Yale University

Bass 432A, 266 Whitney Ave.
PO Box 208114
New Haven, CT 06520-8114

203 432 6105
360 838 7861 (fax)
Mark.Gerstein@yale.edu
http://GersteinLab.org

Dear Editor,

We thank you for the opportunity to respond to referee comments and submit a revised manuscript. We have now addressed all the referee concerns: we provide a brief overview of our responses below, followed by a point-by-point response. We hope you find that the methods and resources presented in our manuscript contribute to the investigation of noncoding variants in cancer research.

Yours sincerely,

Mark Gerstein

-- Overall response to referee comments --

We thank all the referees for their insightful comments and suggestions. We have made several major and minor revisions to address the comments, which we believe clearly address the reviewers' confusions and significantly strengthen the manuscript.

The main contribution of our LARVA method is not only to improve extend the current state-of-the-art approach to driver candidates discovery in noncoding regions by properly handling overdispersion in the mutation counts data, but also provide a valuable resource to pinpoint the functions of these regions to the best of our effort. In response to comments from both referees, we further investigated our performance comparison in the coding regions by applying LARVA on a total of 5032 exome sequencing samples in detail.

Below we list the response to all comments in a point-by-point fashion. We label each comment as 'Major' or 'Minor' for major and minor comments, respectively.

Referee 1:

Referee general comments:

Reviewer comment	*In the manuscript "LARVA: an integrative framework for Large-scale Analysis of Recurrent Variants in noncoding Annotations", Lochovsky et al. developed an innovative framework to estimate the mutation load of noncoding regions from whole genome sequencing data. They modeled mutation count with a beta-binomial distribution to account for the heterogeneous mutation rates across the genome, and demonstrated that beta-binomial distribution fits the data better than the binomial distribution, and therefore lead to much less false positive hits.* *The manuscript is well written and easy to follow. The description of the methods and data sources is very clear. All calculations and use of statistics throughout the manuscript were properly carried out.*
Author Response	We appreciate the comments of the reviewers.

-- Minor questions/suggestions --

Referee minor comment 1:

Reviewer comment	*Does the different sequencing depth/coverage of individual samples (and even at different loci within the same sample) affect the analysis results?*
Author Response	We thank the reviewers for pointing out this important issue. Sequencing depth/coverage for the individual samples obviously affect the quality of variant calling and potentially affect any downstream analysis in cancer research, not just for LARVA. It's essentially like a garbage-in-garbage-out problem. That is precisely why uniform variant calling is highly recommended, and is under analysis by some working groups, like PCAWG. We have mentioned this caveat in our discussion section. It is our intention that as more and more uniformly processed WGS data is released, we will immediately incorporate such information into our method.
Excerpt from Revised Manuscript	We added a new paragraph in the discussion section in the updated manuscript [Page 12]. "One factor that may affect LARVA's performance is the uneven sequencing depth of the WGS experiments currently available. This may result in undetected variants in regions that are insufficiently covered, or not covered at all. Our plan is to incorporate additional, uniformly processed WGS data into LARVA as it becomes available in the future. Groups such as the TCGA's Pan-Cancer Analysis Working Groups (PCAWG) are currently working to produce such data for higher quality downstream analyses."

Jing Zhang 4/20/2015 7:56 PM
Deleted: issue for LARVA.

Jing Zhang 4/20/2015 7:55 PM
Deleted: would

Jing Zhang 4/20/2015 7:55 PM
Deleted: potentially

Jing Zhang 4/20/2015 7:55 PM
Deleted: , which might generate both false positives and false negatives, especially when analyzing samples from different labs

Jing Zhang 4/20/2015 9:14 PM
Comment [1]: [[JZ2MG:]] I do feel that this is a too strong argument…

Lucas Lochovsky 4/20/2015 2:56 PM
Deleted: the exact reason

Lucas Lochovsky 4/20/2015 2:56 PM
Deleted: being

Lucas Lochovsky 4/20/2015 2:57 PM
Deleted: being analyzed

Lucas Lochovsky 4/20/2015 2:58 PM
Deleted: However, currently not many uniformly processed whole genome sequencing (WGS) samples have been released for different cancer types, hence it is difficult for us to gather the sequencing depth information at each position.

Lucas Lochovsky 4/20/2015 2:58 PM
Deleted: problem

Referee minor comment 2:

Reviewer comment	Supplementary table 2 is missing (I can't find a separate file with the table).
Author response	This problem has been addressed. We thank the reviewers for pointing this out.

Referee 2:

Referee general comments:

Reviewer comment	Lochovsky et al describe a method (LARVA) to identify non-coding regions that accumulate tumor somatic mutations more than expected, which could point to driver mutations. They compare their method to a simple binomial test which assumes equal probability of mutations across the genomes and instead introduce a beta-binomial approach, which they claim can better control false positives. They also take into account replication timing to control for different mutation rates in different genomic regions. All the ideas presented in this article have already been proposed before, including the fact that mutation rates are variable across the genome and that this should be accounted in a proper statistical test to find significantly mutated regions. Using a beta-binomial distribution and comparing it to a binomial test doesn't seem to me a significant improvement over existing knowledge or methodology.
Author Response	We thank the reviewer for this comment. We disagree with the reviewer about the novelty of LARVA. We challenge the reviewer to point out the specific reference that actually implements the noncoding mutation burden analysis. Currently, there have been extensive investigations of mutation burden in the coding regions, such as Lawrence *et al.* (2013). However, the first large-scale analysis of noncoding driver discovery was published in Weinhold *et al.* (2014), where a simple binomial test was used for p-value evaluation, and incomplete interpretation of noncoding regions was provided. After its publication for only 6 months, it has been cited 9 times (11/1/2014-4/15/2015), and provoked extensive discussions in the cancer research community. Other scientists may realize that simple binomial test might not be the best choice, but to our current knowledge there is no public software that handles the overdispersion specifically designed for the noncoding variant analysis. ***We emphasize our contribution in the following listed points.*** 1. We are among the first to implement the somatic burden

Jing Zhang 4/20/2015 7:58 PM
Deleted: .

Jing Zhang 4/20/2015 7:59 PM
Deleted: in cancer research

Jing Zhang 4/20/2015 8:00 PM
Deleted: , and they have successfully identified driver mutations in those regions

Jing Zhang 4/20/2015 8:28 PM
Deleted: not many whole genome noncoding results have been published due to three main difficulties: 1) The background mutation rate is not as easy to derive in noncoding regions compared to coding regions, where the synonymous sites may serve as a natural and biologically meaningful control; 2) the poor quality of interpretation of noncoding results due to the currently limited understanding of noncoding regions; 3) in coding regions, genes are the natural units to gather the variants for the test, but it's still a debatable question how to pool the variants to perform the same test in the noncoding regions.

Jing Zhang 4/20/2015 8:27 PM
Deleted: ... [1]

Jing Zhang 4/20/2015 8:27 PM
Deleted: across the whole genome

Jing Zhang 4/20/2015 9:16 PM
Formatted: Font:Bold, Italic, Underline

	test with overdispersion control, which is specifically designed for noncoding somatic variant analysis. 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. 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 this region. To emphasize our contributions, we have added a new paragragh in the discussion section (highlighted in the updated manuscript) for clarity. For this reviewer's other concerns, we provided our responses in a point-by-point layout in the following section.
Excerpt from Revised Manuscript	We added a new paragraph in the discussion section in the updated manuscript [Page 12]. "LARVA's complete design, in terms of both software and provided data, offers a new, convenient processing engine for whole genome mutation burden tests. Exome burden tests may be conducted with naturally defined regions—genes—to test for mutation burden. Whole genome burden tests, however, are hindered by the fact that many noncoding functional regions are poorly defined, if at all. LARVA unifies multiple noncoding annotation sets derived from a set of uniformly processed pipelines and experiments. These annotations are tested for mutation burden, and make it easy to understand the functional significance of each highly mutated region."

Jing Zhang 4/20/2015 8:29 PM
Deleted: All the provided regions were carefully obtained through uniformly processed pipelines from real experiments.

Jing Zhang 4/20/2015 8:29 PM
Deleted: This may prove to be beneficial for the drug discovery process.

Lucas Lochovsky 4/20/2015 4:15 PM
Deleted: point

Jing Zhang 4/20/2015 9:17 PM
Deleted: added two sentences

Jing Zhang 4/20/2015 8:30 PM
Deleted: If this reviewer finds our explanation unsatisfactory, we challenge this reviewer to point out a specific example from previously published literature that addresses the same issues that LARVA addresses.

-- Major questions/suggestions --

Referee major comment 1:

Reviewer comment	To address that, it would be desirable to test the method in protein coding genes to demonstrate that it is able to finds well known cancer genes and it is not selecting too many false positives.
Author Response	We thank the reviewers for pointing this out and we agree that it's a good idea to test our method on the coding regions. Although the accurate false positive and false negative rates are difficult to estimate, it does give us good sense of performance calibration. In the updated manuscript, we mentioned the coding analysis in the result section and more details in Text S1. As suggested by the reviewer, we applied our method to the coding regions for the sake of comparison with the binomial test. We downloaded the whole exome sequencing data from the TCGA website, which incorporates 20 cancer types and 5032 samples in total. We first used all the coding transcripts in Gencode V19 annotation to define the gene regions. In total, 3,547,350 variants were found in these regions with the average mutation rate as 0.0141 for the pooled samples. As a result, 6 out of 7 genes claimed as highly mutated by LARVA were clearly documented to be associated with some types of cancer (Table S3 in Text S1). On the other hand, the p-values for the binomial test method were heavily inflated. After p-value adjustment, there are 6759 out of 18,826 genes, roughly 35.90%, with p-value less than 0.05. It is very unlikely that all such genes are associated with cancer. This result shows that LARVA may effectively find meaningful results in the coding regions. In terms of the real false positive and negative rate estimation, currently there is no golden standard dataset for a benchmark comparison, so it is difficult for us to obtain. We added some sentences in the discussion section in the updated manuscript (also highlighted).
Excerpt from Revised Manuscript	We added a paragraph in the result section in updated manuscript and a new section 3 (Coding Region Mutation Burden Analysis) in the updated Text S1. Details about we performed the coding region analysis were given in section 3 Text S1

Jing Zhang 4/20/2015 8:32 PM
Deleted: does give us a sense of how our proposed method works

Jing Zhang 4/20/2015 8:44 PM
Deleted: did

Jing Zhang 4/20/2015 8:44 PM
Deleted: y

Jing Zhang 4/20/2015 8:38 PM
Formatted: Font:(Default) Arial

Jing Zhang 4/20/2015 8:38 PM
Formatted: Font:(Default) Arial

Jing Zhang 4/20/2015 8:43 PM
Deleted: The detailed data is given in Figure R 1.
Adjusted P value ... [2]

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Arial

Unknown
Formatted: Font:(Default) Helvetica

Jing Zhang 4/20/2015 9:17 PM
Formatted: Justified, Line spacing: single, Tabs:Not at 7.62 cm + 15.24 cm

Jing Zhang 4/20/2015 8:43 PM
Deleted: ... [3]

Jing Zhang 4/20/2015 9:17 PM
Formatted: List Paragraph, Justified, Space After: 12 pt

Jing Zhang 4/20/2015 8:47 PM
Deleted: even in the coding regions. The discovery of meaningful genes depends on lots of varying factors, including the samples used, sequencing depth and read coverage, variant calling methods, and lots of covariate correction in the coding regions. These factors are out of our control in the current LARVA version.

Jing Zhang 4/20/2015 9:05 PM
Formatted: Font:9 pt

Jing Zhang 4/20/2015 9:04 PM
Deleted: [Page 10] ... [4]

Referee major comment 2:

Reviewer comment	It would be necessary also to provide evidence that the obtained pvalues from their test follow a uniform distribution, with few exceptions that would be the regions with driver mutations.
Author Response	We thank the reviewers for pointing out this important issue. The QQ plots of p-values are provided in the following figures.

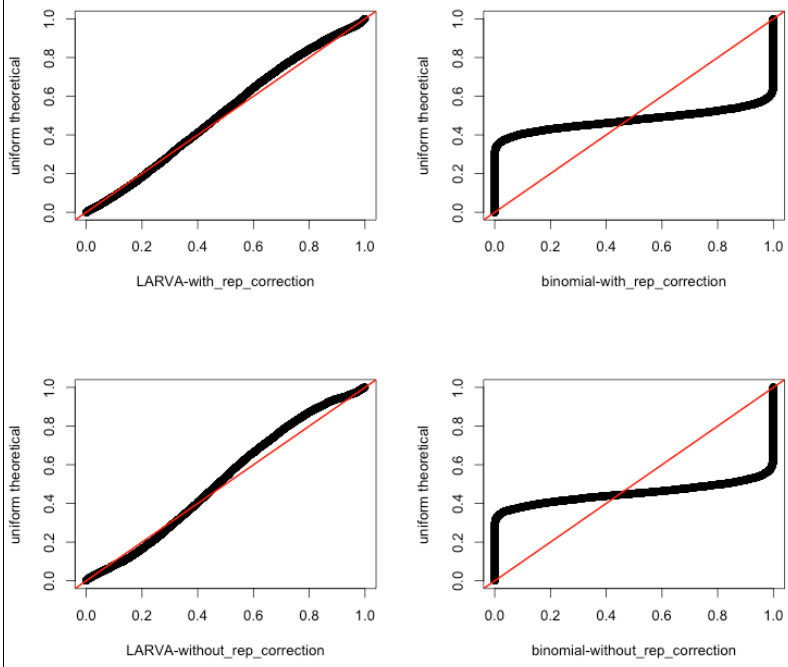

Figure R 1: coding region p-values vs. theoretical. The red line is the diagonal line.

Lucas Lochovsky 4/20/2015 3:15 PM
Deleted: 5

Jing Zhang 4/20/2015 9:02 PM
Deleted: 4

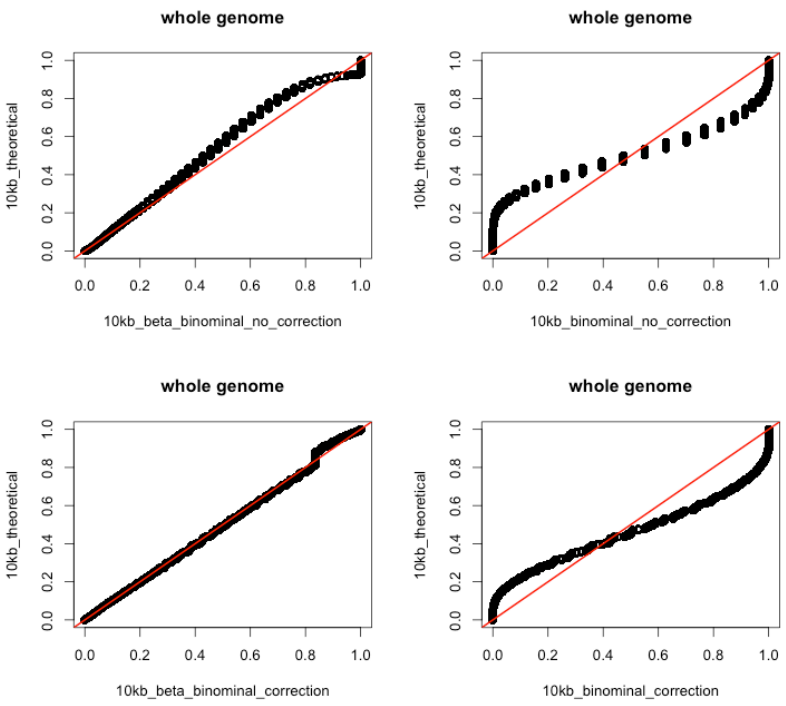

Figure R 2: QQplot of the pvalues from genome wide 1kb bin

In Figure R 1, it is shown that the p-values from binomial test severely violate the uniform distribution, which is consistent with its bad fitting of the data. On the other hand, the p-values from the LARVA method (figures on the left hand side) roughly follow a uniform distribution. It is worth mentioning that after replication timing correction, the p-values from LARVA method have improved concordance with the theoretical distribution, indicating the importance of correction. We also provided the QQ plot of the 10kb bin regions from the whole genome sequencing analysis. Even at this resolution, we observed an improved p-value distribution in LARVA vs. binomial test. The discrete dots in

Lucas Lochovsky 4/20/2015 3:15 PM
Deleted: 6

Jing Zhang 4/20/2015 9:02 PM
Deleted: 5

Lucas Lochovsky 4/20/2015 3:17 PM
Deleted: Figure R 5

Lucas Lochovsky 4/20/2015 9:55 PM
Deleted: Figure R

Jing Zhang 4/20/2015 9:02 PM
Deleted: 4

Jing Zhang 4/20/2015 9:02 PM
Deleted: violates

Lucas Lochovsky 4/20/2015 3:17 PM
Deleted: the

Lucas Lochovsky 4/20/2015 3:17 PM
Deleted: s

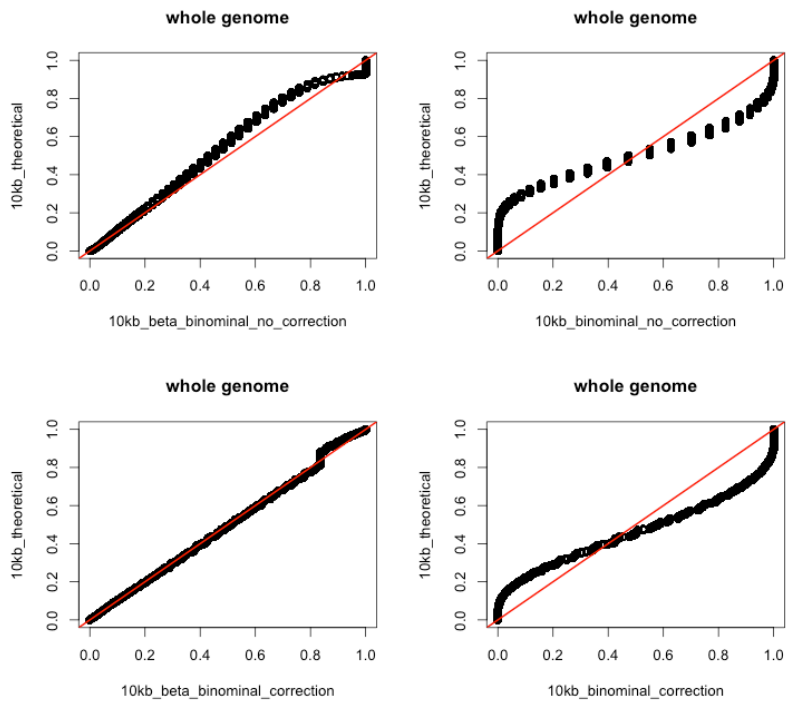

Figure R 2 is due to the limited number of genomes (785 WGS data). Only 137 unique variants counts values were observed in the 10kb region analysis. Similar to the coding region analysis, replication-timing correction improves the p-value distribution.

Excerpt from Revised Manuscript	These two figures were given in Fig. S12 and Fig. S13 in Text S1.

Lucas Lochovsky 4/20/2015 3:17 PM
Deleted: Figure R 6

Lucas Lochovsky 4/20/2015 9:55 PM
Deleted: Figure R

Jing Zhang 4/20/2015 9:03 PM
Deleted: 5

Jing Zhang 4/20/2015 9:03 PM
Deleted: replication timing

Jing Zhang 4/20/2015 9:02 PM
Deleted: [Page 10]

Jing Zhang 4/20/2015 9:06 PM
Deleted: ... [5]