# **Response letter for resubmission**

# **Reviewer 1**

### **-- Ref 1.1 –MCF7 analysis--**

|  |  |
| --- | --- |
| Reviewer  Comment | In the response letter, "Ref 1.3, MCF7 analysis", the author said that "the contact maps provided by Barutcu et al. 2015 were already ICED at the beginning. The effect of CNV should at least be partially reduced". The sentence is not accurate. ICE, as a matrix-balancing algorithm, is relied on the "equal visibility" assumption. It assumes that if there is no systematic bias, each bin has the same number of "total" interactions. However, in cancer cells, bins with CNV will have more interactions than bins without CNV. Therefore, the ICE and "equal visibility" assumption are not valid here. I suggest the authors to down-tune it and demonstrate the limitation of ICE normalization. One alternative approach is to apply explicit Hi-C normalization method, such as Yaffe and Tanay, Nature Genetics, 2011. |
| Author  Response | First of all, as we have mentioned in our last response, we agree with the reviewer’s concern in this aspect, and we have already added a caveat in the manuscript in our last revision (see excerpt). We mentioned in the method section that the contact map of MCF7 cell line was normalized by the ICE algorithm, but we actually did not say in the text that the ICE normalization can reduce the effects of CNV (we said in our response to the reviewer).  Now, let us clarify our thought. Like most of the TAD analysis in the literature, we called TADs by mapping reads to the reference genome. In what scenario the presence of CNVs may affect TAD calling? We can imagine there is a bin which is not a boundary at the first place. However, because of its multiple copy numbers, the number of reads mapped to the bin increases, and therefore the TAD calling algorithm called it. As the reviewer pointed out, the ICE normalization assumes equal visibility, the increases in the number of reads should therefore be normalized and the number of false positives we hypothesized should be reduced. We did not claim the ICE algorithm can solve the problem. Indeed, algorithm mentioned by the reviewer in the last revision, HiCapp, that makes use of the distance dependence of interactions to correct the effect of CNVs is probably a better method in removing the false positives we imagined. Perhaps a possible way to address the issue is to simulate contact maps with CNVs. Nevertheless, it is a rather non-trivial problem and we believe it is beyond the scope of this paper.  Back to our original analysis in Figure 7, what we observed was an abrupt change of mutational load across TAD boundaries. We believe such a pattern is unlikely to be the effects of domain boundaries that are false positives. The presence of certain false positives should only weaken the pattern of interest. To sum up, we share the reviewer’s concern that the effect of CNVs in TAD calling is an important but not fully addressed problem in this paper, and probably in literature. Nevertheless, we believe it does not affect that the conclusion in Figure 7, as well as the main idea of our manuscript. |
| Excerpt From Revised Manuscript | In P. 12  Nevertheless, it is worthwhile to point out that mapping Hi-C reads from cancer cell lines like MCF7 to the reference genome is not perfect because quite some reads may come from translocations or copy number variations. Computational approaches have recently been developed to perform correction as well as to infer those large scale genomic alterations [27][28]    In P.17  Hi-C data and contact maps in MCF7 cells were reported in Ref. [48]. The whole-genome contact map provided was binned with 40kb bin size and was normalized by the ICE algorithm. . |

### **-- Ref 1.2 –Reproducibility--**

|  |  |
| --- | --- |
| Reviewer  Comment | In the response letter, “Ref 1.5, reproducibility”, the authors down-sampled the reads to 30M, 25M, …, 1M reads, and evaluate the reproducibility at different sequence levels. Figure S9 shows the normalized MI. I understand that they did not include this analysis in the manuscript due to the data is not published, but there are many published Hi-C datasets which can be used for reproducibility analysis. I think it is very useful to include the reproducibility analysis using published Hi-C datasets. In addition, I feel the down-sampled sequencing depths are relatively low. For example, Rao et al, Cell, 2014 paper contains GM12878 Hi-C data with 4.9 billion reads, Jin et al, Nature, 2013 paper contains IMR90 Hi-C data with 3.4 billion reads. A typical Hi-C study contain 200~300 million reads. I am curious to see the reproducibility of TAD calling results for higher sequencing depths. |
| Author  Response | We agreed with the reviewer’s concern. We performed a new reproducibility analysis using GM12878 data from Rao et al, Cell 2014 and included the analysis in Figure S10. More specifically, we made use of the raw contact matrices mapped from individual Hi-C libraries of the same replicate. The superposition of 15 libraries results in a contact map constructed from 2.4 billion reads. We then applied MrTADFinder to obtain the TADs. We then reduced the number of libraries one by one and performed TAD calling. The new sets of TADs were compared to the original TADs based on mutual information. As shown in Figure S10, the discrepancy is rather low. If we compare the TADs obtained from 2.4 billion reads to the TADs obtained from 480 million reads, the normalized mutual information is 0.88. More importantly, we do observe a gradual increase in mutual information when the sequencing depth gets close to the original map. On the contrary, the trend was not observed and the discrepancy is in general larger (MI roughly equals to 0.7) in our last analysis in which the down-sampling were started from 30 million reads. We would like to thank the reviewer for pushing this analysis. It is indeed important to gauge the reproducibility of TAD calling algorithms with respect the sequencing depth. As the reviewer pointed out, a typical Hi-C experiment contains 300-400 million reads, the reproducibility of MrTADFinder should be quite well behaved in such regime. |
| Excerpt From Revised Manuscript | P.13  Secondly, we explored the effects of sequencing depth to our algorithm. Specifically, we applied MrTADFinder to identify TADs from a deeply sequenced Hi-C data of GM12878 [21]. We then reduced the number of reads included and called TADs again. We found that the TADs identified using a subset of reads are slightly different from the original, and in general, the discrepancy increases as fewer reads were used (Figure S10 and Methods). Despite a certain level of discrepancy, nevertheless, the resultant TADs agree well. For instance, in the extreme case, by comparing using contact maps constructed from 2.4 billion reads and 480 million reads respectively, the mean normalized mutual information of various pairs of chromosomes is about 0.88. If we compare the TADs called from 2.4 billion reads to the TADs called from 1 billion reads, the normalized mutual information is higher than 0.95. |

### 