GigaScience

Chromatin conformation capture (Hi-C) sequencing of patient-derived xenografts: analysis guidelines

--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00305R1
Full Title:	Chromatin conformation capture (Hi-C) sequencing of patient-derived xenografts: analysis guidelines
Article Type:	Research
Funding Information:	Pharmaceutical Research and Manufacturers of America Foundation (Research Informatics Award) — Dr Mikhail Dozmorov National Cancer Institute (1R01CA246182-01A1) — Dr J. Chuck Harrell Susan G. Komen (CCR19608826) — Dr J. Chuck Harrell
Abstract:	Sequencing of patient-derived xenograft (PDX) mouse models allows investigation of the molecular mechanisms of human tumor samples engrafted in a mouse host. Thus, both human and mouse genetic material is sequenced. Several methods have been developed to remove mouse sequencing reads from RNA-seq or exome sequencing PDX data and improve the downstream signal. However, for more recent chromatin conformation capture technologies (Hi-C), the effect of mouse reads remains undefined. We evaluated the effect of mouse read removal on the quality of Hi-C data using in silico created PDX Hi-C data with 10% and 30% mouse reads. Additionally, we generated two experimental PDX Hi-C datasets using different library preparation strategies. We evaluated three alignment strategies (Direct, Xenome, Combined) and three pipelines (Juicer, HiC-Pro, HiCExplorer) on Hi-C data quality. Removal of mouse reads had little-to-no effect on data quality than the results obtained with the Direct alignment strategy. Juicer extracted more valid chromatin interactions for Hi-C matrices, regardless of the mouse read removal strategy. However, the pipeline effect was minimal, while the library preparation strategy had the largest effect on all quality metrics. Together, our study presents comprehensive guidelines on PDX Hi-C data processing.
Corresponding Author:	Mikhail Dozmorov Virginia Commonwealth University Richmond, UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Virginia Commonwealth University
Corresponding Author's Secondary Institution:	
First Author:	Mikhail Dozmorov
First Author Secondary Information:	
Order of Authors:	Mikhail Dozmorov
	Katarzyna M. Tyc
	Nathan C. Sheffield
	David C. Boyd
	Amy L. Olex
	Jason Reed
	J. Chuck Harrell

Order of Authors Secondary Information:

Response to Reviewers:

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.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum	Yes

Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Chromatin conformation capture (Hi-C) sequencing of patient-derived xenografts: analysis guidelines

Mikhail G. Dozmorov[1,2,3*], Katarzyna M. Tyc[1,2,4], Nathan C. Sheffield[5], David C. Boyd[3,6], Amy L. Olex[7], Jason Reed[4,8], J. Chuck Harrell[3*]

[1]These authors contributed equally

[2]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, 23298, USA.

[3]Department of Pathology, Virginia Commonwealth University, Richmond, VA, 23284, USA.

[4]Current address: Department of Pharmacology and Toxicology, Virginia Commonwealth University, Richmond, VA, 23298, USA

[5]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, 22908, USA.

[6]Integrative Life Sciences Doctoral Program, Virginia Commonwealth University, Richmond, VA, 23298, USA.

[7]C. Kenneth and Dianne Wright Center for Clinical and Translational Research, Virginia Commonwealth University, Richmond, VA, 23298, USA.

[8]Department of Physics, Virginia Commonwealth University, Richmond, VA, 23220, USA

[*]To whom correspondence should be addressed: Mikhail.Dozmorov@vcuhealth.org, Joshua.Harrell@vcuhealth.org

Abstract

Sequencing of patient-derived xenograft (PDX) mouse models allows investigation of the molecular mechanisms of human tumor samples engrafted in a mouse host. Thus, both human and mouse genetic material is sequenced. Several methods have been developed to remove mouse sequencing reads from RNA-seq or exome sequencing PDX data and improve the downstream signal. However, for more recent chromatin conformation capture technologies (Hi-C), the effect of mouse reads remains undefined.

We evaluated the effect of mouse read removal on the quality of Hi-C data using *in silico* created PDX Hi-C data with 10% and 30% mouse reads. Additionally, we generated two experimental PDX Hi-C datasets using different library preparation strategies. We evaluated three alignment strategies (Direct, Xenome, Combined) and three pipelines (Juicer, HiC-Pro, HiCExplorer) on Hi-C data quality.

Removal of mouse reads had little-to-no effect on data quality than the results obtained with the Direct alignment strategy. Juicer extracted more valid chromatin interactions for Hi-C matrices, regardless of the mouse read removal strategy. However, the pipeline effect was minimal, while the library preparation strategy had the largest effect on all quality metrics. Together, our study presents comprehensive guidelines on PDX Hi-C data processing.

Keywords: Hi-C, chromatin conformation capture, Xenografts, PDX, Xenome

Introduction

Patient-derived tumor xenograft (PDX) mouse models are indispensable in preclinical and translational cancer research. Previous studies have demonstrated that human tumors engrafted in immunocompromised mouse models preserve each patient’s genetic heterogeneity [1] and response to treatment [2,3]. Consequently, the main application of PDX systems is to understand the molecular mechanisms of human cancers within controlled *in vivo* conditions. With the wide adoption of sequencing technologies, sequencing of PDX samples is now a standard [4–7].

High-throughput sequencing of PDX samples faces challenges not present when sequencing cell lines and homogeneous tissues. Engraftment of human cancer tissue fragments into mice leads to the rapid loss of human stroma and invasion of mouse stromal cells [1,3]. Consequently, sequencing of PDX tumor samples produces reads derived from both human and mouse genomes, with mouse read contamination ranging from 4-7% up to 20% for RNA-seq and exome data [8], and even 47% on average for whole-genome sequencing data [9]. Metastases are even more variable, and we previously identified up to 99% mouse reads in PDX RNA-seq data from lung, liver, or brain metastases [4]. Given the high similarity of human and mouse genomes, with orthologous gene products on average 85% identical [10], the presence of mouse reads introduces uncertainty in the alignment of PDX sequencing data.

Three strategies have been developed to address the removal of mouse reads from PDX sequencing data. The first strategy, referred to hereafter as “Direct,” is the direct alignment of PDX sequencing data to the human genome. The second, filtering strategy, includes separation of human and mouse reads and using only human data for downstream analysis. Xenome was among the first tools implementing filtering strategy. It classifies reads into the human, mouse, both, neither, or ambiguous categories using a 25-mer matching algorithm [11]. Despite being relatively old and lacking maintenance, Xenome remains widely used in bioinformatics pipelines [12]. We refer to this strategy as “Xenome” throughout. The third strategy involves the alignment of reads to human and mouse genomes simultaneously and then filtering reads by best alignment match [8]. This approach has been implemented in Disambiguate [13], bamcmp [14], and XenoCP [15] tools. This strategy, referred to hereafter as “Combined,” includes alignment to the *in silico* combined human-mouse reference genome to disambiguate human and mouse reads at the alignment step [4,16].

Each strategy for mouse read removal from PDX sequencing data has its own advantages and disadvantages. The Xenome and Combined strategies require extra efforts, more processing time, and in some cases doubling the requirements for computational resources. Several studies investigated the benefits of removal of contaminating mouse reads from PDX sequencing data. In DNA-seq PDX data, the removal of mouse reads reduced the false-positive rate of somatic mutation detection, especially when matching normal samples are not available [8,12,13,15,17]. In RNA-seq data, the removal of mouse reads improved gene expression quantification [15], correlation with pure human gene expression [8], and enrichment in relevant pathways [14]. Benchmarking of all three strategies using DNA sequencing convincingly demonstrated that the Xenome and Combined strategies are necessary to minimize false discovery rates in detecting genomic variants, with exome sequencing data benefiting the most [17]. The general consensus is that the removal of mouse reads from PDX sequencing data improves the extraction of human-specific signal from RNA-seq and DNA-seq PDX sequencing data [8,11–16].

Chromatin conformation capture technology and its high-throughput derivatives, such as Hi-C [18], have recently emerged as tools to assess the three-dimensional (3D) structure of the genome. Changes in the 3D structure of the genome are an established hallmark of cancer [19–21]. However, the majority of the 3D cancer genomics studies have been performed *in vitro* using cell lines [22–24]. Hi-C sequencing of PDX samples opens novel ways for understanding mechanisms of human cancers under controlled *in vivo* conditions. However, the effect of contaminating mouse reads on the quality of PDX Hi-C data, and the choice of pipeline, remains undefined.

Hi-C sequencing data possesses unique qualities that need to be considered when evaluating the effect of mouse reads in Hi-C PDX data. First, Hi-C paired-end reads are processed individually, as single-end data. Second, Hi-C data undergo extensive filtering to extract “valid pairs,” i.e., reads representative or two ligated DNA fragments with proper orientation and distance between them [25,26]. Furthermore, in contrast to typical sequencing experiments, processing of Hi-C data requires high-performance computational resources as one Hi-C experiment produces more than 20X number of reads of a typical RNA-seq experiment [27]. It remains uncertain whether efforts for removing mouse reads from PDX Hi-C data are justified and meaningfully improve the quality of human Hi-C data.

To address the effect of mouse read removal in PDX sequencing data, we evaluated three strategies for preprocessing PDX Hi-C data: Direct, Xenome, and Combined. Using different library preparation strategies, we generated two deeply sequenced Hi-C datasets of a carboplatin-resistant UCD52 breast cancer cell line [4,5]. We further created three *in silico* PDX Hi-C datasets with either 10% or 30% of mouse read contamination, mirroring the percent of mouse reads observed in our experimental Hi-C data. In particular, we used Hi-C data from normal and cancer cells to investigate whether the biological properties, such as copy number variations inherent to cancer genomes, impact the quality of Hi-C data. Human Hi-C data without mouse read contamination was used as a baseline. This design allowed us to comprehensively quantify the effect of contaminating mouse reads on the quality of Hi-C data and the downstream results.

Although several studies discuss how to process Hi-C data and what pipeline to use [25,28,29], they have not evaluated the effect of mouse read contamination. We evaluated three leading pipelines, Juicer [30], HiC-Pro [31], and HiCExplorer [32] in terms of Hi-C data quality, their ability to extract biological information, and computational runtime.

In total, we tested nine combinations of strategies–all pairwise combinations of three strategies for mouse read handling (Direct, Xenome, and Combined), and three pipelines (Juicer, HiC-Pro, and HiCExplorer)–to generate contact matrices from nine *in silico* and two experimental PDX Hi-C datasets. Furthermore, we assessed the effect of library preparation strategies on the quality of downstream results from Hi-C data. We found that removing mouse reads using Xenome or Combined strategies minimally affects the quality of Hi-C matrices and information extracted from them, while the Direct alignment yielded comparable-quality results without the additional computational overhead. The choice of processing pipeline had negligible impact on data quality and the downstream results. Ultimately, the choice of library preparation was the single variable with largest effect on data quality. From these studies, we recommend using the Direct alignment of PDX Hi-C data to the human genome. The choice of the library preparation strategy should be given priority.

Results

A comprehensive workflow for assessing the impact of mouse read contamination in PDX Hi-C data

Sequencing of biological samples from patient-derived xenograft (PDX) mouse models face a challenge of mixed genomic context derived from host (mouse) and graft (human) cells. Naturally, the goal is to sequence human-specific genomic information; however, highly homologous mouse reads may hinder the identification of human genomic information. We investigated whether the presence of mouse reads in human Hi-C data negatively affects Hi-C data quality, and whether the removal of mouse reads improves the detection of topologically associating domains (TADs) and chromatin loops. We created *in silico* PDX Hi-C data and generated two experimental PDX Hi-C datasets (Table 1, Additional File 1: Table). We assessed three alignment strategies for mouse read removal and three common pipelines to generate Hi-C matrices (Figure 1).

Figure 1. PDX Hi-C data analysis workflow. In silico (controlled mixture of human and 10/30% mouse Hi-C reads) and experimental PDX Hi-C data (two library preparation strategies) were processed using three read-alignment strategies (Direct: read alignment directly to the human genome, Xenome: human reads retrieved with Xenome tool, and Combined: all reads aligned to the combined human-mouse genome, respectively). Three pipelines (Juicer, HiC-Pro, HiCExplorer) were used to obtain Hi-C matrices. Hi-C data quality and runtime metrics were assessed following each processing step.

Table 1. Summary of *in silico* and experimental PDX Hi-C data. The proportion of mouse reads within experimental PDX Hi-C data was estimated using Xenome/Combined alignment, respectively. The optimal resolution was estimated following Direct/Xenome/Combined alignment strategy, respectively.

Hi-C data	Description	Total reads	Proportion of mouse reads	Optimal resolution (kb)
Baseline				
GM12878	Human B-lymphoblastoids	486,848,169	0%	7.0
HMEC	Human Mammary Epithelial	456,577,383	0%	7.9
KBM7	Human Myelogenous Leukemia	431,368,621	0%	8.3

CH12-LX (rep 1)	Mouse lymphoma cell line	45,594,869	100%	n/a
CH12-LX (rep 2)	Mouse lymphoma cell line	175,930,719	100%	n/a
in silico PDX				
GM12878 (10%)	GM12878 + CH12-LX (rep 1)	532,443,038	8.56%	7.0/7.1/7.0
GM12878 (30%)	GM12878 + CH12-LX (rep 2)	662,778,888	26.54%	7.0/7.1/7.0
HMEC (10%)	HMEC + CH12-LX (rep 1)	502,172,252	9.08%	7.9/7.9/7.9
HMEC (30%)	HMEC + CH12-LX (rep 2)	632,508,102	27.81%	7.9/7.9/7.9
KBM7 (10%)	KBM7 + CH12-LX (rep 1)	476,963,490	9.56%	8.3/8.3/8.3
KBM7 (30%)	KBM7 + CH12-LX (rep 2)	607,299,340	28.97%	8.3/8.3/8.3
Experimental PDX				
UCD52 Library 1	Basal-like BRCA cell line	873,892,191	12.16%/12.38%	11.5/11.9/11.7
UCD52 Library 2	Basal-like BRCA cell line	708,069,622	25.78%/29.14%	8.9/9.1/9.0

The *in silico* PDX Hi-C data were created by concatenating FASTA reads from previously published mouse and human Hi-C data [27] (see Methods). Human Hi-C data from GM12878 B-lymphoblastoid cells (nearly normal karyotype) and KBM7 myelogenous leukemia (near-haploid karyotype) were selected to assess the effect of mouse read contamination in normal and cancer Hi-C data, respectively. HMEC human mammary epithelial cells were selected to parallel breast cancer origin of our experimental PDX Hi-C data. Mouse Hi-C data from B-lymphoblast CH12-LX cells were used to create the *in silico* PDX Hi-C data with $\sim 10\%$ and $\sim 30\%$ level of mouse read contamination. Human Hi-C data for the corresponding cell lines without mouse reads were used as a baseline.

The main limitation of *in-silico* PDX Hi-C data is that human and mouse reads originate from different libraries. Although *in silico* PDX Hi-C data may be sufficient to test the performance of aligners on a mixture of highly homologous human and mouse reads, it is unknown whether this mixture can recapitulate the complexity of experimental PDX Hi-C data, where, theoretically, crosslinking and ligation of human and mouse DNA can occur. To investigate whether the removal of mouse reads from experimental PDX Hi-C data improves the quality of Hi-C matrices, we generated replicates of Hi-C data from a triple-negative breast cancer PDX (UCD52 cells), obtained with two different library preparation strategies (Library 1 and Library 2, see Methods). As expected, human-specific replicates of experimental PDX Hi-C data prepared with the same library preparation strategy showed high correlation, in contrast to those prepared with a different strategy (mean Pearson Correlation Coefficient PCC = 0.9963 and 0.9547, respectively). Mouse matrices were uniformly correlated irrespective of the library preparation strategy (mean PCC = 0.9870, Additional File 2: Figure). Therefore, replicates of Hi-C data were merged for downstream processing. In total, we processed 11 PDX Hi-C datasets (Table 1).

We applied three alignment strategies to remove mouse reads contamination: the Direct alignment of PDX Hi-C reads to the human reference genome (“Direct”), the alignment of data cleaned of mouse reads data using the Xenome tool [11] ("Xenome"), or using pre-alignment to a combined human and mouse genomes ("Combined," see Methods, Figure 1). We then applied three pipelines for processing of Hi-C data: Juicer [30], HiC-Pro [31], and HiCExplorer [32] (Figure 1). The use of different methods for mouse read removal and pipelines allowed us to establish the optimal strategy for analyzing Hi-C data derived from PDX mouse models.

Experimental PDX Hi-C data have higher proportion of ambiguously mapped reads

Xenome accurately estimated the 10%/30% proportion of mouse reads in our *in silico* PDX Hi-C data (Figure 2, Additional File 3: Table). We observed a similar proportion of mouse reads in our experimental PDX data (approximately 12% and 30%, Table 1). Less than 1% of reads were mapped to both or neither human nor mouse genomes, and these results were consistent in the *in silico* and experimental PDX Hi-C data. Compared with *in silico* PDX data, the number of "ambiguous" reads in the experimental data was higher (4-5% vs. 1%, Additional File 3: Table S2). 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. Overall, our results indicate that *in silico* PDX Hi-C data reflect the level of mouse read contamination observed in experimental settings. However, the higher level of ambiguously mapped reads suggests unique biological properties in experimental PDX Hi-C data and justifies the need for their analysis.

Figure 2. Proportions of human and mouse reads in experimental and *in silico* PDX Hi-C data. Dashed lines indicate the 10% and 30% mouse read contamination thresholds. Details of Xenome read separation statistics are collected in Additional File 3: Table.

Removal of mouse reads has negligible impact on the retrieval rate and quality of Hi-C contacts

Following data processing using all combinations of alignment strategies and pipelines, we investigated the level of residual mouse reads mismapped to the human genome. For that, we used *in silico* PDX Hi-C data where the identity of human and mouse reads are possible to track. Expectedly, following Xenome and Combined mouse read removal strategies, the data processed by any pipeline had, on average, 0.0064% of mouse reads, and these results were independent on the initial level of mouse read contamination (range 0.0002 - 0.0125%, Additional File 4: Table). Furthermore, using the Direct alignment strategy resulted in a higher level of residual mouse reads (average 0.0625%, range 0.0037-0.2402%). Juicer retained the largest proportion of mouse reads with, on average, 0.1032%/0.2250% of the initial 10% and 30% mouse read contamination, respectively, while HiC-Pro retained the smallest proportion of mouse reads (Additional File 4: Table). Thus, both HiC-Pro and HiCExplorer pipelines effectively eliminated contaminating mouse reads with direct alignment of Hi-C reads to the human genome.

We extracted four Hi-C quality metrics from the log files produced by each pipeline (all QC metrics are shown in Additional File 5: Table). **Alignment rate** is the proportion of reads aligned to the human genome. **Valid interaction pairs** is the proportion of reads marked as Hi-C contacts by each pipeline considering the valid restriction site within a reasonable distance. Higher values of those metrics indicate better data quality. **Cis/trans ratio** is the ratio of intra- vs. inter-chromosomal interacting reads. A higher cis/trans ratio indicates enrichment for within-chromosomal reads, expected in the Hi-C experiments. **Long/short ratio** is the ratio of cis interactions more than 20kb away vs. those less than 20kb away. The expectation is to capture more long-distance chromatin interactions, i.e., a long/short ratio with a value higher than 1, while the long/short ratio less than 1 indicates long interactions were lost, prompting a cautious interpretation of the results. These Hi-C quality metrics allow for the comprehensive definition of optimal alignment strategy and the effect of mouse read removal.

The removal of mouse reads had minimal-to-no effect on the alignment quality metrics of *in silico* and experimental PDX Hi-C data (Figure 3, Additional File 6: Figure). Expectedly, the alignment rate and the proportion of valid interaction pairs in *in silico* PDX Hi-C data were diminished proportionally to the percent of mouse read contamination (10% or 30%), as compared with those in pure human Hi-C data for the corresponding cell lines (dashed lines in Figure 3A, B). The removal of mouse reads from *in silico* PDX Hi-C data did not markedly affect the cis/trans ratio and long/short ratio (Figure 3C, D). These results were consistent across cell lines (Additional File 6: Figure), and suggest that, while the Direct alignment strategy retains more mismapped reads, the downstream Hi-C quality metrics perform similarly to those from data with explicitly removed mouse reads.

Similar to the results obtained with *in silico* PDX Hi-C data, the removal of mouse reads from experimental PDX Hi-C data did not markedly affect quality metrics (Figure 3), although more variability was observed (~2-4%). Interestingly, although the alignment rate of data prepared with the Library 2 strategy was lower than that of Library 1-prepared data (Figure 3A), the proportion of valid interaction pairs, cis/trans ratio, and, in particular, long/short ratio were higher (Figure 3B-D). These results suggest that the Library 2-prepared data contain more information about intra-chromosomal long- and short-distance chromatin interactions. In summary, these results indicate that the removal of mouse reads does not substantially improve or change the alignment quality of PDX Hi-C data, but the library preparation strategy has a significant effect.

Figure 3. Quality metrics for selecting the optimal pipeline for processing PDX Hi-C data. All metrics are stratified by the pipeline (Juicer, HiC-Pro, and HiCExplorer) and color-coded by the alignment strategy (Green: Direct, Blue: Xenome, Red: Combined). (A) Alignment rate representing the proportion of all aligned reads. (B) The proportion of valid interaction pairs as

determined by each pipeline. (C) The ratio of cis interacting pairs (i.e., occurring on the same chromosome) vs. trans interacting pairs (i.e., between chromosome interactions). (D) The ratio of long- vs. short-interacting Hi-C contacts. Dashed lines correspond to the baseline alignment quality metrics for human Hi-C data without mouse reads.

Evaluation of pipelines in terms of their ability to recover information from PDX Hi-C data

Although removing mouse reads using either strategy did not substantially affect the alignment quality of PDX Hi-C data (Figure 3A), we noted pipeline-specific differences (Figure 3, Additional File 7: Figure), referred by their names for brevity. Specifically, Juicer produced a similar alignment rate as HiC-Pro in *in silico* PDX Hi-C data. However, it recovered nearly 15% more alignable reads in experimental PDX Hi-C data compared to HiC-Pro. On the other hand, HiCExplorer yielded ~20% lower alignment rate for *in silico* PDX Hi-C data. Yet, HiCExplorer performed nearly as good as Juicer in the alignment of experimental PDX Hi-C data (Additional File 7: Figure A). Similarly, Juicer recovered up to 10% more valid interaction pairs in *in silico* PDX data as compared to HiC-Pro and HiCExplorer (Additional File 7: Figure B). However, in experimental PDX Hi-C data, Juicer recovered nearly twice as many valid interaction pairs as the HiC-Pro, and outperformed HiCExplorer by a ~2% margin (Additional File 7: Figure B). These results indicate that Juicer can recover more alignable reads and recover a higher proportion of valid interaction pairs. These improvements were particularly pronounced when processing experimental PDX Hi-C data.

A typical Hi-C experiment is expected to detect the majority of interactions within chromosomes (cis interactions) as compared with between chromosome (trans) interactions. This should be reflected by a high cis/trans ratio. Juicer produced Hi-C data with a higher cis/trans ratio than HiC-Pro and HiCExplorer pipelines. These results were consistent between *in silico* and experimental PDX Hi-C data (Figure 3, Additional File 7: Figure C). Juicer yielded lower long/short ratios compared to the other two pipelines (Figure 3D), which reflects the fact that

Juicer captured overwhelmingly more and most probably unwanted short-distance cis interaction (Figure 3C). These results were consistent in *in silico* and experimental PDX Hi-C data (Additional File 7: Figure D). Interestingly, HiCExplorer gave the highest long/short ratios in all *in silico* PDXs and in the experimental PDX using the Library 2 preparation strategy. Notably, all quality metrics were superior in Hi-C data obtained using the Library 2 preparation strategy. These results suggest that, altogether, HiCExplorer may offer most reliable information from PDX Hi-C data, and highlight the importance of library preparation strategy.

The presence of mouse reads has a negligible effect on the detection of TADs and chromatin loops

The most typical use of Hi-C data is to detect chromatin 3D structures, such as topologically associating domains (TADs) and chromatin loops. Given that mouse read removal strategies had negligible impact on Hi-C data quality, we used the Direct alignment strategy for the following tests. We evaluated the number of TADs and loops detected from data processed by the three pipelines. To focus on the data- and pipeline-specific differences, we used the same TAD/loop calling algorithms throughout our work (see Methods). 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.

Compared to the baseline (pure human Hi-C data), the number of cell-type-specific TADs and loops was nearly identical at the 10% or 30% level of *in silico* mouse read contamination (Figure 4, Additional File 8: Table). We also observed that TAD and loop boundaries detected from *in silico* PDX Hi-C data were highly overlapping in a condition-specific manner, and this overlap was unaffected by mouse read contamination (Additional File 9: Figure A-C, Additional File 10: Figure A-C). These results were consistent irrespectively of the pipeline and support the notion that mouse reads do not markedly affect TAD and loop boundary detection.

Library preparation strategy has the largest effect on TAD and loop detection

We observed nearly twice as many TADs and loops detected in Library 2-prepared data than those detected in Library 1-prepared data (Figure 4), paralleling our observation that Library 2-prepared data has better quality metrics (Figure 3). Using experimental Hi-C data, HiC-Pro detected the least number of TADs but a largest number of loops, while HiCExplorer detected the most TADs. Notably, the pipeline-specific differences in the number of TADs and loops were most pronounced for Library 1-prepared data (Figure 4). These results suggest that, with the optimal library preparation strategy, the differences in pipelines are negligible, further emphasizing the importance of library preparation strategy.

Similar to the analysis we did on in silico PDX Hi-C data, we investigated the agreement between TAD and loop boundaries detected from experimental PDX Hi-C data, processed with different pipelines. Given the same biological origin of experimental PDX Hi-C data, we expected a high overlap of boundaries also between the two libraries. We found boundaries detected from data prepared with the Library 2 strategy to be highly consistent irrespectively of the pipelines (Additional File 9: Figure D, Additional File 10: Figure D). In contrast, boundaries detected from Hi-C data prepared with the Library 1 strategy were most distinct and more variable. Notably, Juicer and HiCExplorer boundaries were most similar, while HiC-Pro boundaries were distinct from them (Additional File 9: Figure D, Additional File 10: Figure D). These results suggest that pipeline selection is less critical when working with high-quality data (Library 2). Of note, Juicer and HiCExplorer appear to detect concordant boundaries irrespectively of data quality.

Finally, we investigated the enrichment of CTCF, a known boundary mark, at TAD and loop boundaries. Expectedly, co-localization enrichment of CTCF was highly significant (chi-square p-value < 2.225E-308) and similar irrespectively of the initial mouse read contamination level. However, cell-line- and library-specific differences were more pronounced (Additional File 11:

Figure). Similarly, enrichment of CTCF signal was highly similar (Additional File 12-13: Figure). We observed slightly higher variability in undersequenced KBM7 data and Library 1-prepared experimental PDX data, with less significant CTCF co-localization and signal enrichment in those samples (Additional File 12-13: Figure). These results suggest that boundaries supported by biological evidence can be detected irrespectively of mouse read contamination and pipeline, and the library preparation strategy is essential for improved TAD/loop boundary detection.

Figure 4. The library preparation strategy has the largest effect on TAD/loop detection. The number of TADs (A) and loops (B) are similar at different levels of mouse reads and across pipelines in all *in silico* PDX Hi-C data, whereas experimental PDX Hi-C data produced variable results. Results for the Direct alignment strategy are shown.

Figure 5. Removal of mouse reads carries a significant computational overhead. An example of runtime (A) and storage (B) resources required for processing PDX Hi-C data to obtain Hi-C matrices. Only within-pipeline runtime comparisons are valid, as each pipelines used different computational resources (see Methods). Results for processing Library 2-prepared PDX Hi-C data are presented. Extra: accounting for time and storage space required to filter mouse reads. Main: time and storage determined for processing human reads.

Technical and runtime considerations

We compared the runtime and storage requirements for each alignment strategy and pipeline. Removal of mouse reads with either Xenome or Combined strategy resulted in smaller files and, consequently, faster processing time (Figure 5A). However, when considering the additional time needed to remove mouse reads (longest for the Combined strategy), processing of the original data (Direct) was the fastest. Together with previous observations of the minimal effect of mouse read removal on Hi-C data quality, these results indicate that extra computational time used to remove mouse reads does not appear to be beneficial for the quality of downstream results.

The removal of mouse reads requires considerable extra storage space, with the Combined strategy requiring the largest amount of additional storage (Figure 5B). Interestingly, the Juicer pipeline required the largest storage space even when processing the original data (Direct); however, it can be minimized by compressing text files produced by it. Together with additional time requirements, extra space for removing mouse reads creates a significant computational overhead with negligible benefits as compared with the Direct alignment strategy.

The choice of tools for mouse read removal is an important technical consideration requiring significant human time. Xenome, a part of the Gossamer bioinformatics suite, has not been updated since January 5, 2017 (as of October 15, 2020). It requires dependencies that can only be installed using administrative privileges, which are rarely available for bioinformaticians working in a high-performance computing environment. Furthermore, Xenome requires creating its own genome index, which also contributes to the storage and processing time, and was not included in Figure 5. The Combined strategy can be implemented ad hoc, and the combined genomes and indexes can be downloaded using Refgenie [33] (see Methods). However, the extra hard drive space and time required for mouse read removal create an unnecessary human and computational burden and can contribute to delays in a project. We recommend using the Direct alignment strategy for the most optimal computational processing of experimental PDX Hi-C data.

Discussion

We have assessed the impact of mouse read contamination on the performance of three leading pipelines for Hi-C data processing. Using quality control metrics at the alignment stage, we showed that, unlike whole-exome and RNA-seq data from PDX models, Hi-C PDX data are largely unaffected by mouse read contamination. This is not unexpected as Hi-C data processing pipelines include a series of filters to select valid pairs [25]. It is highly unlikely for experimental PDX Hi-C data to contain human-mouse chimeric reads, and even if such a read pair occurs, the probability that it would be recognized as a valid Hi-C contact (i.e., mapped in the proper orientation, within a certain distance from the nearest restriction site, etc.) is negligible. Our study confirms this reasoning and recommends the Direct alignment of PDX Hi-C data to the graft (human) genome.

Our results indicate that the Juicer pipeline may recover more alignable reads, valid interaction pairs, and achieves better cis/trans but worse long/short interaction ratios. Given that Juicer retains more misaligned mouse reads within *in silico* PDX Hi-C data (Additional File 4: Table), it remains unclear whether these reads represent true human chromatin interactions in experimental PDX Hi-C data. This performance of Juicer can be attributed to the use of the `bwa mem` aligner that can efficiently handle split-read alignment. In contrast, HiC-Pro uses `bowtie2` aligner with the default "–end-to-end" mapping settings. The documentation for HiCExplore pipeline discourages end-to-end alignment of Hi-C reads, as the alignment needs to accommodate for potential ligation junctions. Consequently, we used the `bwa mem` aligner with HiCExplorer. Given that Juicer and HiCExplorer both detected similar TAD/loop boundaries even in the poorer quality Library 1-prepared data (Additional File 9-10: Figure D), both emerge as leading tools in our study. More generally, our results suggest the use of `bwa mem`-based pipelines when processing experimental PDX Hi-C data.

Even though Juicer initially produced poor results in terms of long/short ratio metric (Figure 3D), this did not appear to affect the final number of TAD and loop detected, as well as their boundaries. Between Juicer and HiCExplorer, we find Juicer the easiest to set up for running. On the other hand, HiCExplorer comes with a comprehensive suite of tools for downstream analysis of the Hi-C matrices with no need of changing the Hi-C matrix format. Both tools perform well and we leave the choice to the user based on his/her experience to install and run the tools, as well as the ability to change between different Hi-C matrix data formats.

We identified library preparation strategy as a major determinant of the downstream data quality. While differences in quality metrics between *in silico* PDX Hi-C datasets can be attributed to the differences in sequencing depth (Additional File 1: Table), differences in our experimental PDX Hi-C data can be directly attributed to the library preparation strategies. Although our experimental PDX Hi-C data had nearly twice as many reads as the *in silico* PDX Hi-C data (Table 1), their quality metrics were inferior compared to *in silico* constructed Hi-C data (Figure 3). This was most pronounced for Library 1-prepared data, which we speculate is due to the presence of nearly 40% read duplicates, as compared to 12-15% duplicates in other datasets (Additional File 1: Table). However, the higher proportion of dangling ends, self circles, dumped reads, singletons, etc., may have contributed to the inferior quality of Library 1-prepared data (Additional File 5: Table). Similar to the ENCODE guidelines [34], our observations suggest the importance of controlling duplicates in Hi-C data.

Despite the lower number of sequencing reads and alignment rate, data obtained with Library 2 preparation strategy recovered more cis interacting Hi-C contacts spanning longer distances (cis/trans ratio and long/short ratio metrics in Figure 3C and Figure 3D, respectively). Furthermore, the number and size of TADs detected from the Library 2-prepared data was similar to that detected in *in silico* PDX Hi-C data (Figure 4). This can be attributed to multiple enzymes cutting the human genome in more than 16M sites. In contrast, the single-enzyme Library 1 preparation strategy digests the genome in about 7.2M sites. Given that Hi-C data quality significantly affects downstream results, we suggest careful inspection of the shallow sequenced library before the deep-sequencing experiment, giving particular weight to the metrics presented in Figure 3. The choice of restriction enzymes should be given primary consideration in designing PDX Hi-C experiments.

According to the ENCODE guidelines [34], we expected to recover about 58% of sequenced reads as valid Hi-C interactions. While our *in silico* PDX Hi-C data [27] almost always achieved this threshold, our experimental PDXs did not meet these criteria (~ 28 and ~ 45 for Library 1 and Library 2 preparation strategies, respectively, Additional File 5: Table). Of note, other studies report a much lower rate of valid Hi-C interactions. For instance, the average number of valid interactions across 93 Hi-C datasets was 17.72 ± 13.04 [35]. The overall lower percentage of valid interactions in our experimental Hi-C data can be partially explained by the fact that the genome of carboplatin-resistant UCD52 cells may be affected by genome rearrangements. The presence of duplications, deletions, and inversions is known to affect the genome's 3D organization [36] and may have negatively affected the performance of our experimental PDX Hi-C data. Our results suggest the need to consider the effect of large-scale genome variation in the processing of PDX Hi-C data, in addition to the standard Hi-C data quality metrics.

Methods

Generation of experimental PDX Hi-C data

UCD52 tumors were implanted in mice and once palpable treated with a single dose of 40mg/kg carboplatin, as previously described [4,5]. Once the tumors began growing again, they were treated with another dose of carboplatin. This was repeated until the tumor was no longer responsive to carboplatin. Xenograft tissue samples were processed by Phase Genomics (Seattle, WA) and Arima Genomics (San Diego, CA). Data generated using Phase Genomics/Arima Genomics library preparation strategy are referred to as "Library 1"/"Library 2," respectively. The following protocols detail each strategy, as provided by the respective service providers.

Phase Genomics (Library 1) preparation strategy

Approximately 200 mg of xenograft tissue was finely chopped and then crosslinked for 20 min at room temperature (RT) with end-over-end mixing in 1 ml of Proximo Crosslinking solution. The crosslinking reaction was terminated with a quenching solution for 20 min at RT with end-over-end mixing. Quenched tissue was rinsed once with 1X Chromatin Rinse Buffer (CRB), resuspended in Proximo Animal Lysis Buffer 1, and then transferred to Dounce Homogenizer (Kontes) and homogenized with 12 strokes using the ‘A’ homogenizer. Disrupted tissue in lysis buffer was incubated 20 min at RT. Large debris was removed following a 1 min 500xg spin. Lysate was recovered and transferred to a clean tube and pelleted by spinning at 17,000xg for 5 min. The supernatant was removed and pellet washed once with 1X CRB. After removing 1X CRB wash, the pellet was resuspended in 100 μl Proximo Lysis Buffer 2 and incubated at 65°C for 10 min Chromatin was irreversibly bound to SPRI beads by adding 100 μl SPRI beads to lysate, incubating for 10 min at RT. Beads were then washed once with 1X CRB. Beads were resuspended in 150 μl of Proximo fragmentation buffer and 5 μl of Proximo fragmentation enzyme (PN LS0027; 5,000 U/ml Sau3AI cutting at ‘GATC’) was added and incubated for 1 hour at 37°C. The sample was cooled to 12°C, and 2.5 µl of Phase Genomics Finishing Enzyme was added (PN LS0030). Sample was incubated 30 minutes at 12°C, adding 6 µl of Stop Solution (PN LS0004) at the completion of the incubation. The beads were then washed with 1X CRB and resuspended in 100 µl of Proximo Ligation Buffer supplemented with 5 µl of Proximity ligation enzyme. The proximity ligation reaction was incubated at RT for 4 hours with occasional gentle mixing. After the ligation step, 5 µl of Reverse Crosslinks enzyme (PN BR0012) was added and the reaction incubated at 65°C for 1 hour. After reversing crosslinks, the free DNA was recovered by adding 100 µl of SPRI buffer to the reaction. Beads were washed twice with 80% ethanol, air dried, and proximity ligation products were eluted (Elution Buffer, PN BR0014). DNA fragments containing proximity ligation junctions were enriched with streptavidin beads

(PN LS0020). After washing streptavidin beads twice with PG Wash Buffer 2 (PN BR0004), once with PG Wash Buffer 1 (PN BR0016), and once with molecular biology grade water, library was constructed using Proximo library reagents (PNs LS0034, LS0035, and BR0017) amplified with high-fidelity polymerase (PN BR0018), and size selected using SPRI enriching for fragments between 300 and 700 bp. Pooled libraries were sequenced on an Illumina NovaSeq 6000 instrument using an S4 flow cell. Libraries were de-multiplexed using unique dual indexes following standard Illumina methods.

Arima Genomics (Library 2) preparation strategy

Hi-C experiments were performed by Arima Genomics (San Diego, CA) according to the Arima-HiC protocols described in the Arima-HiC kit (P/N: A510008). After the Arima-HiC protocol, Illumina-compatible sequencing libraries were prepared by first shearing purified Arima-HiC proximally-ligated DNA and then size-selecting DNA fragments from ~200-600bp using SPRI beads. The size-selected fragments were then enriched for biotin and converted into Illumina-compatible sequencing libraries using the KAPA Hyper Prep kit (P/N: KK8504). After adapter ligation, DNA was PCR amplified and purified using SPRI beads. The purified DNA underwent standard QC (qPCR and Bioanalyzer) and was sequenced on the HiSeq X following the manufacturer's protocols.

Construction of ***in silico*** PDX Hi-C data

Publicly available Hi-C data from Rao et al. 2014 study [27] (GSE63525) were used to construct *in silico* PDX Hi-C data. Three human and one mouse cell line Hi-C data were selected (Table S1). To construct *in silico* PDX data containing a mixture of human and mouse reads, FASTA files from human and mouse cell lines were concatenated to form Hi-C datasets containing approximately 10% and 30% mouse reads (Table 1). If read length differed between human and mouse datasets, reads were trimmed from 3’ end to 96bp (smallest read length) using cutadapt (v2.7, [37]) before concatenation.

Removal of mouse reads from PDX Hi-C data

Three mouse read removal strategies were evaluated: Direct, Xenome, and Combined (Figure 1). In the Direct alignment strategy, all reads were mapped to the human reference genome version GRCh38/hg38 using only autosomal and sex chromosomes. In the Xenome approach, PDX Hi-C reads were processed with the `Xenome` tool [11] from the gossamer GitHub repository [38], and human only FASTA reads were kept. In the Combined strategy, the combined human-mouse genome was created by concatenating autosomal and sex chromosomes from hg38 and mm10 genomes. Chromosome names were renamed with “hg38_” or “mm10_” prefixes. Both species-specific and combined genomes, as well as the corresponding bowtie2 and bwa indexes, are available for download using `refgenie` v.0.9.3 [33]. Scripts to download and organize refgenie’s assets are provided (see “Data and code availability” section).

Raw reads were first mapped with `bwa mem -SP5` (v.0.7.17 [39]) to the combined genome, and the resulting BAM files were then subsetted with `samtools` (v.1.3.1 [40]) to keep reads mapping to the hg38 chromosomes. `bedtools bamtofastq` (v.v2.17.0 [41]) was then applied to convert the hg38-BAM files back to FASTQ format.

Processing human Hi-C data and PDX Hi-C data

All Hi-C data were processed with three pipelines with default settings: (1) `Juicer` (v.1.6 [30]), (2) `HiC-Pro` (v.3.0.0 [31]); and (3) `HiCExplorer` (v. 3.5.1 [32]). Sites for Phase Genomics cutting enzyme (GATC) were detected using (1) `generate_site_positions.py`, (2) `digest_genome.py`, and (3) `findRestSite` scripts that come with each tool, respectively. Sites for Arima Genomics cutting enzyme (^GATC, G^ANTC) were obtained from [42] (used for HiC-Pro and HiCExplorer), and generated with the `generate_site_positions.py` for Juicer pipeline. The optimal data resolution was identified using `Juicer`'s script `calculate_map_resolution.sh` and set to 10 Kb to analyze 3D genome structures for all Hi-C data.

Switching between Hi-C file formats and matrix normalization

Each pipeline adapts its own format for storing the data. Juicer saves the contact matrices into a binary `.hic` format. HiC-Pro stores results as a text file in the sparse data matrix `.matrix` and genomic coordinate `.bed` formats. HiCExplorer uses an HDF5-based binary `.h5` file format. To compare data produced by each pipeline, the data at 10kb resolution were converted to the HiCExplorer-compatible `.h5` format. HiC-Pro raw text-based contact matrices were directly converted to `h5` format with the HiCExplorer's `hicConvertFormat` tool with the default settings. Juicer's toolbox was used to extract raw text-based contact matrices with the following command: `juicer_tools_1.13.02.jar dump observed NONE file.hic chrom chrom BP 10000 outputName.txt`. The text files were then converted to HiC-Pro format using a customized R script and converted to `h5` format with the HiCExplorer's `hicConvertFormat` tool. All `h5` files were then normalized using the HiCExplorer's `hicCorrectMatrix` tool on a per chromosome basis using the Knight and Ruiz (KR) method.

Analysis of Topologically Associating Domains (TADs) and chromatin loops

HiCExplorer's `hicFindTADs` tool was applied on the KR-normalized matrices to calculate a genome-wide TAD separation score with 'minDepth,' 'maxDepth,' and 'step' set to 30 Kb, 100 Kb, and 10 Kb, respectively. 'thresholdComparisons,' and 'delta' were set to 0.05 and 0.01, 'fdr' method was chosen for 'correctForMultipleTesting.'

Similarly, HiCExplorer's `hicDetectLoops` tool was used to detect chromatin loops with the following settings: 'maxLoopDistance' set to 2000000, 'windowSize' set to 10, 'peakWidth' set to 6, 'peakInteractionsThreshold' set to 10, 'pValuePreselection' and 'pValue' both set to 0.05.

CTCF co-localization (or overlap) enrichment was assessed using GenomeRunner [43,44]. Briefly, genomic coordinates of TAD and loop boundaries were converted to the hg19 genome assembly and tested for enrichment in the consolidated Transcription Factor ChIP-seq data from ENCODE (wgEncodeRegTfbsClusteredV2 table in the UCSC genome browser). Chi-square test was used to assess co-localization enrichment and enrichment odds ratios were presented for across condition-comparisons.

CTCF signal was plotted using HiCExplorer's `computeMatrix` and `plotProfile` tools with the default settings. The `ENCFF414WYX.bigWig` CTCF track was downloaded from https://www.encodeproject.org/experiments/ENCSR000DZN/ on 12-14-2020.

Technical considerations

All jobs were run on a high-performance computer cluster under the CentOS v.6.7 operating system and the PBS job submission system PBSPro_12.2.1.140292. The Juicer pipeline was run on 1 CPU; the other pipelines were run on 16 CPUs. Due to administrative restrictions, only time and storage space were captured. The processing scripts are available at [45].

Additional Files

Additional File 1: Table. Datasets used in the current study. Selected quality metrics were obtained using FastQC v.0.11.8.

Additional File 2: Figure. Correlation between Hi-C matrices obtained from each replicate of experimental PDX samples. Experimental PDX Hi-C data were processed through Xenome to separate human and mouse reads. Human Hi-C matrices showed very high correlation, most pronounced for Library 2 preparation strategy (A). As expected, mouse Hi-C matrices were similar irrespectively of library preparation strategy. Pearson correlation coefficients were calculated for 1Mb matrices (non-zero elements only) and averaged across all chromosomes.

Additional File 3: Table. Xenome filtering statistics. The values represent the proportions of total reads in each PDX as indicated.

Additional File 4: Table. The proportion of mouse reads mismapped to the human genome in *in silico* PDX Hi-C data. 10% and 30% initial mouse read contamination, processed with Direct, Xenome, and Combined strategies, and Juicer, HiC-Pro, and HiCExplorer pipelines, 0-100% range. The % values are calculated with respect to the total number of reads that define each PDX.

Additional File 5: Table. Summary statistics used to compare the efficacy of the three Hi-C pipelines. Pipeline-specific alignment statistics are shown in the corresponding worksheets. Statistics shown in Figure 3 are highlighted in red.

Additional File 6: Figure. Quality metrics assessed to select the optimal pipeline for processing PDX Hi-C data. Observations using HMEC and KBM7 cell lines confirm the results shown in Figure 3. All metrics are stratified by the pipeline (Juicer, HiC-Pro, and HiCExplorer) and color-coded by the alignment strategy (Green: Direct alignment. Blue: Xenome selected alignment of human reads. Red: Combined human-mouse genome alignment strategy). (A) Alignment rate representing the proportion of all aligned reads. (B) The proportion of valid interaction pairs as determined by each pipeline. (C) The ratio of Cis interacting pairs (i.e., occurring on the same chromosome) vs. trans interacting pairs (i.e., between chromosome interactions). (D) The ratio of long- vs. short-interacting Hi-C contacts. Dashed lines correspond to the baseline alignment quality metrics for Hi-C data without mouse reads.

Additional File 7: Figure. Comparison of information extracted from in silico and experimental PDX Hi-C data by the alignment strategy. The same data as shown in Figure 3 and Additional File 6: Figure grouped by the mouse read removal strategy. Green: Juicer. Blue: HiC-Pro. Red: HiCExplorer. Dashed line: threshold marking the ratios equal to one.

Additional File 8: Table. The number of TADs and loops detected in each PDX Hi-C sample by each pipeline. Results for the Direct alignment strategy are shown.

Additional File 9: Figure. Overlap between TAD boundaries detected from PDX data processed by Juicer, HiC-Pro, and HiCExplorer. Multi-dimensional scaling (MDS) plots of the (1 - Jaccard overlap) distance matrices are shown. Pipeline-specific data are shown on panels A-C. Panel D shows the overlap between TAD boundaries detected in experimental PDX Hi-C data. Results for the Direct alignment strategy are shown.

Additional File 10: Figure. Overlap between loop boundaries detected from PDX data processed by Juicer, HiC-Pro, and HiCExplorer. Multi-dimensional scaling (MDS) plots of the (1 - Jaccard overlap) distance matrices are shown. Pipeline-specific data are shown on panels A-C. Panel D shows the overlap between TAD boundaries detected in experimental PDX Hi-C data. Results for the Direct alignment strategy are shown.

Additional File 11: Figure. CTCF overlap enrichment odds ratio. The CTCF enrichment odds ratios are shown at TAD (A) and loop (B) boundaries detected from the in silico and experimental PDX Hi-C data. The pipelines (X-axis) are color-coded as: green: Juicer, blue: HiC-Pro, red: HiCExplorer. Results for the Direct alignment strategy are shown.

Additional File 12: Figure. CTCF signal enrichment at TAD boundaries. CTCF signal was calculated up to 25 kb upstream and downstream from the TAD boundary (referred to as “center”) and the mean values across all TAD boundaries are plotted for each PDX tested. The pipeline-specific mean signals are color-coded as: dark blue: HiC-Pro, light Blue: HiCExplorer, yellow: Juicer. Results for the Direct alignment strategy are shown.

Additional File 13: Figure. CTCF signal enrichment at loop boundaries. CTCF signal was calculated up to 25 kb upstream and downstream from the TAD boundary (referred to as “center”) and the mean values across all TAD boundaries are plotted for each PDX tested. The pipeline-specific mean signals are color-coded as: dark blue: HiC-Pro, light Blue: HiCExplorer, yellow: Juicer. Results for the Direct alignment strategy are shown.

Availability of source code and requirements

Project name: PDX Hi-C processing

Project home page: https://github.com/dozmorovlab/PDX-HiC_processingScripts

Commented [NN1]: Make into reference.

Operating systems(s): Linux

Programming language: Shell, R (> = 4.0)

Other requirements: None

License: MIT

Any restrictions to use by non-academics: None

Availability of supporting data

Accession numbers to download the publicly available Hi-C data used in this study are listed in Table S1. Experimental PDX Hi-C data will be available at SRA upon publication (submitted). All codes necessary to reproduce the analyses are available at [45]. Snapshots of our code and other supporting data are openly available in the *GigaScience* repository, GigaDB [46].

Abbreviations

PDX - patient-derived xenograft; TADs - topologically associating domains

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by the PhRMA Foundation Research Informatics Award and the George and Lavinia Blick Research Scholarship to MD, the NIH/NCI (1R01CA246182-01A1) grant and the Susan G. Komen Foundation (CCR19608826) award to JCH.

Author contributions

M.D. and C.H. conceived the project. C.H., D.B., J.R. collected samples. N.S. created all genomic references. M.D., K.T., A.O. analyzed the data. M.D. and K.T. wrote the manuscript. All authors commented on the manuscript.

References

1. Bruna A, Rueda OM, Greenwood W, Batra AS, Callari M, Batra RN, et al. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell*. 2016; doi: 10.1016/j.cell.2016.08.041.

2. Izumchenko E, Paz K, Ciznadija D, Sloma I, Katz A, Vasquez-Dunddel D, et al. Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors. *Ann Oncol*. 2017; doi: 10.1093/annonc/mdx416.

3. DeRose YS, Wang G, Lin Y-C, Bernard PS, Buys SS, Ebbert MTW, et al. Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. *Nat Med*. 2011; doi: 10.1038/nm.2454.

4. Turner TH, Alzubi MA, Sohal SS, Olex AL, Dozmorov MG, Harrell JC. Characterizing the efficacy of cancer therapeutics in patient-derived xenograft models of metastatic breast cancer. *Breast Cancer Res Treat*. 2018; doi: 10.1007/s10549-018-4748-4.

5. Alzubi MA, Turner TH, Olex AL, Sohal SS, Tobin NP, Recio SG, et al. Separation of breast cancer and organ microenvironment transcriptomes in metastases. *Breast Cancer Res*. 2019; doi: 10.1186/s13058-019-1123-2.

6. Girotti MR, Gremel G, Lee R, Galvani E, Rothwell D, Viros A, et al. Application of sequencing, liquid biopsies, and patient-derived xenografts for personalized medicine in melanoma. *Cancer Discov*. 2016; doi: 10.1158/2159-8290.CD-15-1336.

7. Li S, Shen D, Shao J, Crowder R, Liu W, Prat A, et al. Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep*. 2013; doi: 10.1016/j.celrep.2013.08.022.

8. Rossello FJ, Tothill RW, Britt K, Marini KD, Falzon J, Thomas DM, et al. Next-generation sequence analysis of cancer xenograft models. *PLoS ONE*. 8:e744322013;

9. Lin M-T, Tseng L-H, Kamiyama H, Kamiyama M, Lim P, Hidalgo M, et al. Quantifying the relative amount of mouse and human DNA in cancer xenografts using species-specific variation in gene length. *Biotechniques*. 2010; doi: 10.2144/000113363.

10. Makałowski W, Zhang J, Boguski MS. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res*. 6:846–571996;

11. Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, Williams ED, et al. Xenome–a tool for classifying reads from xenograft samples. *Bioinformatics*. 28:i172–1782012;

12. Woo XY, Srivastava A, Graber JH, Yadav V, Sarsani VK, Simons A, et al. Genomic data analysis workflows for tumors from patient-derived xenografts (PDXs): challenges and guidelines. *BMC Med Genomics*. 12:922019;

13. Ahdesmäki MJ, Gray SR, Johnson JH, Lai Z. Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples. *F1000Res*. 2016; doi: 10.12688/f1000research.10082.2.

14. Khandelwal G, Girotti MR, Smowton C, Taylor S, Wirth C, Dynowski M, et al. Next-Generation Sequencing Analysis and Algorithms for PDX and CDX Models. *Mol Cancer Res*. 15:1012–62017;

15. Rusch M, Ding L, Arunachalam S, Thrasher A, Jin H, Macias M, et al. XenoCP: Cloud-based BAM cleansing tool for RNA and DNA from xenograft. *bioRxiv*. Cold Spring Harbor Laboratory; 2020; doi: 10.1101/843250.

16. Callari M, Batra AS, Batra RN, Sammut SJ, Greenwood W, Clifford H, et al. Computational approach to discriminate human and mouse sequences in patient-derived tumour xenografts. *BMC Genomics*. 19:192018;

17. Tso KY, Lee SD, Lo KW, Yip KY. Are special read alignment strategies necessary and cost-effective when handling sequencing reads from patient-derived tumor xenografts? *BMC Genomics*. 15:11722014;

18. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009; doi: 10.1126/science.1181369.

19. Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, et al. Oncogene-mediated alterations in chromatin conformation. *Proc Natl Acad Sci U S A*. 2012; doi: 10.1073/pnas.1112570109.

20. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*. 2016; doi: 10.1126/science.aad9024.

21. Valton A-L, Dekker J. TAD disruption as oncogenic driver. *Curr Opin Genet Dev*. 2016; doi: 10.1016/j.gde.2016.03.008.

22. Fritz AJ, Ghule PN, Boyd JR, Tye CE, Page NA, Hong D, et al. Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. *J Cell Physiol*. 2017; doi: 10.1002/jcp.25996.

23. Johnston MJ, Nikolic A, Ninkovic N, Guilhamon P, Cavalli FMG, Seaman S, et al. High-resolution structural genomics reveals new therapeutic vulnerabilities in glioblastoma. *Genome Res*. 2019; doi: 10.1101/gr.246520.118.

24. Kantidze OL, Gurova KV, Studitsky VM, Razin SV. The 3D Genome as a Target for Anticancer Therapy. *Trends Mol Med*. 26:141–92020;

25. Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to hi-c analysis: Practical guidelines. *Methods*. 2015; doi: 10.1016/j.ymeth.2014.10.031.

26. Zheng Y, Ay F, Keles S. Generative modeling of multi-mapping reads with mHi-c advances analysis of hi-c studies. *Elife*. 2019; doi: 10.7554/eLife.38070.

27. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014; doi: https://doi.org/10.1016/j.cell.2014.11.021.

28. Pal K, Forcato M, Ferrari F. Hi-C analysis: from data generation to integration. *Biophys Rev*. 11:67–782019;

29. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nat Methods*. 14:679–852017;

30. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. 3:95–82016;

31. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 16:2592015;

32. Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. 9:1892018;

33. Stolarczyk M, Reuter VP, Smith JP, Magee NE, Sheffield NC. Refgenie: A reference genome resource manager. *Gigascience*. 2020; doi: 10.1093/gigascience/giz149.

34. ENCODE project. Data Production and Processing Standard of the Hi-C Mapping Center.

35. Yang D, Jang I, Choi J, Kim MS, Lee AJ, Kim H, et al. 3DIV: A 3D-genome Interaction Viewer and database. *Nucleic Acids Res*. 46:D52–72018;

36. Chakraborty A, Ay F. Identification of copy number variations and translocations in cancer cells from hi-c data. *Bioinformatics*. 2017; doi: 10.1093/bioinformatics/btx664.

37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*. 2011; doi: 10.14806/ej.17.1.200.

38. Gossamer. *Available from https://githubcom/data61/gossamer*. Accessed 12-11-2019.

39. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 2009; doi: 10.1093/bioinformatics/btp324.

40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25:2078–92009;

41. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26:841–22010;

42. Arima restriction enzyme files. Accessed 04-09-2020.

43. Dozmorov MG, Cara LR, Giles CB, Wren JD. GenomeRunner: Automating genome exploration. *Bioinformatics*. 2012; doi: 10.1093/bioinformatics/btr666.

44. Dozmorov MG, Cara LR, Giles CB, Wren JD. GenomeRunner web server: Regulatory similarity and differences define the functional impact of SNP sets. *Bioinformatics*. 2016; doi: 10.1093/bioinformatics/btw169.

45. PDX-HiC project homepage: https://github.com/dozmorovlab/PDX-HiC_processingScripts.

46. Dozmorov M; Tyc KM; Sheffield NC; Boyd DC; Olex AL; Reed J; Harrell JC: Supporting data for "Chromatin conformation capture (Hi-C) sequencing of patient-derived xenografts: analysis guidelines" GigaScience Database. 2021. http://dx.doi.org/10.5524/100870.

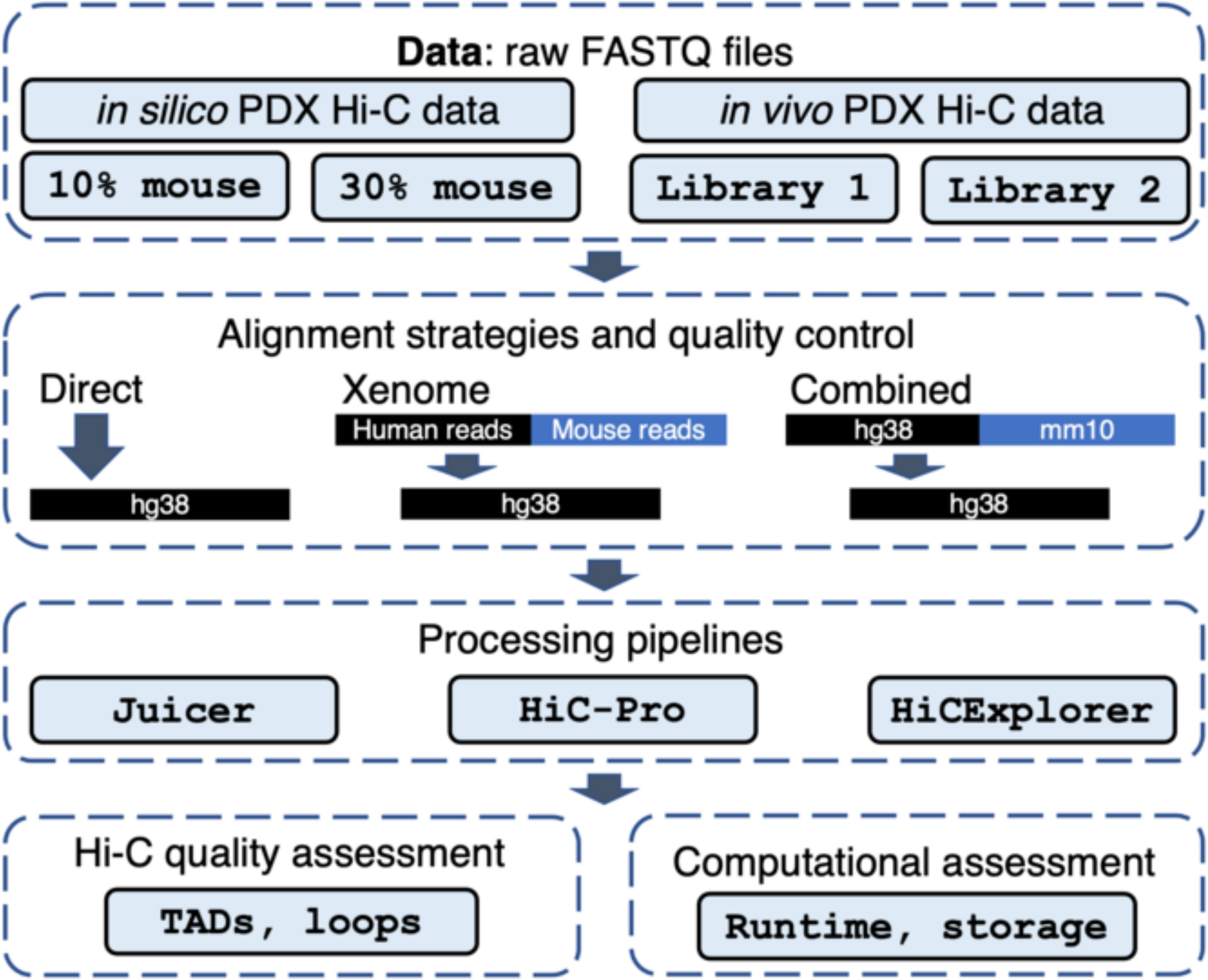
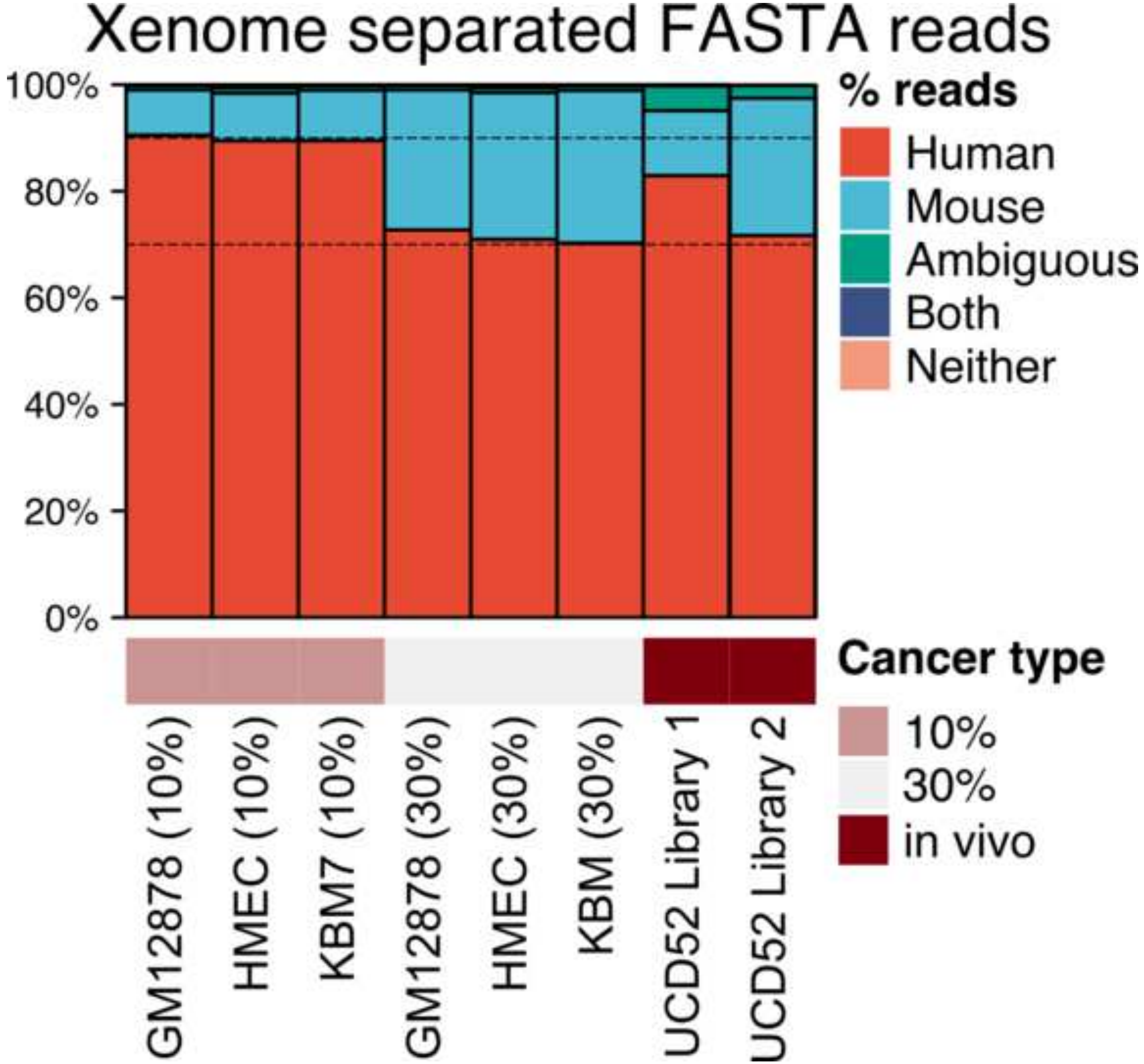

Data: raw FASTQ files
in silico PDX Hi-C data
in vivo PDX Hi-C data
10% mouse
30% mouse
Library 1
Library 2
Alignment strategies and quality control
Direct
Xenome
Combined
Human reads
Mouse reads
hg38
mm10
hg38
hg38
hg38
Processing pipelines
Juicer
HiC-Pro
HiCExplorer
Hi-C quality assessment
TADs, loops
Computational assessment
Runtime, storage

Figure 2


Xenome separated FASTA reads
100%
80%
60%
40%
20%
0%
% reads
Human
Mouse
Ambiguous
Both
Neither
Cancer type
10%
30%
in vivo
GM12878 (10%)
HMEC (10%)
KBM7 (10%)
GM12878 (30%)
HMEC (30%)
KBM (30%)
UCD52 Library 1
UCD52 Library 2

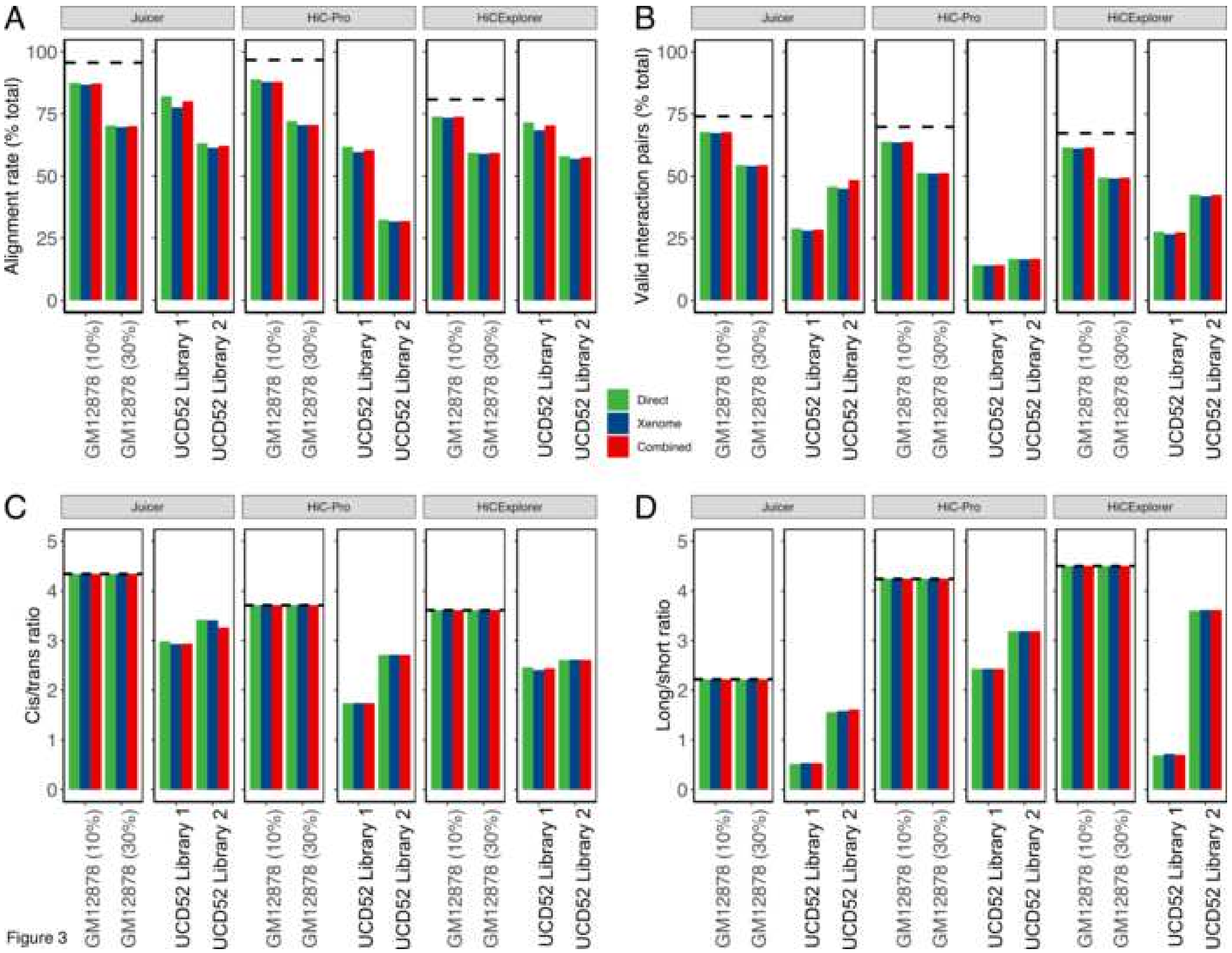
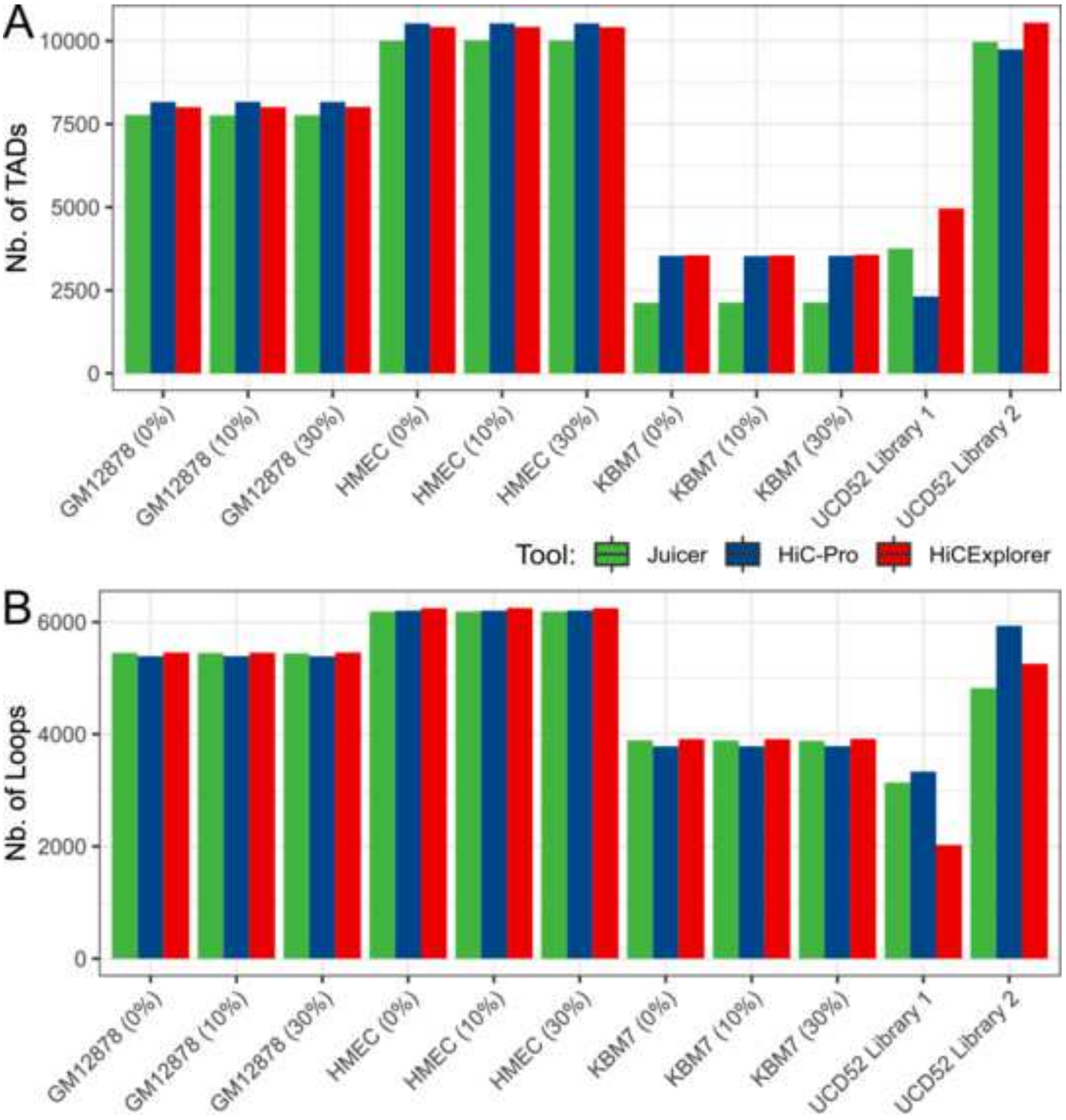
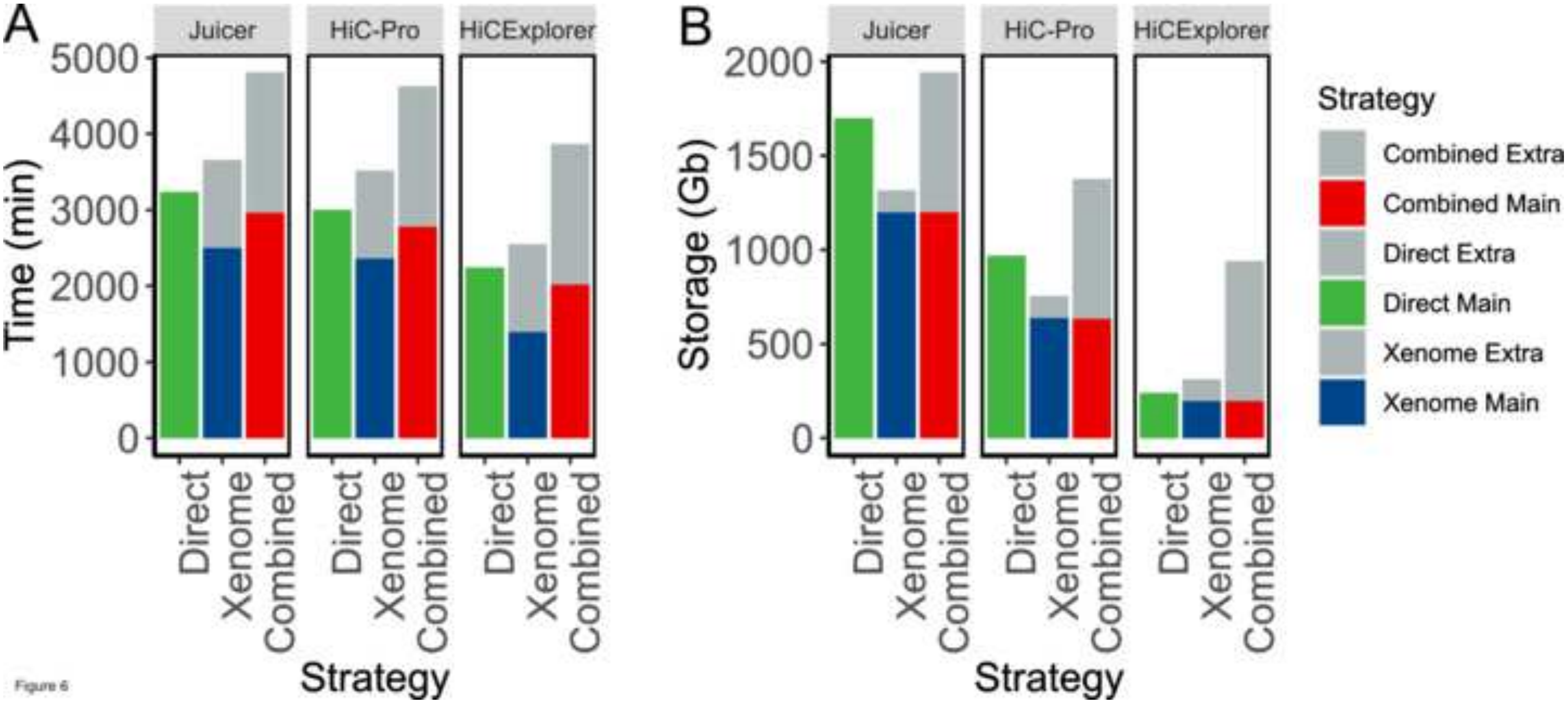

A
B
C
D
Juicer
HiC-Pro
HiCExplorer
Alignment rate (% total)
Valid interaction pairs (% total)
Cis/trans ratio
Long/short ratio
0
25
50
75
100
0
1
2
3
4
5
GM12878 (10%)
GM12878 (30%)
UCD52 Library 1
UCD52 Library 2
Direct
Xenome
Combined

Figure 3


A
Nb. of TADs
0
2500
5000
7500
10000
GM12878 (0%)
GM12878 (10%)
GM12878 (30%)
HMEC (0%)
HMEC (10%)
HMEC (30%)
KBM7 (0%)
KBM7 (10%)
KBM7 (30%)
UCD52 Library 1
UCD52 Library 2
Tool:
Juicer
HiC-Pro
HiCExplorer
B
Nb. of Loops
0
2000
4000
6000
GM12878 (0%)
GM12878 (10%)
GM12878 (30%)
HMEC (0%)
HMEC (10%)
HMEC (30%)
KBM7 (0%)
KBM7 (10%)
KBM7 (30%)
UCD52 Library 1
UCD52 Library 2


A
Juicer
HiC-Pro
HiCExplorer
5000
4000
3000
2000
1000
0
Time (min)
Direct
Xenome
Combined
Strategy
B
Storage (Gb)
2000
1500
1000
500
0
Strategy
Combined Extra
Combined Main
Direct Extra
Direct Main
Xenome Extra
Xenome Main
Figure 6

Additional File 1

Additional File 2

Additional File 3

Additional File 4

Additional File 5

Additional File 6

Additional File 7

Additional File 8

Additional File 9

Additional File 10

Additional File 11

Click here to access/download
Supplementary Material
Additional_File_11_GF_TADs_loops.png

Additional File 12

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.