency' first appears to read: "Contrary to expectations, the percentage overlap between technical replicates did not decrease with increasing numbers of putative mutations, suggesting real variation in mutation frequency (the absolute number of SNVs found in a tumor) between samples from the same tumor."

As a note, mutation frequency is commonly calculated in the literature, but it is generally called a mutation rate. Although researchers are calculating and reporting a per base pair sequencing rate, it is often misinterpreted as a rate over time. We use "mutation frequency" to clarify.

5) Their raw data is downloaded from TCGA, and the authors claimed that the analyzing python and R scripts are available in a public git repository (github.com/clauswilke/GBM_genomics), which I attempted to open, but failed with '404 Page not found'.

Our apologies for the oversight. The git repository has now been made public.

6) The submission is within the scope of PeerJ.

The authors tried to assess the quality of TCGA sequencing data and somatic SNV-calling, by applying a general somatic SNV calling pipeline and several filters to technical replicates of 55 GBM samples, and analyzing the size of overlap.

The analysis is a good attempt, which is, however, quite rough. The details are put in the 'General Comments for the Author' part.

The methods look probably reproducible.

We do believe that our methods are reproducible. It is true that in this study, we did our best to benchmark the quality of the TCGA data and methods using analysis of technical replicates. However, our goal was not specifically to asses the quality of the TCGA data that we used; rather, our goal was to begin to develop robust methods for benchmarking next-generation sequencing data and analysis, which could be widely used in many areas of biology.

7) The authors used public data of technical replicates from TCGA.

Their analysis may need to be improved. And the conclusions and discussions may need revision.

Page1-2, Abstract and Introduction. When I read the Results, I suddenly recognized that the authors seemed to talk about somatic mutations by comparing tumor and blood, and figure out the overlap of called somatic mutations between two technical replicates. If this is what they mean, I think it is better to emphasize somatic mutation or somatic SNV in Abstract and Introduction.

We agree that this was unclear. To clarify, we have significantly changed the first two paragraphs of the introduction. They now read:

"Glioblastoma multiforme (GBM) is the most common and deadly primary brain tumor with a median survival time of 14 months and a 5-year survival rate of 5%. Prognosis for patients with this disease remains poor despite significant research investment, due to the difficulty of surgical resection and the limited number of effective chemotherapeutics (Wilson et al., 2014). Recently, research to improve treatments for patients with this devastating disease has focused on the idea of precision medicine, that is using large- scale genomics data to discover how the disease arises and progresses, and how to stop it. The past six years have seen an explosion of data in cancer genomics, an effort led by TCGA, an archive of publicly-available data that includes sequencing of paired tumor- normal samples from individual patients for thousands of tumors (The Cancer Genome Atlas Research Network, 2008; Brennan et al., 2013), including over 500 GBMs. While these and other similar data (Parsons et al., 2008), researchers have discovered genes and pathways mutated in GBM (Cerami et al., 2010), discover different GBM subtypes (Verhaak et al., 2010), and develop a variety of computational models to find GBM driver mutations (Gevaert and Plevritis, 2013). Despite widespread use of the data and adoption of the methods, however, efforts to benchmark the data–to assess the quality and repeatability of these and subsequent analyses–have lagged or been non-existent. Here, our goal is to begin to develop robust methods for benchmarking next-generation cancer sequencing data and analysis.

“It is well known that sequencing and variant-calling pipelines are not error-free. For example, different pipelines for calling single-nucleotide variants (SNVs) can return dif- ferent results on the same data (Yu and Sun, 2013). Given the heterogeneity in cancer genomes (Kumar et al., 2014; Friedmann-Morvinski, 2014), and the presence of func- tional low-frequency variants in GBM (Nishikawa et al., 2004), the signal-to-noise ratio in the TCGA dataset may be particularly low. Yet despite widespread use of this dataset, and significant monetary investment in collecting and analyzing the data, we know little about how to maximize the quality of the sequence data and SNV-calls. One way to ad- dress this question would be to analyze technical replicates of sequencing data (Robasky et al., 2014). Here, we addressed one aspect of this question, by asking whether the same SNV-calling pipeline will return comparable results on two sequencing runs from the same tissue. And further, we asked whether added filtering after SNV-calling increases or decreases the degree of similarity among replicates. “

8) Also, in Fig. 1, it is unclear what is the use of the blood sample. So the diagram is better to be redrawn to show that.

Thank you for pointing out where our methods and analysis are unclear. We have heavily revised Figure 1 to make our pipeline clearer, as well as to clear up other parts of the manuscript and reduce redundancy. Specifically, we added boxes for all three hg19 aligned bamfiles, and showed paired tumor/normal samples leading into the SomaticSniper step, which is the step where tumor and blood are compared. We hope these changes increase the clarity of our methods.

9) Page 3, Results. The authors seem to use the phrase 'polymorphic mutation' to mean a somatic mutation site. I am not sure whether 'polymorphic mutation' is a correct, commonly-used term.

Please refer to comment (4).

10) Page 3, Results. The authors said 'Theoretically, any fixed mutation will appear in all replicates, while errors due to (i) sequencing errors, (ii) amplification errors, or (iii) alignment errors will not.' However, in my opinion, system errors in alignment and variant-calling can actually appear in replicates. Although the authors discussed the errors in both replicates in

Discussion, such as DNA degradation and sequencing center factor, they did not mention any system errors due to alignment and variant-calling, which I think might dampen their analysis, and must be analyzed or discussed. In my experience, searching for somatic mutations will enrich more such artifacts than for germline variants, and thus, PCR is required to validate the putative mutations. Therefore, I will not regard the overlap as true positives, although I believe true positives will be in the overlap, if they exist.

We agree with this comment completely: appearing in two samples does not necessarily mean that a mutation is a true positive. However, we do think that it is *more likely to be* a true positive than a mutation appearing in only one sample, so that a pipeline which calls more overlapping SNVs probably has more true positives than a pipeline that calls more non-overlapping SNVs. Because our analysis is not concerned with the validity of individual mutations, but is rather about trying to assess the overall quality of mutation calling pipelines and sequencing data, we think that this is sufficient.

Nevertheless, your point that searching for somatic mutations can enrich for artifacts is well taken. In the discussion, where we discuss the limits of our analysis, we have added the following so that the paragraph now reads: “Several factors limit the conclusions we may draw from this analysis. First, in this analysis we used repeatability between technical replicates (being in the WGS and WGA samples) as a measure of confidence in a putative SNV. This metric is potentially problematic for many reasons, including but not limited to: (i) cancer is highly heterogeneous, and so a legitimate somatic SNV might show up in one replicate but not another; (ii) if the DNA sample is degraded to some extent, due to surgery conditions or some other factor out of the hands of the sequencing center, the same errors may appear in both replicates; (iii) the SNV calling process may enrich for artifacts such as germ line variants, which are nearly indistinguishable from somatic SNVs in computational analyses; cross-referencing to gold standards such as dbSNP is the only way to identify germline SNPs; (iv) a putative mutation that is present in both samples may be a somatic mutation that arose before the tumor. Although repeatability across technical replicates cannot guarantee that a putative mutation is a true somatic mutation, it does increase the likelihood. Having an independent measure of confidence in SNV calls, even an imperfect one, can help us gauge the accuracy of other measures, specifically of different filtering approaches.”

We also agree that validation of a mutation with an orthogonal method (*e.g.* PCR, Sanger sequencing, or some arrays) is necessary when one is concerned with the validity of individual mutations, and indeed we say so in the manuscript. To make this more clear, we have altered the relevant paragraph in the Discussion to read: “We have shown that there is significant overlap between technical replicates of whole exome sequencing in the TCGA GBM dataset, comprising about 50% of putative SNVs in WGS samples and about 30% in WGA samples. The overlap exists even for samples with a high number of putative SNVs, suggesting that some GBMs may have significantly more somatic mutation than others. We acknowledge that the high rate of non-concordance between replicates in our analyses indicates that even the best computational analyses are insufficient to validate any single SNV; when the identity of a single SNV in a single sample is important, validation by an orthogonal method (*e.g.* PCR, Sanger sequencing) remains necessary. Nevertheless, when less fine-tuned methods are acceptable, or are the only option available, one may wish to employ other methods of validation rather than nothing, such as the six data filters that are commonly applied to validate SNVs that we examined. We found that some of these filters remove principally those mutations found in one sample or the other, while other filers remove primarily those in the overlap. We suggest that when orthogonal validation is not an option, only the filters that removed little overlap between samples should be used for computational SNV validation.”

11) For Fig. 2 and Fig. 4 and related analysis, I am not sure how the authors computed the Pearson correlation coefficient, whether it is computed on the original count number scale or on the log scale. (According to the distribution, I think the latter is more reasonable). Computing on different scale may give different Pearson correlation coefficient. Therefore, I suggest to use Spearman correlation coefficient and corresponding test, which will not change on different scale.

We agree this could be clarified. We actually computed the correlation between variables three times (Spearman, Pearson on raw data, and Pearson on log-transformed data) with very little difference in results. In the original manuscript we used the Pearson correlation on the raw data, but your point that this is the least conservative measure, and therefore possibly the least trustworthy, is well taken. We have changed the manuscript to use the Spearman correlation, and have also specified that this is the test that we used.

The results of all three tests (and the R code) can be found in the script Figures.R in the git repository, but to summarize:

- Figure 2 (all with $p<0.005$): Pearson (raw) cor=0.51, Pearson (log transformed) cor=0.44, Spearman ρ=0.42
- Figure 4 (all with $p>0.5$): Pearson (raw) cor=-0.04, Pearson (log transformed) cor=-0.01, Spearman ρ=-0.06

12) Page 4. The authors said 'We further found that the percent overlap between the two samples, calculated as ..., was fairly consistent, on average 31% in WGA replicates and 44% in WGS replicates (Table 1).' However, in Table 1, it is clear to see that the percent overlap ranged from 1-74% for WGS (avg. 31%, sd 20%) and 3-71% for WGA (avg. 44%, sd 13%). Can such huge variation range be said 'fairly consistent'? And the average percent is not consistent between the main text in Page 4 and the number in Table 1.

The inconsistency between Page 4 and Table 1 is a typo, and has been fixed. The point that such a wide range is not “fairly consistent” is a good one, and we have changed the manuscript to better represent the data. The relevant sentences now read: “We found that the percent overlap between the two samples, calculated as [equations], varied widely, from 1%-74%, but was fairly evenly distributed around the average of 31% in WGA replicates and 44% in WGS replicates (Table 1). As expected, the distribution of the percentage overlap was narrower and taller in the WGS replicates, because on the whole the WGA samples had more amplification errors than the WGS samples (Figure 3).”

13) For Fig. 3, I am not sure whether Fig. 3 is meaningful after given Fig. 2. Because the number of putative SNVs is larger in WGA than in WGS, the overlap proportion is certainly higher in WGS than WGA. And as shown in Fig. 2 and Table 1, the variance of the number of putative SNVs is also larger in WGA than in WGS, so it might be reasonable to see the variance of proportion of overlap is larger in WGA than in WGS. Thus, the meaning of Fig. 3 seems to be vague, and the sentence describing Fig. 3 in the text seems not to provide any more information. It might be better to use a scatter plot, showing the proportion of overlap against number of putative SNVs in WGA (just like Fig. 4, proportion of overlap against number of

putative SNVs in WGS), which may merged with Fig. 4 with different dot color or shape, and segments or arrows connecting corresponding dots on the merged figure may be added.

We like this idea. We have followed the suggestion given here nearly exactly, which entailed:

(1) Moving the original figure 3 to supplementary materials (now Supplementary Figure 1)
(2) Adding points (in red) to the original figure 4 (now figure 3) representing the comparison with WGA
(3) Changed the text of Results reflect these changes: "As expected, for each sample the WGA replicate (with the additional amplification step) had slightly more mutations overall, with a slightly smaller percentage appearing in the overlap (Figures 2 and 3). … As expected, the distribution of the percentage overlap was narrower and taller in the WGS replicates, because on the whole the WGA samples had more amplification errors than the WGS samples (Supplementary Figure 1)."
(4) Changed the caption of the new figure 3 (old figure 4): "Number of putative SNVs per sample does not correlate with the number of putative SNVs recoverable in both replicates. The percentage of putative SNVs in a given sample that are in the overlap between replicates is not correlated to the number of mutations in that sample (Spearman rho=0.05, S=29268, P=0.68). We calculated the percent overlap in two ways: with reference to the total number of putative SNVs in the WGS sample (red) and with reference to the total number of putative SNVs in the WGA sample (black). The correlation was calculated with respect to WGA."

14) Page 5-6. 'Does more sophisticated SNV filtering software increase or decrease the degree of similarity between replicates?' In this part, the authors analyzed the influence of several filters on the number of putative mutations and the overlap between technical replicates. They focused on the overall number of filtered putative mutations and the percent of filtered among overlap, and correspondingly plotted Fig. 5-8. In Page 6, they mentioned to answer 'what percentage of putative SNVs removed by a given filter was in the category more likely to be true positives (overlap), versus the category more likely to be false positives (only present in one replicate)?'. However, they only demonstrated some filters removed many putative sites and corresponding much percentage among overlap. They did not directly compare the percentage between the overlap and the list only presenting in one replicate (I calls it 'diffset' in the following), or whether each filters removed more proportion of false positives than true positives or vice versa, which I think should be done. This might be done by comparing the Jaccard similarity coefficient (= overlap / union) before and after each filter, or doing Fisher's exact test on two-by-two contingency table of how many sites were filtered/not filtered for overlap/diffset.
And thus, Table 1 might be too summary for further examining and explaining their results. I suggest to make a supplementary table containing the sample name and the statistics in Table 1 of amount (proportion is not necessary) for each of 55*2=110 tumor samples.

This is an excellent idea, which we think significantly deepens the analysis. We did the analysis of the 'diffset' as suggested by the Reviewer, and made the following changes to the manuscript:

(1) Added a new Figure 6, which shows an analysis of the "diffset"

(2) Modified the text to reflect these changes in the following ways:
    (1) Figure 5 caption: "**Independent performance of individual filters of the difference between 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) before and after the action of each filter (each filter was run on the whole data set independently)."

(2) Results: "We next looked at the individual effects of six of eight filters on the difference between replicates of each sample (those mutations found only in WGS *or* WGA, *but not* both. We looked at the ratio, per sample per filter, of putative SNVs removed from the difference versus the overlap of WGS and WGA ($WGS \triangle WGA / WGS \cap WGA $), where a higher ratio indicates that more mutations are removed from the difference than the overlap. We first plotted this ratio for each sample for each filter (Figure 5, *top left*). To better compare the filters to each other, we looked at the distribution of ratios across samples for each filter (Figure5, *top right*). We found that the LOH and VAQ filters scored worst, followed by the 10bp-SNV, 10bp-INDEL, and dbSNP filters, in that order. (We did not look at the <10\% filter for this analysis, because it did not catch enough data to be meaningful.) Thus, those filters that remove the fewest putative SNVs in the overlap, also remove the highest number of SNVs in the difference relative to the overlap.

We next looked at the similarity of WGS to WGA as a whole as measured by the difference, normalized to 100, in the Jaccard similarity coefficient (WGS ∩ WGA / WGS ∪ WGA) before and after filtering (Jaccard after-Jaccard before /Jaccard before). By this measure, and in contrast to the previous two measures, we found that LOH and VAQ performed the best by an order of magnitude, followed by 10bp-SNV, 10bp-INDEL, and dbSP, in that order (Figure5, *below*). We see this effect because these two filters remove so much data: half or more of the overlap and as much or more of the difference. Removing so much of the data makes the intersection much, much smaller, and thus makes the Jaccard coefficient between samples much larger, mostly independently of the size of the overlap."

Regarding the summary statistics in Table 1, the more complete data (summary statistics for each individual sample) are available in the git repository under FIGURES/FIGURE_DATA/.

15) Besides, as shown in Fig. 7 and Fig. 8, the two figures looked like mirror, which is quite interesting. And as mentioned in the last paragraph of Results, the authors' explanation implied that the percentage was calculated when filtering putative sites only once using all six filters sequentially. If this is true, the order of the six filters may affect the results, and Fig. 7 and Fig. 8 might not be so meaningful. In fact, I think it is better to evaluate each filter on the same original putative sites, and report their effect on overlap and also on diffset. And I am not sure why the authors said '... we found that for samples with overlap of < 100 there was no strong trend.' (why the authors focused on overlap number < 100), for Fig. 7 seems to have a clear positive correlation trend and Fig. 8 with a clear negative correlation trend.

We apologize that our methods were insufficiently clear in the original manuscript, but to clarify: the filters were run independently, not sequentially. Thus, any SNV that we caught by more than one filter was counted more than once. Thus, the sum of the percentage of SNVs removed by each filter for a single sample does NOT equal 100%; in all cases it equals much more than 100%; and this data is available in the git repo. We have clarified this point in the manuscript, by adding the following sentence to Results: "We ran each filter independently on the whole dataset to see which putative SNVs it caught."

We additionally added the following sentence to Methods: “Each of the eight filters was run independently on the whole data set.”

In light of that, it is interesting that Figures 7 and 8 are mirror images of each other. We do not currently have an explanation for why these two filters (VAQ and LOH) remove almost disjoint sets of SNVs, and we say this in the manuscript. We would love to hear ideas or suggestions on this point.

16) By the way, as shown in Table 1, each statistic seemed to vary much between samples. And the authors reported that the VAQ filter removed a large proportion of putative mutations, as shown in Fig. 8. So I wonder if these 110 samples of GBM in TCGA have much variation or 'heterogeneity' on their sequencing quality, so that the putative mutations from the samples with very low quality will probably be filtered. Therefore, I think the authors may need to describe more about the data and their quality (homogeneity).

We address this point more fully in Reviewer 2’s comment 2. To summarize: we have added a two supplementary figures, in which we graph the read coverage and percent mapped reads in the alignment of each sample agains the total number of SNVs called in that alignment. We find no correlation in either case (by Pearson or Spearman), and although we haven’t included it in the manuscript (to limit redundancy), we found the same results when graphing both coverage and percent mapped reads against the size of the overlap.

17) In summary, I list the two most important problems: (1) overlap does not mean all true positives, thus filtering in overlap might not always be wrong, so it might be more clear to compare the filtering effect in overlap and diffset; (2) filters may need to be evaluated separately, and if Fig. 7 and Fig. 8 are still in that mirror shape, it will be interesting.

We thank Reviewer 1 for these comments, and hope that we have adequately addressed your two main concerns in points (10) and (15).

Reviewer 2

It’s a very comprehensive study on QC in WGS analysis. I concern on two points:

Thank you for the compliment on our work.

1) There are too many figures and tables in the main text. It seems a little bit duplication among some of them. For example, both table3 and figure 1 are focused on data processing process. The workflow in figure is clear enough for reader to get the whole experimental design. The Table3 which contains the technical details of the workflow could be move to supplementary.

In response to both reviewers and the editor’s comments on redundancy in figures, we have made the following changes:

- Heavily modified Figure 1
- Moved Table 3 to Supplementary Table 1, and changed relevant text
- Moved density plots from Figure 3 to Supplementary Figure 1

- Altered what was Figure 4 (now Figure 3) to include overlap statistics as calculated with respect to WGS as well as WGA, and modified the figure caption accordingly (see comment (13) above)
- Combined what were formerly figures 5 and 6 into a compound Figure 4, and modified the figure caption to read: "**Independent performance of individual filters of the overlap between replicates.** The *x*-axis gives the name of each filter (detail on Table 1), and the *y*-axis gives (*above)* the number of mutations removed by a given filter in a given sample on a log scale, and *(below)* the percentage of the WGS--WGA overlap removed by the filter, 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."
- Added a new Figure 5 to address the analysis suggested by Reviewer 1 in comment (14) above, and modified the text accordingly (see comment (14) above)
- Combined figures 7 and 8 into a compound Figure 6, and modified the caption to read: "**Replicate SNVs filtered out by LOH and VAQ are almost completely non-overlapping and cover most of the sample.** As the length of the overlap between WGS and WGA increases (x-axis), *(above)* the percentage of the overlap filtered out by LOH increases, and *(below)* the percentage of the overlap filtered out by VAQ decreases. The two are almost (but not quite) perfect inverses; putative SNPs filtered by LOH and VAQ cover almost the entire overlap, and together sum to nearly 100\% of the overlap."

We hope that this has sufficiently improved the readability and interest of the manuscript.

2) Generally, the number of SNVs in WGA are larger than WGS. The author mentioned that it may caused by the greater number of amplification errors. Herein, I'm interested on the percentage of mapped reads among the bam files of WGS and WGA groups after local alignment process in GATK. The higher percentage of mapped reads may also indicate more SNVs. If the mapped reads are in the similar level of WGA and WGS, it would be more reliable to conclude that the larger number of SNVs was caused by amplification errors.

This is an excellent point. To address this, we have added two things to the Supplementary materials: (i) a Supplementary Figure 2, in which we graph coverage as a function of the total number of SNVs per sample, and (ii) a Supplementary Figure 3, in which we graph the percentage of mapped reads as a function of SNVs per sample. Supplementary Figure 1 shows that there is no correlation between coverage and the number of SNVs predicted per sample. Figure 2 shows that the percentage of mapped reads is universally high (only below 90% in two samples, and generally well above 98%), again suggesting that mapping quality is not a major source of variation in the data.

We have also added two sentences to methods stating that mapping quality is universally high and that it does not predict the number of SNVs or the size of the overlap: "For all alignments, average read coverage was 30X, with a low of 5X and a high of 60X (stdev=15.95898). The percentage of mapped reads in our alignments was universally high (mean=98.03545, stdev=3.257872), with only two samples below 90% (68% and 74%), and only 15 samples below 98%. We found no correlation between the read coverage or the percentage of mapped reads and the total number of SNVs called (Supplementary Figures 2 and 3)."