RESPONSE LETTER

Referee 1

-- Ref1.1 – General comments –

Reviewer' comment:

The authors have addressed all my concerns.

Author's response:

We appreciate referee 1's comments.

Referee 2

-- Ref2.1 – General comments –

Reviewer' comment:

The authors are right in stating that there is no other reference that implements a noncoding mutation burden analysis. The only one I know is Weinhold et al., (2014), and I also agree with the authors that a simple binomial test as applied in that reference is not good enough to correctly compute the mutation burden in noncoding regions. What I wonder is if the change from a binomial to a beta-binomial distribution is a good enough solution. Unfortunately, the controls provided in the new version don't seem to be enough to prove that, see comments below. Also, we have tried to run the software and we found many problems and unsatisfactory results in the only case we managed to run it (described below).

Overall I agree with the authors that it would be an important contribution to describe and provide a method that does the noncoding mutation burden analysis correctly. I am not convinced that LARVA does it well, at least in its current version, based on our test on running the software (see below) and on the description in the manuscript.

Author's response:

We thank the reviewer for agreeing with contributions. We have further investigated the false positive and negative problems through simulation and permutation studies. The reviewer's concerns were answered point by point here.

1. "What I wonder is if the change from a binomial to a beta-binomial distribution is a good enough solution"

We added more discussions about importance of covariate correction (paragraph 5 in discussion and section 5 in Text S1) and how to interpret the usage of the beta-binomial model (paragraph 4 in discussion). We actually also tried other distributions, like the negative binomial, and Poisson inverse Gaussian. The performance of these distributions is similar to that of the beta-binomial distribution. Eventually we selected beta-binomial because of its immediate interpretability.

2. "we have tried to run the software and we found many problems and unsatisfactory results in the only case we managed to run it"

Lucas Lochovsky 7/11/2015 2:58 PM
Deleted: covariant

Lucas Lochovsky 7/11/2015 2:58 PM
Deleted: y

Lucas Lochovsky 7/11/2015 2:58 PM
Deleted: performed

Lucas Lochovsky 7/11/2015 2:58 PM
Deleted: in a

Lucas Lochovsky 7/11/2015 2:59 PM
Deleted: way as

Lucas Lochovsky 7/10/2015 10:48 PM
Deleted: their

We have addressed the software issues raised by the reviewer. Detailed answers are provided in section Ref2.5.

Lucas Lochovsky 7/11/2015 2:59 PM
Deleted: were

Lucas Lochovsky 7/10/2015 10:48 PM
Deleted: response in

-- Ref2.2 – False positive and false negative rate –

Reviewer' comment:

```
AUTHOR'S RESPONSE
We emphasize our contribution in the following listed points.
1. We are among the first to implement the somatic burden test with
overdispersion control, which is specifically designed for noncoding
somatic variant analysis.

MY NEW COMMENT
I agree with that. It is important not only to be among the first but
more importantly to make sure that the test is correct, give a good
control of false positives and false negatives, and provide a code that
users can run.

AUTHOR'S RESPONSE
2. We release a convenient annotation resource for the whole community
by gathering all the noncoding regulatory regions from more than 122
experiments from the ENCODE project. Notably, this data has never been
collected in one place before, which will greatly facilitate subsequent
research.

3. Our released noncoding regulatory element corpus provides a natural
and meaningful solution about how to pool biologically relevant regions
to perform the mutation burden test. We do not have to rely on the bin
procedure, which is a relatively ad-hoc method.

4. Once highly mutated regions are detected in a certain cancer type,
users can immediately understand the functions of these regions.

MY NEW COMMENT
I agree with authors that 2, 3 and 4 are useful additional resources
provided with the code of LARVA, however the first and more important
think is that authors convince that LARVA is able to detect noncoding
recurrently mutated drivers, which I understand from the description of
the paper it is the main aim, with an acceptable rate of false
positives and false negatives. This is not clear in this version of the
software and manuscript.
```

Author's response:

We thank the reviewer for agreeing with our contribution. We added some simulation and permutation studies for these questions.

1. Simulation studies

- For all 2.5kb bins on the genome, remove those intersecting gap regions and blacklist regions. For the remaining 1,139,452 bins, count the variants.
- Use negative binomial regression to build the mutation model $NB(\mu_i, \sigma_i)$ for the i^{th} bin by correcting replication timing, GC content, and chromatin status.
- Simulate the variant counts in 1,139,452 bins using local $NB(\mu_i, \sigma_i)$

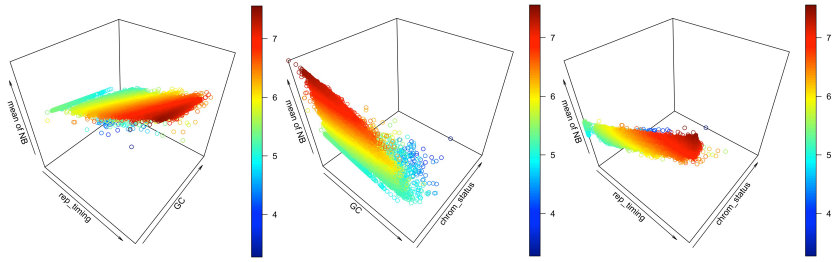
Lucas Lochovsky 7/10/2015 10:49 PM
Deleted: with

Lucas Lochovsky 7/10/2015 10:49 PM
Deleted:

- Randomly selected 100 bins as the true signal, replace the variant count as the top 1e-4 quantile of local NB$(10 * \mu_i, \sigma_i)$

Figure R1 gives the 3d plot on how the average mutation rate changes with these three covariates.

Figure R 1. 3D plot of background model with three covariates



Then we used beta-binomial and binomial model by only correcting the replication timing effect. Our method gives 108 positives—including all true positives—and the binomial test gives 48,505 positives (BH adjusted P<0.05).

2. Permutation studies

For each variant in a set of whole genome sequencing data, find a new position in a 25kb neighboring region (12.5k and 12.5k up and downstream each). Then we tested all the noncoding regions on the original and permuted data set. Since the permuted size 25kb is relatively large as compared to the test region, a better method is supposed to give less or even no positives on the permuted data set. The P values were given in the updated Fig. S13 in Text S1. From the Q-Q plots of the P values, it can be seen that LARVA yields fewer positives in larger regions, such as DHS, TF, Promoter and UTRs. In very small regions such as ultra sensitive and ultra conserved regions, the two methods gives similar performance.

Lucas Lochovsky 7/10/2015 10:50 PM
Deleted: positives

Lucas Lochovsky 7/10/2015 10:50 PM
Deleted: ,

Lucas Lochovsky 7/10/2015 10:51 PM
Deleted: provides

Lucas Lochovsky 7/10/2015 10:51 PM
Deleted: less number of

--Ref2.3 – P-values for all genes –

Reviewer' comment:

```
Finding only 7 significantly mutated coding genes analyzing 5032 tumors
is a surprising low number. I agree that 6759 significantly mutated
genes with the binomial test is a not an acceptable number of genes,
surely full of false positives. It would be useful if authors provide a
supplementary table with the obtained pvalue per gene, not only for the
7 genes claimed as highly mutated by LARVA.
```

Author's response:

We thank the reviewer for this comment.

1). As requested, we provide the P values on our website

2). We agree that 7 is a rather low number but this is understandable since our method is not optimized in the coding region analysis. We clearly mentioned this point in

updated manuscript and explained why (the second last paragraph in the Discussion section in the manuscript and first paragraph in section 3 in Text S1).

-- Ref2.4 – QQ plots –

Reviewer' comment:

> QQ plots should be in - log10 scale to be able to see in detail the most important part of the plot, which correspond to the significant regions. With the QQ plot provided it is not clear if the distribution of pvalues is correct. Authors could use this code for example: http://www.broadinstitute.org/files/shared/diabetes/scandinavs/qqplot.R

Author's response:

We thank the reviewer for this comment. We have updated the QQ plots in Fig. S12 and Fig. S13 in Text S1 in accordance with these suggestions.

-- Ref2.5 – Software errors I –

Reviewer' comment:

> Since I wasn't convinced myself of the validity of the method by reading the new version of the manuscript I thought the best would be to run the software ourself. We decided to run LARVA on a pancancer dataset retrieved from tumorportal (http://www.tumorportal.org/load/data/per_ttype_mafs/PanCan.maf). Unfortunately we were not able to get any results as the program halted the execution raising errors.
>
> We first tried to analyze the coding regions of the pancancer dataset. The program kept running for more than 100 hours (> 4 days) and eventually halted raising an R error.
>
> Error in if (any(mu <= 0) | any(mu >= 1)) stop(paste("mu must be between 0 and 1 ", :
>
> missing value where TRUE/FALSE needed
>
> Calls: pval_varying_length -> pBB
>
> Execution halted

Author's response:

We thank the reviewer for bringing this to our attention. We have addressed the long running time by profiling our code, and optimizing the computations in portions of the code where the running time did not scale well with the size of the input. We have released revised code along with our revised manuscript.

Furthermore, we have migrated our R codebase into C++, giving us more direct control over the source code. Our new code is not prone to the error the reviewer encountered.

-- Ref2.6 – Software Errors II –

Reviewer' comment:

> We next tried to run LARVA with a dataset of 505 tumor whole-genomes

Lucas Lochovsky 7/10/2015 10:51 PM
Deleted: but I

Lucas Lochovsky 7/10/2015 10:52 PM
Deleted: discussion

Lucas Lochovsky 7/10/2015 10:52 PM
Formatted: Left, None, Space Before: 0 pt, Don't keep with next, Don't keep lines together

Lucas Lochovsky 7/10/2015 10:52 PM
Formatted: Left

Lucas Lochovsky 7/10/2015 10:52 PM
Formatted: Left, None, Space Before: 0 pt, Don't keep with next, Don't keep lines together

Lucas Lochovsky 7/10/2015 10:53 PM
Formatted: Left, None, Space Before: 0 pt, Don't keep with next, Don't keep lines together

across 14 cancer types as reported in Fredriksson et al., 2014 in promoters and ultra-sensitive regions. For both promoters and ultra-sensitive regions we used the annotations present in the folder data/annotations/ of LARVA. The program didn't run successfully on promoters and raised an error after approx. 12 hours. Following is the trace of the error:

```
Error in d$p.bbd.cor[d$p.bbd.cor <= 0] = rep(d$p.tiny, sum(d$p.bbd.cor
<=  :
   replacement has length zero
Execution halted
```

Lucas Lochovsky 7/10/2015 10:53 PM
Formatted: Left

Lucas Lochovsky 7/10/2015 10:53 PM
Formatted: Left, None, Space Before: 0 pt, Don't keep with next, Don't keep lines together

Author's response:

We thank the reviewer for bring this to our attention. We have determined that this error can occur in rare boundary conditions in our R code. We have migrated our R codebase into C++, and now have more direct control over the functioning of our code. Our new code handles these conditions properly.

-- Ref2.7 – Software P-value Output –

Reviewer' comment:

We finally managed to run LARVA with this dataset in ultra-sensitive regions. In this case the program performed the analysis quickly. However, when we check the files with the results we found cases, with the exception of 'p.bbd.cor.adj', where the pvalues were greater than 1. How can this be possible? Following are the maximum values of each pvalue type:

p.bbd	3.488000
p.binomial	16.425000
p.bbd.cor	3.286000
p.binomial.cor	20.596000
p.bbd.adj	1.148000
p.bbd.cor.adj	0.357000
p.binomial.adj	14.031000
p.binomial.cor.adj	17.464000

Author's response:

We thank the reviewer for bringing this to our attention. We didn't clearly mention in our software documentation that these numbers are, in fact, –log10-transformed p-values, hence the observed output is correct. This has been rectified in the current version's documentation.

Lucas Lochovsky 7/10/2015 10:57 PM
Deleted: is numerical output is

Lucas Lochovsky 7/10/2015 10:57 PM
Deleted:

Lucas Lochovsky 7/10/2015 10:57 PM
Deleted: the

Lucas Lochovsky 7/10/2015 10:54 PM
Deleted: already

Lucas Lochovsky 7/10/2015 10:55 PM
Formatted: Left, None, Space Before: 0 pt, Don't keep with next, Don't keep lines together

-- Ref2.8 – Software P-value QQ Plots –

Reviewer Comment	After filtering the results for regions that overlapped genes or pseudogenes and for regions without mutations, we did QQplots as follow: we discarded pvalues > 1 (considering them wrong) and we plot on the y axis the -log10 of the sorted observed pvalues and on the x axis the -log10 of a uniform distribution of expected pvalues between 0 and 1. The QQplots were generated by using the code provided

	here:http://www.broadinstitute.org/files/shared/diabetes/scandinavs/qqplot.R) The resulting plots showed that the both the 'pbb' pvalues distributions (p.ddb and p.ddb.cor, top row of the figure) are deflated respect to a perfect correlation between observed and expected pvalues (red diagonal line) and thus the methods are finding less significant genes that what expected by the null model. On the other hand the binomial method (bottom row of the figure) is somehow inflated respect to the red diagonal. While the binomial method is likely to find a number of false positive candidates, the method proposed by the authors is likely to miss many true positive candidates.

Author's response:

We thank the reviewer for the careful checking of Q-Q plots. As mentioned in section Ref2.7, our provided P values are already log transformed P values. Hence taking the –log10(P) again does not reflect the real P value here. We plotted the Q-Q plots as suggested by the reviewer in Fig. S13. Our P values follow the uniform distribution line.

Lucas Lochovsky 7/10/2015 10:56 PM
Deleted: the P value problem

Lucas Lochovsky 7/10/2015 10:56 PM
Deleted: the

Lucas Lochovsky 7/10/2015 10:56 PM
Deleted: to take