Dear Editor,

Thank you very much for your detailed reviewed provided by yourself and the associate editor. We have revised the article to address each concern of the review team.

----------------------- REVIEW 1 ---------------------

1. You allude to the differences between RVD and RVD2 but do not show any comparable RVD results?

**We have added the comparable RVD results and analysis in Section 5.2 for synthetic sequence data. As to the clinical dataset, since the average read depth 90x does not satisfy the high read depth requirement of RVD, the program failed and therefore the result was not shown.**

2. It was not clear to me whether the model applies to complex clonal mixtures, or to binary tumor/normal mixtures, with only one variable ratio.

**The model was applied to complex clonal mixture. In the Synthetic DNA sequence data, sample of the case and control DNA were mixed at defined fractions to yield defined MAFs of 0.1%, 0.3%, 1%, 10% and 100% (Section 4.1 sentence 2). The estimated tumor purity was reported to be 0.8 in clinical HCC1187 sequence data (Section 4.2 sentence 3).**

----------------------- REVIEW 2 ---------------------

Major comments:

1. This paper should include discussion on methods developed for heterogeneous samples. Examples include:

“A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data” (Yau et al. 2010) “SNP Arrays in Heterogeneous Tissue: Highly Accurate Collection of Both Germline and Somatic Genetic Information from Unpaired Single Tumor Samples”, (Assie et al., 2008) “Analyzing cancer samples with SNP arrays”, (Loo et al., 2012)

**We clarified Section 1 Paragraph 4 is about algorithms developed for heterogeneous samples. Also, we added the suggested reference “A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data” (Yau et al. 2010) to this paragraph. This paper is about a variant detection method where tumor sample purity is estimated as the estimation is embedded to the calling algorithm. We talked about four more algorithms, VarScan2, Strelka, muTect and RVD which do not estimate the tumor purity along the detecting process. We didn’t include the two other papers because they are less relative to RVD2 due to the type of dataset they target at or the goal of their method.**

2. The method part should be written more clearly. For example, what is replicate in page 2? What is the motivation of the specific graphical model design? How the sample heterogeneity is considered in the graphical model? Also, a small concrete example to illustrate the ideas behind the proposed graphical model would be helpful.

**Regarding what is replicate, we added a clarification in Section 2 paragraph 2 sentence 3. We have designed the graphical model firstly because it is a generative model which can well capture the error rate distribution in next-generation sequence data using latent and observable variables. We have assigned biological meaning to each latent parameter, as shown in Section 2 paragraph 4. We make it clearer what graphical model is and why use graphical model in Section 2 paragraph 7. We clarified in Paragraph 4 Section 1 that RVD2 does not estimate tumor sample heterogeneity during the variant calling algorithm. By using normal-tumor paired sample, RVD2 is able to call variants regardless the unknown tumor purity. Last but not least, we added a small simulation example to illustrate the graphical model and generative process in Supplementary Information 1.**

3. In the experiments, the paper should contain comparison with other methods that can take into account sample heterogeneity.

**We have made it more clearly in the Introduction section (Section 1 Paragraph 4) and Result section (Section 5.2 Sentence 2) that VarScan2, Strelka, muTect and RVD are methods that can take into account sample heterogeneity. We have shown comparable results with those algorithms in Section 5.2 for synthetic sequence data. As to the clinical dataset, we showed the comparable result and analysis to VarScan2, Strelka, and muTect in Section 5.3.2. Since the average read depth 90x does not satisfy the high read depth requirement for RVD, the program failed and therefore the result was not shown.**

4. In the real data analysis, the proposed method should be compared with other existing methods. For example, only RVD2 was applied to HCC1187 data.

**As mentioned in above, for the clinical HCC1187 data, we added the comparable result and analysis to VarScan2, Strelka, and muTect in Section 5.3.2. Since the average read depth 90x does not satisfy the high read depth requirement for RVD, the program failed and therefore the result was not shown.**

5. How fast and memory efficient is this algorithm compared to other existing method?

**We provided more detailed information on the time and memory efficiency for RVD2 in Section 6 Paragraph 2.**

Minor comments:

1. Fig3 and 4 can be plots to visualize the comparison results.

**We have considered plotting the comparison results for Fig3 and Fig4. However, we haven’t found a type of plot which can convey the information as efficiently as the table.**

2. In page 6, it seems that positions are followed by mutations. It would be good to explicitly define the notations.

**We added a clarification for the notations in the Section 5.3.1 Paragraph 2 Sentence 2.**

3. Section 5.3 is too brief. Rigorous analysis on the real data should improve this paper.

**We added more rigorous analysis on the real data in Section 5.3. This sections includes analysis on the performance of RVD2 (section 5.3.1) and performance comparison with other algorithms including VarScan2, Strelka and muTect.**

----------------------- REVIEW 3 ---------------------

Major comments:

1. It was difficult for me to ascertain the methodological contribution. How is this method different from RVD and the other ones? What are the ideas which make it better? Currently, the model is presented without relating it to existing techniques. It was thus hard to know if there is any novelty in the approach.

**We have included a new paragraph (the last paragraph in Section 2) specially talking about the improvement of RVD2 upon RVD. There are two major improvements: the improvement on the model structure by assigning a prior on local error rate muj and a novel Bayesian Hypothesis Testing comparing to a frequentist normal z-test in RVD.**

2. The results showed superior performance for RVD2, but I believe that there are several limitations of the experiments. There was a synthetic dataset and a HCC1187 dataset. I could not draw many conclusions from the HCC1187 dataset about the accuracy of the method. The results were not validated (experimentally or computationally). A comparison to one other tool was made, but not to any other ones. I think some kind of validation is needed before one can draw conclusions about the method's accuracy on this dataset. I also think a comparison to other state-of-the-art tools should be made on this data.

**In order to validate the performance of RVD2 in the clinical dataset, we showed the comparison result and analysis to VarScan2, Strelka, and muTect in Section 5.3.2. Since the average read depth 90x does not satisfy the high read depth requirement for RVD, the program failed and therefore the result was not shown. Moreover, we provide read depth distribution for position called by all tools in Supplementary Table 1 where we can get an intuitive sense of the performance.**

3. Given the claimed accuracy of the algorithm at low MAF, it would have been really convincing to see it perform on real data with low MAF, if such a dataset exists. The HCC1187 dataset has 80% purity, which is high.

**We haven’t found an appropriate real dataset with low purity. However, we do plan to apply RVD2 once we such a dataset is available.**

4. One of the main challenges of these types of algorithm (especially for low MAF) is performing well given uneven sequencing coverage. While the model takes this into account (with the Mj variable), I do not think the experiments really test this. In the synthetic dataset, the genome is of length 400bp, and uneven coverage due to sequence content is unlikely to appear in such a short genome. In the HCC1187 dataset, the results are not validated and it’s hard to say anything about accuracy.

**We provided the analysis for uneven sequencing coverage and Mj in Appendix 3. We provided performance validation as mentioned previously.**

5. The effect of mapping error is also significant and is not properly tested. The 400bp synthetic genome is unlikely to have repeats, and the HCC1187 dataset is not validated.

**We have recognized that the effect of mapping error can be significant. However, we haven’t incorporated mapping error into our statistical model up to now. At current stage, RVD2 is only able to minimizing the effect of mapping error through normal-tumor paired data. We provided performance validation as mentioned previously.**

Minor comments:

1. I do not understand the meaning of the word "targeted" in the title. The data the paper deals with is whole-genome sequencing, so I am not sure what is being targeted. Also, the title does not refer to heterozygosity, which is strange given how the method is presented in the Intro.

**We used the word “targeted” to highlight the fact that RVD2 can handle low read depth, in contrast with RVD which requires high read depth to work properly. In order to be clearer, we removed the “target” and added “heterogeneous” in our title. Therefore, we have our new title as “RVD2: An ultra-sensitive variant detection model for low-depth heterogeneous next-generation sequencing data”.**

2. First section of section 2, you say "RVD uses a two stage..." Do you mean to say RVD2? Otherwise, it is not clear what the relevance of RVD is here.

**Yes, what we meant was RVD2. We have fixed the typo in the manuscript.**

3. In second paragraph of Section 2, J and N are not defined. I also do not understand where the n\_ji are coming from in step (2) of the model. I would think we would need to generate both n\_ji and r\_ji, but what is being generated is r\_ji | n\_ji.

**We added the definition of J and N in Section 2 Paragraph 2. J is the length of the DNA sequence and N is the number of technical replicates in our dataset. In the generative process, n\_ji is given and does not need to be generated. The notation r\_ji|n\_ji means to generate r\_ji give n\_ji. In order to eliminating confusion, we modified the notation r\_ji|n\_ji to r\_ji in the manuscript. Also, we provided a small example in Appendix 1 for easier understanding of the graphical model and generative process.**

4. How are the Mj's determined? Are they learned from the data and if yes, how? This should capture the dependency of coverage bias on the location, and it would be helpful to know exactly what these Mj's can capture, in a biological sense.

**Mj is determined using Method of Moment, as indicated in Section 3.1 and supplementary A. Mathematically Inversely proportional to the local error rate variance, Mj can capture the local experimental precision in biological sense.**

5. What are the running time / memory usage of the algorithm? This should be discussed.

**We provided more detailed information on the time and memory efficiency for RVD2 in Section 6 Paragraph 2.**