Author's Response To Reviewer Comments

Close

January 14, 2021

Dear Dr. Nogoy:

We thank you and the reviewers for considering our manuscript, "Chromatin conformation capture (Hi-C) sequencing of patient-derived xenografts: analysis guidelines." Below, we address each of the reviewers' points and describe the changes we have made to the manuscript. This revision includes 2 new tables, 2 modified figures, and 5 new multi-panel figures as Additional Files. All new results are incorporated in the manuscript, with all figures and additional files renumbered accordingly. We have indicated changes in the manuscript in red typeface.

Thank you very much for your consideration.

Best regards,

Mikhail Dozmorov

Virginia Commonwealth University

Reviewer 1

Q: Dozmorov, Tyc et al. present guidelines for analysis of Hi-C data generated from PDX models with respect to mouse DNA contamination in the sample. They use in silico spike-in of mouse Hi-C reads and actual Hi-C data from PDX samples to show that different approaches to mapping mouse and human reads and read processing do not affect the final Hi-C maps.

This is an important work and will be of high value for the 3D cancer genome field, however I do not think that the results presented by the authors justify the conclusions and therefore more analysis needs to be done before this manuscript can be published.

The key analyses that need to be performed are to look at the effect of mouse spike-in reads or mouse cell contamination on chromatin interactions. Presented results focus only on high-level domain structures (TADs) and are limited to look at the total number/size of TADs called. TAD boundaries called from Hi-C data have been previously shown to be highly overlapping between mouse and human genomes as well as in some other species (as recently discussed in Eres and Gilad, Trends in Genetics, 2020). However, the main correlation between TAD calls in different datasets can be explained by the use of the same calling algorithm. Therefore, TADs and TAD boundaries are not a good measure of the effect of mouse cell contamination in Hi-C data.

A: We agree that the algorithm used to call TADs will drive high correlations across different datasets. However, the choice of the same calling algorithm was intentional - we aimed to see whether different levels of mouse read contamination and processing pipelines affect TAD detection, keeping all other variables unchanged. We focused on human-only TADs, as mouse reads represent a small fraction of PDX Hi-C data and removed d data processing. We now added the following sentence in the Results section: "To focus on the data- and pipeline-specific differences, we used the same TAD/loop calling algorithms throughout our work (see Methods)."

Q: Instead, the analyses should be focused on chromatin interactions (or enhancer-promoter interactions), which are more cell-type specific. Authors need to show how many mouse-specific interactions are present in the final Hi-C data from PDX samples as well as look at the enrichment of all valid interactions for mouse vs human enhancers, promoters and CTCF binding (using public histone mark data or chromHMM and CTCF ChIP-seq).

A: In this revision, we significantly expanded on the downstream biological analysis that explores biological aspects of the data. First, we assessed the number of chromatin loops when using different alignments and pipelines. To avoid redundancy, we removed the TAD/loop size results as they are inversely related to TAD/loop number. Second, we assessed the overlap of the detected TAD/loop boundaries using the Jaccard coefficient, and visualized the results using multi-dimensional scaling. Third, we investigated the enrichment of CTCF co-localization and signal distribution at TAD/loop boundaries. We hope the new results will strengthen our conclusions about mouse read contamination.

Minor comments

Q: 1. The difference between two Hi-C kits used (Library 1 vs Library 2) including names of the kits and restriction enzymes used should be included somewhere at the front of the results section.

A: This is a delicate point. The use of abstract Library 1 and Library 2 labels is an attempt to emphasize the importance of library preparation strategy, not to undermine one company over the other. However, we explicitly mention the companies in the Methods section, describe their protocols, and refer to them immediately when describing the experimental PDX Hi-C data as "two different library preparation strategies (Library 1 and Library 2, see Methods)." We hope that a regular reader will be satisfied with the high-level labeling of library differences. At the same time, for an advanced reading of our paper, we present all the necessary experimental details in the Methods section.

Q: 2. Can the 40% duplication in Library 1 (Phase Genomics kit) be explained by over-sequencing of the library that is not complex enough due to only one RE used in the kit?

A: It is very hard to pinpoint the cause of the high duplication rate in the Library 1-prepared data. For one, a high duplication rate seen in the sequencing data can be an indication that the reaction was not carried out for long enough, rather than the use of a single RE per se. Regardless, we made all efforts to tackle the issue, e.g., quantified k-mer content to assess library complexity, but couldn't decisively conclude what the most plausible cause of the high duplication rate was. We make this data publicly available. As analysis strategies continue to develop, we believe this data will serve as a key testing set when analyzed by others.

Q: 3. Fig. 5 - TAD number and sizes are not a good quality metric for this question as they are mainly driven by the type of the algorithm used to call TADs.

A: Please, see our response above. We intentionally used the same TAD/loop calling algorithms to focus on mouse read contamination- and pipeline-specific differences, keeping all other variables unchanged. We clarified the rationale of comparing TAD/loop numbers as "The number of TADs and loops should be considered as a suggestive indicator of data quality under the hypothesis that a deeper-sequenced high-complexity Hi-C experiment would produce Hi-C matrices where more TADs/loops can be detected." The key message from this experiment is that the mouse read contamination and/or alignment strategies do not significantly impact the TAD/loop calling step.

Q: The authors should instead include analysis of the actual insulation score/directionality index that underlines the TAD calls and show correlation between the scores, PCA/MDS plot and look at overlap between called boundaries to see if there are any mouse-specific TAD boundaries that are present in the in silico Hi-C data and in vivo PDX data.

A: We used Figure 1 to emphasize that the data is processed in such a way that only human Hi-C matrices are considered for the downstream analyses. As such, there are no mouse-specific TAD boundaries. As we show, for example, in Figure 3A, there are no meaningful differences in the alignment rate between the three strategies we adapted, suggesting mouse reads are nearly completely eliminated even during the Direct alignment strategy.

Regarding the first part of the question, we investigated the correlation between eigenvectors that define A/B compartments. We calculated Jaccard scores for TAD and loop boundaries and found nearly identical results in all instances. We believe Jaccard overlap is the most interpretable metric of overlap, and for that, we visualized these results using Multi-Dimensional Scaling (MDS) plots. Given that the results are almost indistinguishable, we present TADs and loops as the most fine-grained details of the 3D genome analysis.

Q: 4. Authors should look at interactions that are associated with mouse-specific genes - can these be

observed in the in the in silico Hi-C data and in vivo PDX data? Some visual examples are needed as well.

A: Again, we are sorry that our message that we are removing the mouse signal from the data got lost. Mouse-specific genes are only on the mouse genome, which we remove upfront from the analysis, so the mouse-specific interactions in the final file simply cannot occur. We now highlight this point in the manuscript and refer more to Figure 1 to help to illustrate this fact.

Q: 5. It is expected that PDX Hi-C data will show more intra-population heterogeneity as compared to cell line Hi-C data. This will affect "background" noise interactions, which may be present only in small sub-populations of cells and therefore affect the signal to noise ratio. Can this be clarified from the different analysis pipelines used and therefore be a key consideration for researchers when deciding on the best pipeline to use for PDX samples?

A: Thank you for pointing this out. This is indeed an important consideration, and we added the following clarifying statement early in the Results section: "This higher intra-population heterogeneity is expected because, in contrast to cell lines, experimental PDX samples contain a mixture of different cell types and cell states. This will introduce the background noise interactions and should be considered when comparing experimental and in silico PDX Hi-C analysis results."

Q: 6. In PDX tumour samples, mouse fibroblasts have been shown to infiltrate tumours and introduce mouse signal in the analyses data. Can the authors look at fibroblast-specific interactions (e.g. based on fibroblast genes) in the PDX data to see if these can be detected?

A: Our study specifically focused on removing mouse reads from PDX Hi-C data. In any case, looking at mouse-specific interactions is possible in theory, but there are multiple issues that ought to be considered. First, mouse reads are only 10-30% of the total Hi-C reads, making mouse Hi-C matrices extremely sparse. Second, we are not sure how homologous are mouse and human fibroblast genes. The latter creates ambiguity in assigning reads mapping to homologous regions in human fibroblast genes as being of either human or mouse origin. We kindly ask to provide a reference illustrating the suggested analysis and are committed to implementing it given the example.

Reviewer 2

In recent years, Hi-C has been applied to cancer genomes, with the aim of both characterizing cancer-specific alterations of 3D genome organization as well as changes in the 1D genome sequence such as structural variations. Patient-derived xeongrafts (PDX) provide an important system for studying cancer, and is associated with unique technical problems as the human tumor cells are contaminated with mouse DNA. In the current manuscript, the authors test a number of different techniques, both computational and experimental, and try to evaluate which combination provides better quality data. Such a study can be quite useful for other groups pursuing Hi-C in PDX, and while currently this work will be of interest to a relatively small group of specialists, the potential applications of Hi-C in cancer may widen the interest in this type of work in the future. The authors test 3 different techniques for differentiating mouse from human reads, 3 different Hi-C computational pipelines, and 2 different Hi-C protocols (commercial kits). The authors look at a number of different quality statistics, and conclude that the best combination of approaches is probably "Direct" mapping of both human and mouse reads, Juicer Hi-C pipeline, and the Arima Hi-C kit. In general the paper is well written, clear and technically sound. My main issue is with the interpretation of the quality statistics.

Main issues

Q: 1) The authors propose a number of different Hi-C quality statistics, but there is often a tradeoff between these statistics. Thus, in order to show that one method is better than others one needs to show an improvement in all parameters. For example, the authors state that Juicer is better than the other pipelines, and this is supported by more valid reads mapped and better cis/trans ratio. However, the Long/short read ratio worse than the other methods. Cis/trans is a good measure for evaluating the amount of random ligation (which yields more cis than trans). However, another type of common bias in Hi-C data is short range cis interactions that may result from a number of causes such as insufficient digestion or contamination by unligated fragments. It is entirely possible that Juicer maps more of these incorrect short range read pairs, and this is reflected by a higher cis/trans ratio and more "valid reads." Thus, it is not possible to determine based on the current metrics that Juicer is better than the other pipelines. Specifically for this case, A possible control for this bias would be to calculate "valid mapped

reads" and "cis/trans" using only reads>20kb. More generally, I would be careful with drawing strong conclusions about quality unless all statistics point in the same direction (this is not to say the results are not useful, just that the conclusions might need to more careful).

A: We thank the reviewer for the very insightful comment. We agree with all conclusions, and amended that sentence that Juicer is "better." In fact, Juicer indeed mismaps more reads, especially mouse-specific reads (the new Additional File 4: Table). In this revised version of the manuscript, we refrain from mentioning Juicer in the abstract and instead focus on the Direct alignment strategy and the importance of library preparation, limiting our conclusion that "The choice of processing pipeline had negligible impact on data quality and the downstream results." We changed the wording about the long/short ratio to report that: "Juicer yielded lower long/short ratios compared to other two pipelines …," and adjusted the wording about Juicer throughout the manuscript. We hope the complete tool-specific QC outputs provided in Additional File 5 will help PDX Hi-C data analysis practitioners to select the right pipeline based on their preferences. We discuss the potential differences between pipelines in Discussion, describing different short-read aligners used by different pipelines.

Q: 2) It is unclear why the authors can conclude that a "smaller" power-law exponent is better (note that the way the authors use the term "smaller" is confusing here because these are actually negative numbers, -1.83 is not smaller than -1.99). Artifacts like background ligation can cause a shallower decay.

A: Thank you for pointing this out. After extensive discussion, we decided to remove the distance-dependent decay section. This decision is based on the minuscule differences between pipelines, mouse removal strategies, and even library preparations. Together with the need to explain the power-law decay exponent, these results create more confusion than provide illustrative information. Consequently, we focused our results on the easily interpretable and more biologically relevant TAD/loop analysis. Figure and Additional file numbering have been adjusted throughout the manuscript.

Q: 3) The same goes for TADs. TAD calling pipelines can be affected by data biases in different ways, especially since these are often hierarchical overlapping structures, and it is certainly not clear whether finding more TADs is better or worse in terms of data quality. For example, with a higher level of background ligation/mismapped reads, it could be more difficult to identify larger TADs, so only the nested TADs are found resulting, in more TADs.

A: We looked at the number of TADs, and, new in this revision, chromatin loops, to conclude that mouse reads do not affect TAD/loop calling. We agree that the number of TADs/loops is not an indication of data quality and clarified our intuition as "For high-quality Hi-C data (derived from the high-complexity library, deeply sequenced), we expected the algorithm to detect more TADs/loops." Additionally, we investigated the Jaccard overlap between the detected boundaries and visualized the results using Multi-Dimensional Scaling (Additional Files 9 and 10). Furthermore, we investigated the enrichment of CTCF co-localization and signal distribution at TAD/loop boundaries to investigate whether the detection of more TADs/loops improves CTCF enrichment (Additional Files 11, 12, and 13). We updated the figures, added new Additional Files, and described all new observations in the text.

Minor issues

Q: 1) It might be useful in Figure 2 to add a vertical horizontal at the 10% and 30% threshold, where relevant.

A: Thank you for the suggestion - we updated the figure by adding the dashed lines at 10% and 30% thresholds.

Q: 2) Is there any mismapping of reads mouse->human in the in silico data?

A: This is an important question, and we present the new results in Additional File 4. This analysis revealed that Juicer has a higher human-mouse read mismap rate. This may propagate on capturing artifacts as valid pairs and lead to over-optimistic results. Consequently, we adjusted the message that the Juicer pipeline is better, as described above.

Close