----------- REVIEW -----------

Dear Colleagues,

You described in this manuscript an extension to the recently published RVD method to call low frequency variants in sequencing datasets, despite variable sample mixtures and signal noise. This is indeed an important question as current benchmarks are showing glaring disparities between the leading variant callers when they are run on medical tissue for example.

The text itself is very rigorous and well written. I just have two questions:

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

- 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.

Yours sincerely,

Daniel Zerbino

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

PAPER: 41

TITLE: RVD2: An ultra-sensitive variant detection model for low-depth targeted next-generation sequencing data

AUTHORS: Yuting He and Patrick Flaherty

----------- REVIEW -----------

This paper proposes a novel method for detecting SNPs in heterogeneous samples. The proposed method built a graphical model, and then developed an inference algorithm based on Gibbs sampling. Further, they developed a hypothesis testing to distinguish between true mutations and random sequencing errors.

The presented idea is interesting and the results seem promising. However, the paper should be written more clearly. Also, it needs to show more evidence to support their results from real data analysis, and more convincing comparison results.

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)

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 is 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.

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

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.

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

Minor comments:

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

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

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

Overall, the idea presented in the paper and simulation results look promising, but I believe that issues mentioned above need to be addressed before this paper is published.

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

PAPER: 41

TITLE: RVD2: An ultra-sensitive variant detection model for low-depth targeted next-generation sequencing data

AUTHORS: Yuting He and Patrick Flaherty

----------- REVIEW -----------

He et al. describe a SNP caller for heterozygous samples with low coverage. The problem is an important one, as many SNP callers tend to assume that the sample has negligible heterozygocity. The paper is well written and clear.

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.

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.

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.

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 its hard to say anything about accuracy.

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.

Minor comments:

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 heterozygocity, which is strange given how the method is presented in the Intro.

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.

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.

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.

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