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:

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's Response:

We appreciate the comments of the reviewers.

-- Minor questions/suggestions --

Referee minor comment 1:

Does the different sequencing depth/coverage of individual samples (and even at different loci within the same sample) affect the analysis results?

Author's Response:

We thank the reviewers for pointing out this important issue for LARVA. Sequencing depth/coverage for the individual samples would potentially affect the quality of variant calling, which might generate both false positives and false negatives, especially when analyzing samples from different labs. That is the exact reason why uniform variant calling is being highly recommended and is being analyzed by some working groups, like PCAWG. 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. We have mentioned this problem 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.

Jing Zhang 4/15/2015 10:23 AM
Deleted: greatly

Jing Zhang 4/17/2015 1:28 PM
Deleted: could potentially

Jing Zhang 4/15/2015 10:24 AM
Deleted: For example, in a TF binding site that is only 50% covered by sequencing reads, it is impossible to observe the variant calls in the uncovered half. Hence, it is highly possible that LARVA cannot detect strong signals in these instances. ... [1]

Jing Zhang 4/15/2015 10:25 AM
Deleted: added

Referee minor comment 2:

Supplementary table 2 is missing (I can't find a separate file with the table).

Author's Response:

This problem has been addressed. We thank the reviewers for pointing this out.

Referee 2:

Referee general comments:

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's Response:

We thank the reviewer for this comment. Currently, there have been extensive investigations of mutation burden in the coding regions in cancer research, such as Lawrence *et al.* (2013), and they have successfully identified driver mutations in those regions. However, 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.

The first large-scale analysis of noncoding driver discovery across the whole genome 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 test with overdispersion control, which is specifically designed for noncoding somatic variant analysis.
2. We release a convenient resource for the whole community by gathering all the noncoding regulatory regions from more than 122 experiments from the ENCODE project. All the provided regions were carefully obtained through uniformly processed pipelines from real experiments.
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. This may prove to be beneficial for the drug discovery process.

WE SHALL THE REFEREE TO POINT TO PUBLISHED

To emphasize our point, we added two sentences 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.

Lucas Lochovsky 4/17/2015 4:22 PM
Deleted: Yet a

Lucas Lochovsky 4/17/2015 4:22 PM
Deleted: already

Lucas Lochovsky 4/17/2015 4:21 PM
Deleted: for

Lucas Lochovsky 4/17/2015 4:21 PM
Deleted: it also arouses

Lucas Lochovsky 4/17/2015 4:22 PM
Deleted: till now

Lucas Lochovsky 4/17/2015 4:22 PM
Deleted: -

Jing Zhang 4/15/2015 10:49 AM
Comment [1]: Don't want to make it too aggressive. What do you think?

Lucas Lochovsky 4/17/2015 4:23 PM
Deleted: who

Lucas Lochovsky 4/17/2015 4:23 PM
Deleted: s

Lucas Lochovsky 4/17/2015 4:24 PM
Deleted: e

Jing Zhang 4/15/2015 10:45 AM
Deleted: .

-- Major questions/suggestions --

Referee major comment 1:

> *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's Response:

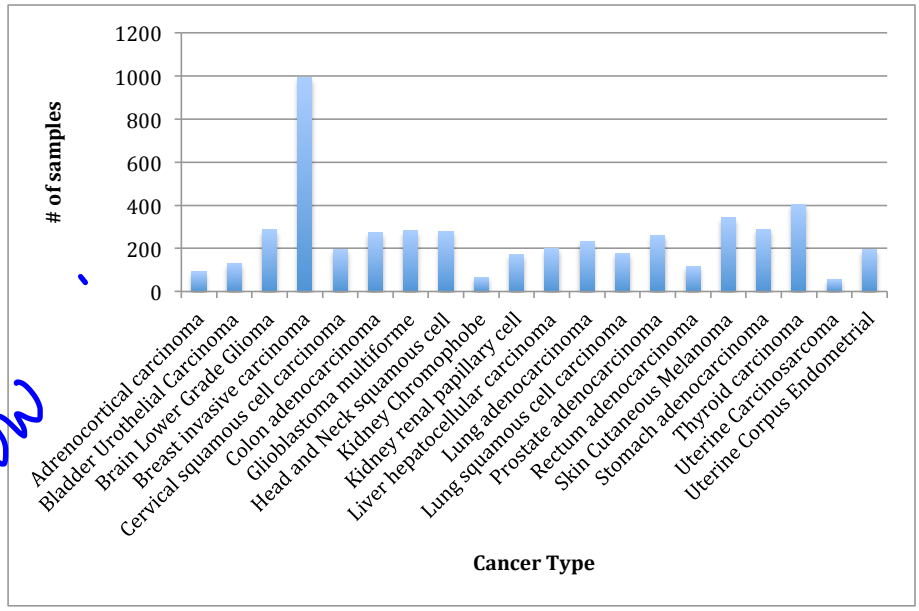

Figure R 1: TCGA Whole Exome Sequencing samples by cancer types.

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 a sense of how our proposed method works.

As suggested by the reviewer, we did apply 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. The detailed data is given in Figure R 1.

Lucas Lochovsky 4/17/2015 4:25 PM
Deleted: to

Jing Zhang 4/15/2015 10:53 AM
Deleted: First, we want to emphasize that LARVA is not optimally designed to analyze the variants in the coding region, although it is possible for LARVA to take coding regions as input. The reason is that coding regions have very biological meaningful background mutation rates that can be used, and naturally researchers selected genes as a unit to give a p-value. Furthermore, other well-known mutation confounders, such as expression level, can be used for more rigorous false positive and false negative control. On the other hand, LARVA's main strength is: 1) find meaningful test units (such as a TF, enhancer, DHS region); 2) false positive and false negative rate control by better overdispersion and replication timing control; 3) immediate function interpretation of the discovered regions.

Jing Zhang 4/15/2015 10:54 AM
Deleted: requested

Since the target of the Whole Exome Sequencing data is whole coding exons, we first picked all the protein coding transcripts in Gencode V19 annotation by requiring that the transcript be protein coding, and the knowledge of the protein coding region should be complete. We then merged all these transcripts for each gene as shown in Figure R 2.

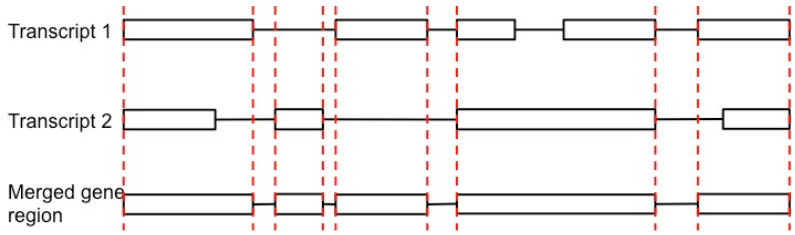

Figure R 2: details of gene region definition. Note that only coding transcripts were used for the Whole Exome Sequencing data analysis.

In the end, we generated regions for 19,822 genes in a total of 252,356,877 nucleotides. The gene length distribution is given in Figure R 3.

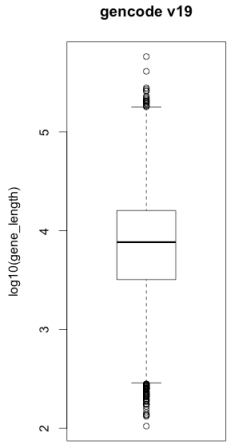

Figure R 3: distribution of the gene length

The total number of mutations falling into the merged gene regions is 3,547,350, and the average mutation rate is 0.0141 for the pooled samples. As with the noncoding regions, we also found huge mutation heterogeneity in the coding regions (as shown in Figure R 4).

Lucas Lochovsky 4/7/2015 11:24 AM
Deleted: .

Jing Zhang 4/17/2015 1:32 PM
Deleted: one

Lucas Lochovsky 4/17/2015 4:25 PM
Deleted: should

Lucas Lochovsky 4/17/2015 4:51 PM
Deleted: Figure R 2

Jing Zhang 4/7/2015 1:19 PM
Formatted: Caption, Centered, Space After: 0 pt, Line spacing: single

Lucas Lochovsky 4/17/2015 4:51 PM
Deleted: 22

Jing Zhang 4/7/2015 1:19 PM
Deleted: .

Lucas Lochovsky 4/7/2015 12:11 PM
Deleted: .

Lucas Lochovsky 4/17/2015 4:26 PM
Deleted: was

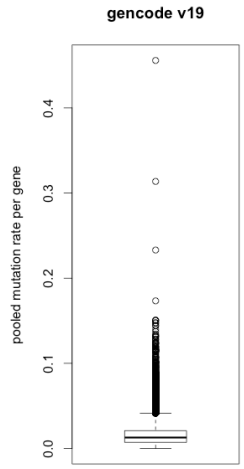

Figure R 4: distribution of the pooled mutation rates

We removed the genes with length less than the bottom 5% of gene lengths for higher annotation confidence, and then compared the performance of LARVA and the binomial test. After p-value adjustment, LARVA found 7 genes with higher mutation burden (results in Table R1).

Lucas Lochovsky 4/17/2015 4:36 PM
Deleted: that are potentially under

Table R1

Gene	Adjusted P value	Simple annotation	Supporting Reference
TP53	0	Well-known oncogene	PMID:20182602
BRAF	2.332696e-04	B-Raf proto-oncogene	http://ghr.nlm.nih.gov/gene/BRAF
KRTAP4-11	3.323269e-03	Unknown	
IDH1	3.323269e-03	Glioblastomas, astrocytomas, oligodendroglial tumors	PMID:19435942
FRG1B	4.860527e-03	lineage-specific mutation patterns in many cancer types	PMID: 24465236
CDKN2A	9.842880e-03	pancreatic cancer	PMID: 21150883

PRSS1	2.341413e-02	pancreatic cancer	PMID:22379635

Out of the 7 genes found by LARVA, we searched each gene's documented reference support online to check its association with some cancer types. Except for KRTAP4-11, we found all the remaining genes to be clearly documented with some cancer association. Note that we only reported one PubmedID per gene, even if there are many more supporting references. It effectively demonstrates that LARVA is capable of finding meaningful results on protein coding regions. 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.

We want to emphasize that LARVA is not optimally designed to analyze the variants in the coding region. Coding regions have very biological meaningful background mutation rates, and researchers naturally selected genes as a unit to test for mutation burden. Furthermore, other well-known mutation confounders, such as expression level, can be used for more rigorous false positive and false negative control. On the other hand, LARVA's main strength is: 1) finding meaningful noncoding test units (such as a TF, enhancer, DHS region); 2) false positive and false negative rate control by better overdispersion and replication timing control; 3) immediate function interpretation of the discovered regions.

In terms of the real false positive and negative rate estimation, currently there is no gold standard dataset for a benchmark comparison 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. We added some sentences in the discussion section in the updated manuscript (also highlighted).

Jing Zhang 4/17/2015 1:34 PM
Deleted: only

Lucas Lochovsky 4/17/2015 4:27 PM
Deleted: Besides

Lucas Lochovsky 4/17/2015 4:29 PM
Deleted: , w

Lucas Lochovsky 4/17/2015 4:27 PM
Deleted: The reason is that c

Lucas Lochovsky 4/17/2015 4:27 PM
Deleted: that can be used

Lucas Lochovsky 4/17/2015 4:27 PM
Deleted: naturally

Lucas Lochovsky 4/17/2015 4:29 PM
Deleted: give a p-value

Lucas Lochovsky 4/17/2015 4:30 PM
Deleted: en

Lucas Lochovsky 4/17/2015 4:30 PM
Deleted: And t

Jing Zhang 4/7/2015 1:25 PM
Comment [2]: Too much?

Jing Zhang 4/15/2015 10:56 AM
Deleted: . Although LARVA is able to find meaningful results on both coding and noncoding regions, we currently have no genuine benchmark dataset available to rigorously evaluate LARVA's results

Jing Zhang 4/15/2015 10:56 AM
Deleted: two

Referee major comment 2:

> *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's Response:

We thank the reviewers for pointing out this important issue. The QQ plots of p-values are provided in the following figures.

Lucas Lochovsky 4/17/2015 4:30 PM
Deleted: were

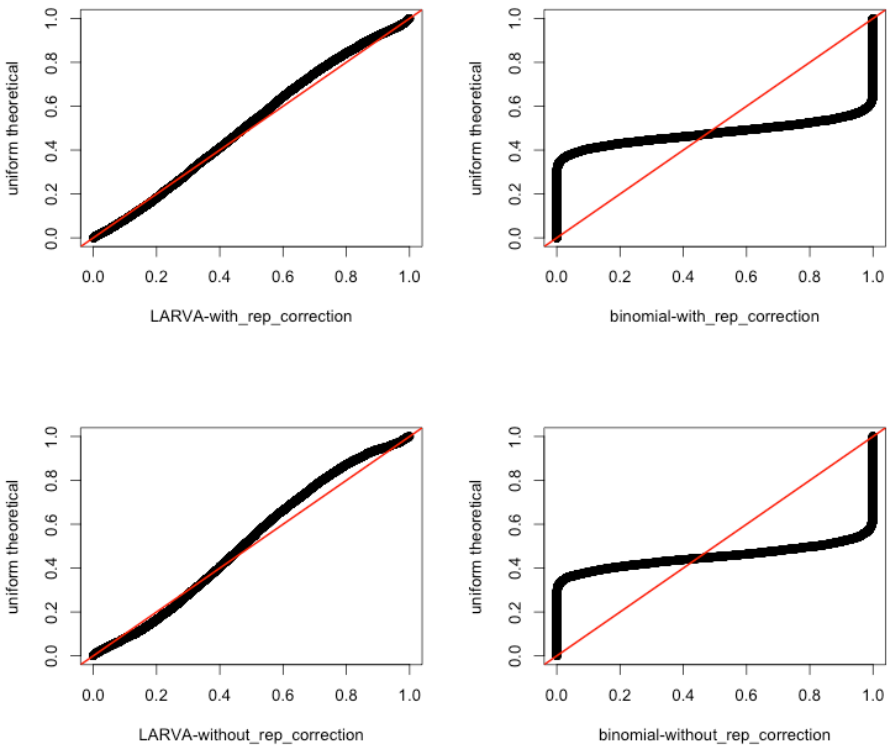

Figure R 5: coding region p-values vs. theoretical. The red line is the diagonal line.

Lucas Lochovsky 4/17/2015 4:31 PM
Deleted: Red

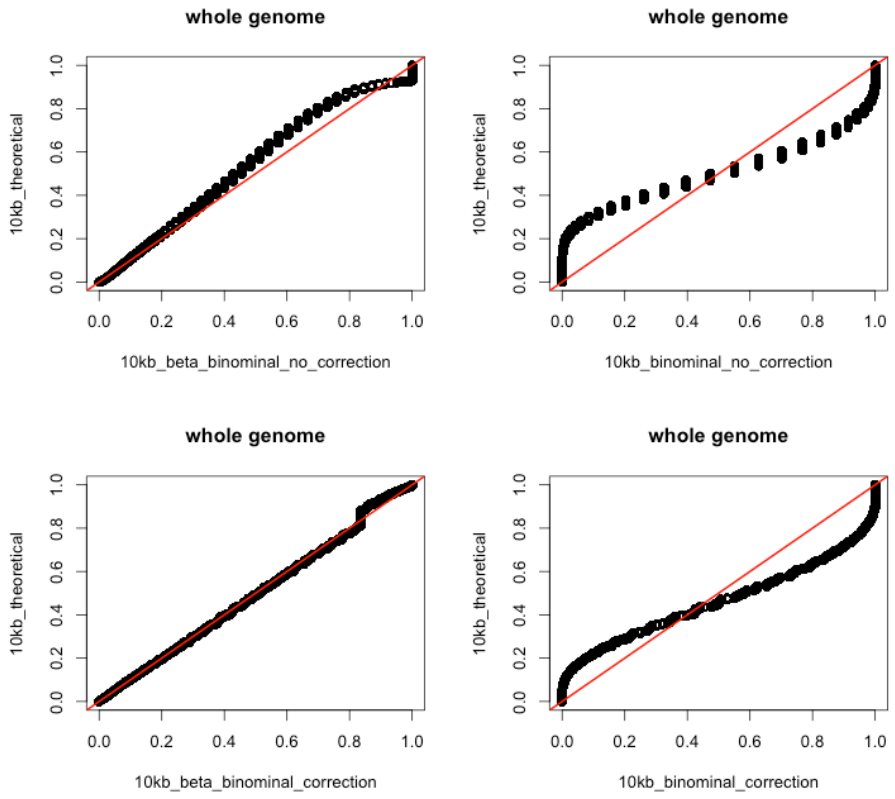

Figure R 6: QQplot of the pvalues from genome wide 1kb bin

In Figure R 5, it is shown that the p-values from binomial test severely violates 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 the 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 improved p-value distribution in LARVA vs. binomial test. The discrete dots in Figure R 6 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.

Unknown
Formatted: Font:(Default) Helvetica

Jing Zhang 4/15/2015 10:58 AM
Formatted: Caption, Centered, Space After: 0 pt, Line spacing: single

Lucas Lochovsky 4/17/2015 4:31 PM
Deleted: Still

Lucas Lochovsky 4/17/2015 4:32 PM
Deleted:

Lucas Lochovsky 4/17/2015 4:32 PM
Deleted: helps to

Lucas Lochovsky 4/17/2015 4:32 PM
Deleted: