Author’s response to reviews

Title: ClickGene: an open cloud-based platform for big pan-cancer data genome-wide association study, visualization and exploration

Authors:

Jia-Hao Bi (jh_bi@tju.edu.cn)
Yi-Fan Tong (tongyfan@gmail.com)
Zhe-Wei Qiu (zw_qiu@tju.edu.cn)
Xing-Feng Yang (nerosoft@outlook.com)
John Minna (john.minna@utsouthwestern.edu)
Adi Gazdar (adi.gazdar@utsouthwestern.edu)
Song Kai (ksong@tju.edu.cn)

Version: 1 Date: 27 Mar 2019

Author’s response to reviews:

Dear Editor

We are so grateful for the reviewer's approval and encouragement. We are also very appreciate for the comments which helped a lot to improve our manuscript. The following are the answers to the reviewer's comments point by point. Please feel free to contact us if you have any questions.

Reviewer #1: Bi et al. present an online application (ClickGene) for plotting processed genomic data of large cancer cohorts. There are multiple available tools of this type, the most well-known perhaps being cBioPortal. These tools are useful for performing quick exploratory analyses of processed cancer data and generating hypothesis. The specific tool introduced in this paper allows the user to create several types of plots that are not included in cBioPortal. It also performs simple statistics. The online application comes pre-loaded with the TCGA dataset, but also allows the user to upload their own dataset.
Although I find the tool is worth publishing, there are multiple aspects with the way it is presented in the paper that I think need to be revised:

1. I find the discussion section in the paper to be very misleading. The authors focus the discussion on results that they claim are new or inconsistent with the current literature. However, as far as I see, none of those results represent an inconsistency, but rather a misinterpretation of the results by the authors. For instance, the authors look at copy number alterations of chromosome arms 1p and 19q in glioblastoma. Co-deletion of 1p and 19q is a well-characterized event in low-grade IDH-mutant oligodendroglioma, to the extent that it is now one of the defining characteristics of IDH-mutant oligodendroglioma according to the WHO Classification of Brain Tumors. In high-grade gliomas, 19q deletions are rare, usually only observed in secondary IDH-mutant glioblastomas. Most glioblastomas therefore do not carry a 1p/19q co-deletion. In the paper, the authors look at the average copy number of 19q in the TCGA glioblastoma cohort and find that it is above 2 (figure 6) and conclude that "19q should not be considered as a deletion". I do not see any inconsistency here with the previous literature. Moreover, looking at the mean is misleading, as it is particularly sensitive to outliers. If instead, we consider the median, we observe that 19q indeed is very often deleted in low-grade gliomas, and occasionally also in high-grade glioblastomas, as evidenced by a similar plot produced with ClickGene but using the median instead of the mean. In my opinion, the paper would improve substantially if the authors focus the discussion section in presenting some examples of analyses of the TCGA data using ClickGene, highlighting the consistency with previous studies.

Answer: We are so sorry for the misleading of this section. We agreed with the reviewer. Therefore, we moved the results of GBM consistent with the published papers back into the main manuscript to prove the performance of CG platform.

We agree that MEAN value is misleading as it is particularly sensitive to outliers. That's the reason we provide Median value as an option. But for fair comparison with the published results, we chose to MEAN value in the manuscript. We added a sentence in the manuscript to remind readers of the sensitivity of mean value to the outliers.

2. In the discussion, the authors present dynamic time wrapping (DTW) scores for the copy number alterations of adenocarcinomas and squamous cell carcinomas (table 1), and highlight differences among different cancer types. However, the differences in score are small (< 5%) and the authors do not perform any statistical test to assess the significance of these differences. Are they consistent with random fluctuations? The authors could use some approach, like bootstrapping, to better assess the significance of these differences.
Answer: The DTW score is calculated to quantify the difference among the CNV Mountain plots of different types of cancer. Since the CNV curve of each cancer is created with the MEAN/MEDIAN value of each gene in each group, so it's dependent only on the CNV curves. Once the curve is determined, the DTW score is determined too. Therefore, it's not necessary to do bootstrapping or something like this. But since MEAN value is sensitive to the outliers, the corresponding DTW score is sensitive to the outliers too. But the DTW score calculated from the MEDIAN values would be better. It's better to use DTW, Deflection plot and Manhattan plot together. We used the difference between CNVs of COAD and READ as an example to explain how to discover the difference by using these tool together.

The mechanism of DTW makes it not more sensitive to the lengths of fluctuations than to the amplitudes. We used the DTW score of 12q in GBM as an example to demonstrate it in the Discussion part.

3. The application includes simple statistical tests. However, significances in the plots produced by ClickGene are not adjusted for multiple hypothesis testing. I find this is an important aspect in this context, as many of the tests are performed thousands of times to produce a plot. In my opinion, the authors should implement at least one of the standard procedures for controlling the false discovery rate (e.g. Benjamini-Hochberg).

Answer: We completely agree with the reviewer. To adjust multiple hypothesis testing, in our previous study of genome-wide CNV patterns between LUAD and LUSC in reference 13, we used Bonferroni correction, the most conservative one. This is the very reason that we allow users to specify the significance level. Therefore, they could adjust the family-wise significance test by giving a different cutoff rather than using 0.05. In this revised version, since we agree with the reviewer, we used Q-value after the controlling of the false discovery rate in Manhattan plot and Deflection plot. The popular way for using Volcano plot is using p-value, so we still use p-value in Volcano plot. We added this in the Discussion section.

4. Although in general the online application works reasonably OK, it often gets stuck without giving an error. For instance, if I try to generate a Beeswarm plot of the copy number variations of gene ADAMTS12 in lung adenocarcinoma (LUAD), the application goes into the "Loading..." screen indefinitely, without giving an error.

Answer: We are so sorry for this bug. We fixed it. Thank you so much!

5. Overall, I find the writing could be much improved. There are many grammatical errors, as well as colloquial expressions that do not seem suited for a scientific publication.
Answer: We invited an native speaker to help us to modify the English this time.