Reviewer #1:

1. Software implementing the model should be made available to enable independent assessment of the results

Thank you for pointing out the need for source code. **The source code in Python is now available in the bitbucket git repository under the “variational” branch**:

<https://bitbucket.org/flahertylab/rvd/branches/>

We have also updated the manuscript text to point readers to this location.

1. Although the software tools CRISP and SNVer are mentioned in the introduction, these tools were not included in the benchmark shown in the results section. SNPs identified with these tools should be included in the comparison taking into account that their models are a-priori more suitable for the described experimental setup compared to tools such as samtools or GATK, which are designed for SNV discovery on individual samples.

Thank you for identifying this important omission. **We have included the results of CRISP and SNVer in the comparison table (Table 2) in Section 3.2.3.** CRISP is able to detect variants at the true non-reference allele frequency (NRAF) of 10.0% and 100.0%, but the sensitivity decreases as the read depth increases at the 10.0% NRAF event. SNVer performs very well in detecting variants when the true NRAFs are 10.0% and 100.0%; it also shows high sensitivity and specificity when the NRAFs are 0.3% and 1.0% at the highest read depth. CRISP and SNVer both show very low sensitivity at 0.1% NRAF events.

1. In the description of the simulation experiments the authors should describe how sequencing error rates were simulated and which values were used to understand the signal to noise ratio of each experiment.

Our description of the experiments in subsection “Synthetic DNA Sequence Data” was confusing and we have clarified it in the manuscript. Briefly, these are in-vitro mixture experiments, not in-silico mixture experiments. We have found that mixing in-silico is a good first step, but does not capture the variation as well as actual in-vitro mixture data. So, we chemically synthesized two DNA fragments – one with 14 variant loci compared to the other. Then we combined these fragments at defined molar fractions to yield the fractions of interest: 0.1%, 0.3%, 1.0%, 10% and 100%. In order to capture the dependence of our method on sequencing depth, we downsampled the dataset in-silico using Picard.

Reviewer #2:

1. My only  rather minor comment relate to Tables 1,2; these tables show sensitivity/specificity results that are very hard to parse unless one of the variables (e.g. specificity) is kept constant across methods; alternatively, ROC curves can be employed.

Thank you for this important observation. **We have generated a ROC plot in Figure 2 under Section 3.2.1.** The results in the ROC curves are generated by varying parameter in the posterior distribution test. It shows that the performance improved with read depth and true mutant mixtures.