Response letter for resubmission

Reviewer 1

-- Ref 1.1 –the choice of gamma--

Reviewer Comment	\gamma is the resolution parameter to determine the size of TADs. In practice, how to choose \gamma? The authors need to provide some guideline of \gamma selection.
Author Response	Thanks for the suggestion. It is indeed an important point. We have added some guideline in the discussion section.
Excerpt From Revised Manuscript	P. 13 From a practical point of view, seems to be the natural starting point. One could increase or decrease the value of in order to explore the intrinsic structure. Nevertheless, because of the different contact maps might have various differences like the read coverage, one should be cautious to directly compare the resolution parameters between different contact maps.

-- Ref 1.2 –TAD boundaries--

Reviewer Comment	In Figure 3, are the enrichment of histone marks statistically significant? The enrichment of some histone marks at TAD boundary regions look very weak. How about the enrichment of house-keeping genes at TAD boundary regions under different resolutions?
Author Response	The enrichment is estimated based on the number of peaks fall in the boundary over an expectation in which peaks are randomly distributed. Naively, we can estimate the statistical significance based on a Poisson model. In this case, the enrichments like H3k27ac in high resolutions are all significant. This is because at high resolution, the expected number of peaks fall at the boundary is quite high as there are more and more boundary). Moreover, it is our message that the enrichment decreases as the resolution increases, the statistical significance of the weak cases is not very relevant. In regard to housekeeping genes, we found performed the additional analysis. We found that housekeeping genes and tissue-specific genes have different characteristic resolutions. We have reported the analysis in the manuscript and included a new figure (Figure 4) to report this interesting results.

Excerpt From Revised Manuscript	P. 8-9 Beside epigenetic signatures, we examined the distribution of protein-coding genes along chromosomes in relation to TAD boundaries formation. Though the starting positions of genes tend to be enriched near TAD boundaries, the enrichment is much stronger for housekeeping genes as compared to tissue-specific genes (Figure 4A). This observation was firstly reported in Ref. [8]. Nevertheless, by extending the idea to multiple resolutions, we found that the distribution of housekeeping genes follows a different length scale compared to tissue-specific genes. As shown in Figure 5B, housekeeping genes in general marks the boundary of TADs up to the resolution .

-- Ref 1.3 –MCF7 analysis--

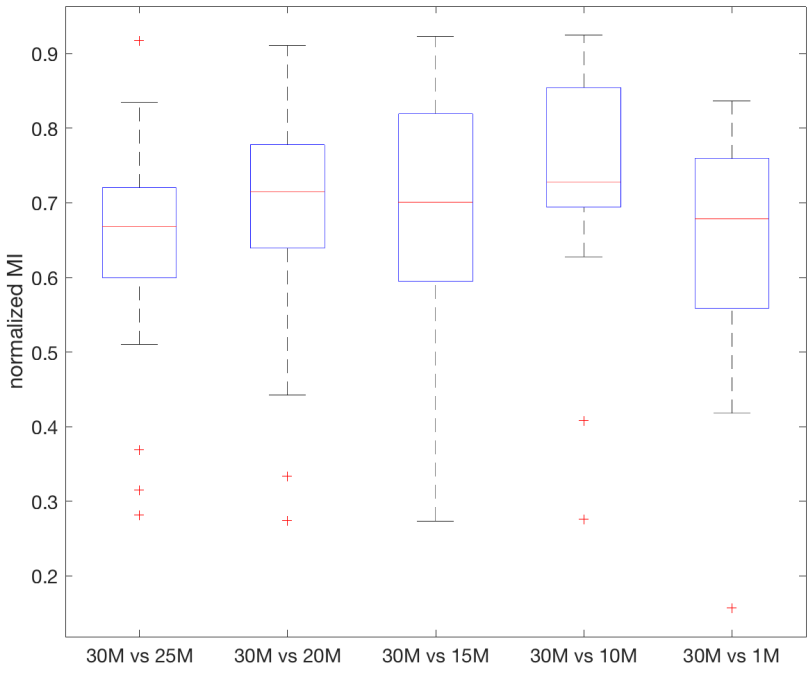
Reviewer Comment	The authors need to provide more details on the analysis of cancer Hi-C data MCF7, since high number of Hi-C reads may come from translocation or copy number variation. In addition, reads mapping will be complicated for the heterogeneous cancer genome. There is a recent method named "calCB" to adjust for CNV in cancer Hi-C data (https://doi.org/10.1093/bioinformatics/btw540). The authors can apply calCB to re-analyze MCF7 data.
Author Response	We agree with the reviewers that this is indeed an issue. However, we were not able to perform the analysis as suggested by the reviewer due to technical issues in running the calCB pipeline. We decided not to pursue further because the analysis of MCF7 is not a central result of the manuscript. However, even though it might be the case that a few TADs were missed, it should not affect an analysis like Figure 5 too much because it is based on accumulating all distributions. Also, we would like to mention 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. Again, we fully recognize the point raised by the reviewer. Therefore, we have included more details and discussed this point in the manuscript.
Excerpt From Revised Manuscript	In P. 11 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 [22][23]. In P.16 Hi-C data and contact maps in MCF cells were reported in Ref. [43]. The whole-genome contact map provided was binned with 40kb bin size and was already passed the ICE normalization.

-- Ref 1.4 –optimal partition--

Reviewer Comment	In page 12, last paragraph, the authors mentioned that "Given the time complexity, finding the optimal partition using a bin size of 40kb is quite impractical". Can they try to find optimal partition in some short chromosomes, such as chr18 and chr19?
Author Response	We performed the calculation for chr21 (the shortest chromosome) using a bin size of 40kb. It takes about an hour, in comparison to a few seconds by the heuristic. We have added this observation to the manuscript to illustrate the meaning of "impractical".
Excerpt From Revised Manuscript	P 12. Given the time complexity, finding the optimal partition using a bin size of 40kb is quite impractical. For instance, the calculation takes about an hour for chromosome 21, as compared to seconds by using the heuristic.

-- Ref 1.5 –reproducibility--

Reviewer Comment	What is the reproducibility of MrTADFinder between two biological replicates? Are the TAD calling results sensitive to the overall sequencing depth?
Author Response	This is indeed an important question. We have performed a new analysis using Hi-C data released by the ENCODE consortium in which biological replicates were performed in each cell line. We have chosen cell lines with a relatively higher coverage, resulting at 8 out of 12 available cell lines. We found that, for each cell line, by running MrTADFinder in each of the two replicates, the two sets of TADs agree reasonably well, with normalized mutual information over 0.85 (see Figure S9). We have added this analysis in the manuscript. Concerning the effect of sequencing depth, we looked at one of the more deeply sequenced cell line in ENCODE Hi-C data, the G401 cell line. Our collaborators have downsampled the reads, and generated contact maps in various sequence levels: 30M, 25M, 20M, 15M, 10M, 1M reads. We ran MrTADFinder for all these maps, and compare the resultant TADs. As shown in the included plot, the overlap is reasonable but we do not see a consistent effect with respect to the sequencing depth. Because our collaborators are working on a separate manuscript based on their downsampled maps, we decided not to include this analysis in the manuscript.

	
Excerpt From Revised Manuscript	See P.13. We further investigated how MrTADFinder performs in replicates. Using Hi-C data released by the ENCODE consortium, we found that TADs called in a pair of biological replicates agree reasonably well, with normalized mutual information about 0.85 (see Figure S9 and Methods).

-- Ref 1.6 –significance of boundaries--

Reviewer Comment	Since MrTADFinder formulates the identification of TAD as a global optimization problem, I would expect that the TAD calls are deterministic. How to model the variability in Hi-C data, and how to evaluate the uncertainty in TAD calls? It would be ideal if the authors can provide p-values to quantify the significance of TAD boundary.
Author Response	The reason why MrTADFinder is not deterministic is because we have employed a heuristic (the modified Louvain algorithm) which is probabilistic in nature. By performing the heuristic multiple times, a boundary score is defined, which simply means the fraction of times a bin is called as a boundary. The boundary score is a part of the output, and it is essentially a confidence measure for quantifying the significance of TAD boundary. Though the boundary score could be used separately as a statistical measure, to simplify the output, we define a set of consensus boundary based on the boundary score (score>0.9),

	and then a set of TADs is defined. We have already provided such details in the methods section (see Heuristic procedures for optimizing Q). To emphasize this point, we have briefly outlined the idea in the main text. Concerning the variability in Hi-C data, see Response 1.5 above.
Excerpt From Revised Manuscript	P 6. To ensure robustness, multiple runs of the modified Louvain algorithm are performed, and a boundary score is defined as the fraction of times a bin is called as a boundary. The final set of TADs is defined based on the set of consensus boundaries (Figure 1 and Methods).

-- Ref 1.7 –--

Reviewer Comment	Page 14, 2nd paragraph. The claim "However, the dependence of intra-chromosomal interactions and genomic distance is not explicitly modeled" is not accurate. For example, Ay et al (http://genome.cshlp.org/content/24/6/999.long) developed a method called Fit-Hi-C, which explicitly models the dependence of intra-chromosomal interactions and genomic distance via a non-parametric spline curve. The authors need to remove or down-tune that sentence.
Author Response	We are fully aware of work mentioned by the reviewer. While methods aiming to identify significant hi-c interactions explicitly model the dependence of intra-chromosomal interactions and genomic distance, we believe that is not the case for TAD calling algorithm. However, to avoid confusion, we have removed the sentence.

Reviewer 2

-- Ref 2.1 –Readability of Fig. 3A--

Reviewer Comment	Fig. 3A, only keep a couple of panels and put the rest of them into a supplementary figure. The figure as it is now is not readable;
Author Response	Thanks for the suggestion. We have made the change.

-- Ref 2.2 –AUC--

Reviewer Comment	Page 9, line 14, add details how AUC is computed;
Author Response	We have added the details in both the main text and the methods section.
Excerpt From Revised Manuscript	Main text: To do so, we formulated a classification problem which aims to distinguish, for each resolution, a set of boundaries identified by MrTADFinder (positive set) from a set of random boundaries obtained by swapping the TADs along the chromosomes (negative set). Using a logistic regression model recently proposed by [17], we integrated the binding signals of 60 transcription factors at a genomic locus to predict if it is TAD boundary (see Methods for details). Generally speaking, with 10-fold cross validation, the model is quite successful (AUC=0.81, Figure 4B). Methods: The influence of individual transcription factors on the formation of domain borders was formulated as a classification problem. For a particular resolution, the set of boundaries called by MrTADFinder was used as a positive set whereas a set of random boundaries obtained by swapping the TADs along the genome was chosen as the negative set. The signal values of 60 transcription factors are used as features for classification. The combined effect of all features was modeled the logistic function, here X represents all features; is a vector determining the coefficients of influence for all features and is a bias parameter. Given a training set, a likelihood function was defined. An optimal was inferred by optimizing the likelihood function using gradient descent with L1-regularization. The inferred logistic function was used to predict the test set. To have a more accurate estimate, 10-fold cross-validation was performed, and the error bars were estimated by multiple negative training sets.

-- Ref 2.3 –--

Reviewer Comment	Page 9, line 15, "predicting power of the model decreases as the resolution increases" – what is the reason for this observation? Does it have anything to do with the accuracies of predicted TAD boundaries?
Author Response	We believe this observation is in relation to the general trend shown in Figure 3 and 4: the decrease of enrichment of various chromatin features as resolution increases. The trend suggests that as the resolution increases, the number of TAD boundaries called increases and some of the high resolution boundaries identified do not have the distinctive chromatin features associated. As a result, the model trained has a weak predictive power.
Excerpt From Revised Manuscript	P. 10 Being consistent with the trend that chromatin features are not always present at the boundaries of high resolution TADs, the predicting power of the model decreases as the resolution increases.